Improved Optimistic Algorithms for Logistic Bandits

Louis Faury [*1 2] **Marc Abeille** [*1] **Clément Calauzènes** [1] **Olivier Fercoq** [2]

Abstract

The generalized linear bandit framework has attracted a lot of attention in recent years by extending the well-understood linear setting and allowing to model richer reward structures. It notably covers the logistic model, widely used when rewards are binary. For logistic bandits, the frequentist regret guarantees of existing algorithms are $\tilde{\mathcal{O}}(\kappa\sqrt{T})$, where κ is a problem-dependent constant. Unfortunately, κ can be arbitrarily large as it scales exponentially with the size of the decision set. This may lead to significantly loose regret bounds and poor empirical performance. In this work, we study the logistic bandit with a focus on the prohibitive dependencies introduced by κ. We propose a new optimistic algorithm based on a finer examination of the non-linearities of the reward function. We show that it enjoys a $\tilde{\mathcal{O}}(\sqrt{T})$ regret with no dependency in κ, but for a second order term. Our analysis is based on a new tail-inequality for self-normalized martingales, of independent interest.

Introduction

Parametric stochastic bandits is a framework for sequential decision making where the reward distributions associated to each arm are assumed to share a structured relationship through a common unknown parameter. It extends the standard Multi-Armed Bandit framework and allows one to address the exploration-exploitation dilemma in settings with large or infinite action space. The Linear Bandit (LB) is the most famous instance of parametrized bandits, where the value of an arm is given as the inner product between the arm feature vector and the unknown parameter. While the theoretical challenges in LB are relatively well understood and addressed (Dani et al., 2008; Rusmevichientong and Tsitsiklis, 2010; Abbasi-Yadkori et al., 2011; Abeille and Lazaric, 2017, and references therein), its practical interest is limited by the linear structure of the reward, which may fail to model real-world problems. As a result, extending LB to allow for richer reward structures and go beyond linearity has attracted a lot of attention from the bandit community in recent years. To this end, two main approaches have been investigated. Following (Valko et al., 2013), the linearity of the reward structure has been relaxed to hold only in a reproducing kernel Hilbert space. Another line of research relies on Generalized Linear Models (GLMs) to encode non-linearity through a link function. We focus in this work on the second approach.

Generalized Linear Bandits. The use of generalized linear models for the bandit setting was first studied by Filippi et al. (2010). They introduced GLM-UCB, a generic optimistic algorithm that achieves a $\tilde{O}(d\sqrt{T})$ frequentist regret. In the finite-arm case, Li et al. (2017) proposed SupCB-GLM for which they proved a $\tilde{O}(\sqrt{d\log K}\sqrt{T})$ regret bound. Similar regret guarantees were also demonstrated for Thompson Sampling, both in the frequentist (Abeille and Lazaric, 2017) and Bayesian (Russo and Van Roy, 2013; 2014; Dong and Van Roy, 2018) settings. In parallel, Jun et al. (2017) focused on improving the time and memory complexity of Generalized Linear Bandit (GLB) algorithms while Dumitrascu et al. (2018) improved posterior sampling for a Bayesian version of Thompson Sampling in the specific logistic bandits setting.

Limitations. At a first glance, existing performance guarantees for GLB seem to coincide with the state-of-the-art regret bounds for LB w.r.t. the dimension d and the horizon T. However, a careful examination of the regret bounds shows that they all depend in an *"unpleasant manner on the form of the link function of the GLM, and it seems there may be significant room for improvement"* (Lattimore and Szepesvári, 2018, §19.4.5). More in detail, they scale linearly with a multiplicative factor κ which characterizes the degree of non-linearity of the link function. As such, for highly non-linear models, κ can be prohibitively large, which drastically worsens the regret guarantees as well as the practical performances of the algorithms.

Logistic bandit. The magnitude of the constant κ is particularly significant for one GLB of crucial practical in-

[*]Equal contribution [1]Criteo AI Lab, Paris, France [2]LTCI, Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France. Correspondence to: Louis Faury <l.faury@criteo.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

terest: the logistic bandit. In this case, the link function of the GLB is the sigmoid function, resulting in a highly non-linear reward model. Hence, the associated problem-dependent constant κ is large even in typical instances. While this reduces the interest of existing guarantees for the logistic bandit, previous work suggests that there is room for improvement. In the Bayesian setting and under a slightly more specific logistic bandit instance, Dong et al. (2019) proposed a refined analysis of Thompson Sampling. Their work suggest that in some problem instances, the impact on the regret of the diameter of the decision set (directly linked to κ) might be reduced. In the frequentist setting, Filippi et al. (2010, §4.2) conjectured that GLM-UCB can be slightly modified in the hope of enjoying an improved regret bound, deflated by a factor $\kappa^{1/2}$. To the best of our knowledge, this is still an open question.

Contributions. In this work, we consider the logistic bandit problem and explicitly study its dependency with respect to κ. We propose a new non-linear study of optimistic algorithms for the logistic bandit. Our main contributions are : **1)** we answer positively to the conjecture of Filippi et al. (2010) showing that a slightly modified version of GLM-UCB enjoys a $\tilde{O}(d\sqrt{\kappa T})$ frequentist regret (Theorem 2). **2)** Further, we propose a new algorithm with yet better dependencies in κ, showing that it can be pushed in a second-order term. This results in a $\tilde{O}(d\sqrt{T}+\kappa)$ regret bound (Theorem 3). **3)** A key ingredient of our analysis is a new Bernstein-like inequality for self-normalized martingales, of independent interest (Theorem 1).

1. Preliminaries

Notations For any vector $x \in \mathbb{R}^d$ and any positive definite matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$, we will note $\|x\|_{\mathbf{M}} = \sqrt{x^\mathsf{T}\mathbf{M}x}$ the ℓ^2-norm of x weighted by $\mathbf{M}$, and $\lambda_{\min}(\mathbf{M}) > 0$ the smallest eigenvalue of $\mathbf{M}$. For two symmetric matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \succ \mathbf{B}$ means that $\mathbf{A}-\mathbf{B}$ is positive semi-definite. We will denote $\mathcal{B}_p(d) = \left\{x \in \mathbb{R}^d : \|x\|_p \leq 1\right\}$ the d-dimensional ball of radius 1 under the norm ℓ^p. For two real-valued functions f and g of a scalar variable t, we will use the notation $f_t = \tilde{\mathcal{O}}(g_t)$ to indicate that $f_t = \mathcal{O}(g_t)$ up to logarithmic factor in t. For an univariate function f we will denote $\dot{f}$ its derivative.

1.1. Setting

We consider the stochastic contextual bandit problem. At each round t, the agent observes a context and is presented a set of actions $\mathcal{X}_t$ (dependent on the context, and potentially infinite). The agent then selects an action $x_t \in \mathcal{X}_t$ and receives a reward r_{t+1}. Her decision is based on the information gathered until time t, which can be formally encoded in the filtration $\mathcal{F}_t := \left(\mathcal{F}_0, \sigma(\{x_s, r_{s+1}\}_{s=1}^{t-1})\right)$ where $\mathcal{F}_0$ represents any prior knowledge. In this paper, we assume that conditionally on the filtration $\mathcal{F}_t$, the reward r_{t+1} is binary, and is drawn from a Bernoulli distribution with parameter $\mu(x_t^\mathsf{T}\theta_*)$. The *fixed* but *unknown* parameter θ_* belongs to $\mathbb{R}^d$, and $\mu(x) := (1+\exp(-x))^{-1}$ is the sigmoid function. Formally:

$$\mathbb{P}\left(r_{t+1} = 1 \mid x_t, \mathcal{F}_t\right) = \mu\left(x_t^\mathsf{T}\theta_*\right) . \tag{1}$$

Let $x_*^t := \arg\max_{x\in\mathcal{X}_t} \mu\left(x^\mathsf{T}\theta_*\right)$ be the optimal arm. When pulling an arm, the agent suffers an instant *pseudo-regret* equal to the difference in expectation between the reward of the optimal arm x_*^t and the reward of the played arm x_t. The agent's goal is to minimize the *cumulative* pseudo-regret up to time T, defined as:

$$R_T := \sum_{t=1}^{T} \mu\left(\theta_*^\mathsf{T}x_*^t\right) - \mu\left(\theta_*^\mathsf{T}x_t\right).$$

Following (Filippi et al., 2010), we work under the subsequent assumptions on the problem structure, necessary for the study of GLBs[1] .

Assumption 1 (Bandit parameter)**.** *$\theta_* \in \Theta$ where Θ is a compact subset of $\mathbb{R}^d$. Further, $S := \max_{\theta\in\Theta}\|\theta\|_2$ is known.*

Assumption 2 (Arm set)**.** *Let $\mathcal{X} = \bigcup_{t=1}^{\infty}\mathcal{X}_t$. For all $x \in \mathcal{X}$, $\|x\|_2 \leq 1$.*

We let $L = M = 1/4$ be upper-bounds on the first and second derivative of the sigmoid function respectively. Finally, we formally introduce the parameter κ which quantifies the degree of non-linearity of the sigmoid function over the decision set $(\mathcal{X}, \Theta)$:

$$\kappa := \sup_{x\in\mathcal{X},\theta\in\Theta} 1/\dot{\mu}(x^\mathsf{T}\theta). \tag{2}$$

This key quantity and its impact are discussed in Section 2.

1.2. Reminders on Optimistic Algorithms

At round t, for a given estimator θ_t of θ_* and a given exploration bonus $\epsilon_t(x)$, we consider optimistic algorithms that play:

$$x_t = \arg\max_{x\in\mathcal{X}_t} \mu(\theta_t^\mathsf{T}x) + \epsilon_t(x) .$$

We will denote $\Delta^{\text{pred}}(x,\theta_t) := \left|\mu(x^\mathsf{T}\theta_*) - \mu(x^\mathsf{T}\theta_t)\right|$ the *prediction error* of θ_t at x. It is known that setting the bonus to be an upper-bound on the prediction error naturally gives a control on the regret. Informally:

$$\Delta^{\text{pred}}(x,\theta_t) \leq \epsilon_t(x) \implies R_T \leq 2\sum_{t=1}^{T}\epsilon_t(x_t).$$

[1] Assumption 2 is made for ease of exposition and can be easily relaxed to $\|x\|_2 \leq X$.

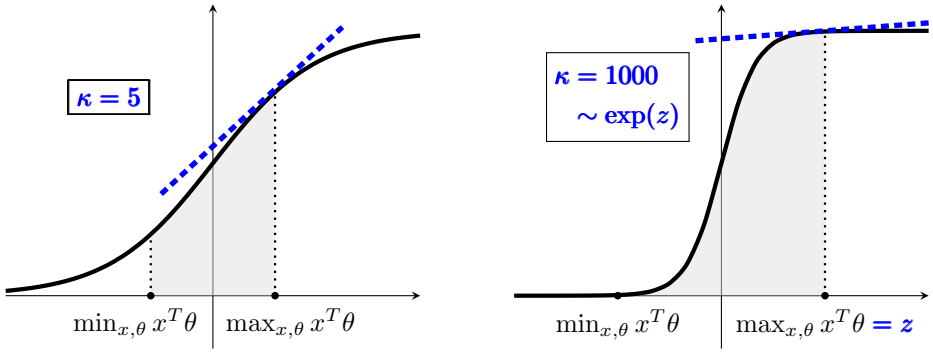


Figure 1: Visualization of the reward signal for different arm-sets and parameter-sets. Left: κ is small as the agent mostly plays in the linear part of the sigmoid, a case of little practical interest. Right: κ is significantly larger as the agent plays on a larger spectrum of the sigmoid. This case is more realistic as there exists both actions of very high and very low value.

This implication is classical and its proof is given in Section C.1 in the supplementary materials. As usual in bandit problems, tighter predictions bounds on $\Delta^{\text{pred}}(x, \theta_t)$ lead to smaller exploration bonus and therefore better regret guarantees, as long as the sequence of bonus can be shown to cumulate sub-linearly. Reciprocally, using large bonus leads to over-explorative algorithms and consequently large regret.

1.3. Maximum-Likelihood Estimate

In the logistic setting, a natural way to compute an estimator for θ_* given $\mathcal{F}_t$ derives from the maximum-likelihood principle. At round t, the regularized log-likelihood (or negative cross-entropy loss) can be written as:

$$\mathcal{L}_t^\lambda(\theta) = \sum_{s=1}^{t-1} \Big[r_{s+1} \log \mu(x_s^\mathsf{T}\theta) + (1 - r_{s+1}) \cdot \log(1 - \mu(x_s^\mathsf{T}\theta)) \Big] - \frac{\lambda}{2} \|\theta\|_2^2.$$

$\mathcal{L}_t^\lambda$ is a strictly concave function of θ for $\lambda > 0$, and the maximum likelihood estimator is defined as $\hat{\theta}_t := \arg\max_{\theta \in \mathbb{R}^d} \mathcal{L}_t^\lambda(\theta)$. In what follows, for $t \geq 1$ and $\theta \in \mathbb{R}^d$ we define $g_t(\theta)$ such as:

$$\nabla_\theta \mathcal{L}_t^\lambda(\theta) = \sum_{s=1}^{t-1} r_{s+1} x_s - \underbrace{\Big(\sum_{s=1}^{t-1} \mu(x_s^\mathsf{T}\theta) x_s + \lambda\theta \Big)}_{:= g_t(\theta)}. \quad (3)$$

We also introduce the Hessian of the log-loss:

$$\mathbf{H}_t(\theta) := \sum_{s=1}^{t-1} \dot{\mu}(x_s^\mathsf{T}\theta) x_s x_s^\mathsf{T} + \lambda \mathbf{I}_d, \quad (4)$$

as well as the design-matrix $\mathbf{V}_t := \sum_{s=1}^{t-1} x_s x_s^\mathsf{T} + \kappa\lambda \mathbf{I}_d$.

The negative log-loss $\mathcal{L}_t^\lambda(\theta)$ is known to be a *generalized self-concordant* function (Bach et al., 2010). For our purpose this boils down to the fact that $|\ddot{\mu}| \leq \dot{\mu}$.

2. Challenges and Contributions

On the scaling of κ. First, we stress the problematic scaling of κ (defined in Equation (2)) with respect to the size of the decision set $\mathcal{X} \times \Theta$. As illustrated in Figure 1, the dependency is exponential and hence prohibitive. From the definition of κ and the definition of the sigmoid function, one can easily see that:

$$\kappa \geq \exp\left(\max_{x \in \mathcal{X}} |x^\mathsf{T}\theta_*| \right). \quad (5)$$

The quantity $x^\mathsf{T}\theta_*$ is directly linked to the probability of receiving a reward when playing x. As a result, this lower bound stresses that κ will be exponentially large as soon as there exists bad (resp. good) arms x associated with a low (resp. high) probability of receiving a reward. This is unfortunately the case of most logistic bandit applications. For instance, it stands as the standard for click predictions, since the probability of observing a click is usually low (and hence κ is large). Typically, in this setting, $\mathbb{P}(\text{click}) = 10^{-3}$ and therefore $\kappa \sim 10^3$. As all existing algorithms display a linear dependency with κ (see Table 1), this narrows down the class of problem they can efficiently address. On the theoretical side, this indicates that the current analyses fail to handle the regime where the reward function is significantly non-linear, which was the primary purpose of extending LB to GLB. Note that (5) is only a lower-bound on κ. In some settings κ can be even larger: for instance when $\mathcal{X} = \mathcal{B}_2(d)$, we have $\kappa \geq \exp(S)$. Even for reasonable values of S, this has a disastrous impact on the regret bounds.

Algorithm	Regret Upper Bound	Note
GLM-UCB (Filippi et al., 2010)	$\mathcal{O}\left(\boldsymbol{\kappa} \cdot d \cdot T^{1/2} \cdot \log(T)^{3/2}\right)$	GLM
Thompson Sampling (Abeille and Lazaric, 2017)	$\mathcal{O}\left(\boldsymbol{\kappa} \cdot d^{3/2} \cdot T^{1/2} \log(T)\right)$	GLM
SupCB-GLM[2] (Li et al., 2017)	$\mathcal{O}\left(\boldsymbol{\kappa} \cdot (d \log K)^{1/2} \cdot T^{1/2} \log(T)\right)$	GLM, K actions
Logistic-UCB-1 **(this paper)**	$\mathcal{O}\left(\boldsymbol{\kappa^{1/2}} \cdot d \cdot T^{1/2} \log(T)\right)$	Logistic model
Logistic-UCB-2 **(this paper)**	$\mathcal{O}\left(d \cdot T^{1/2} \log(T) + \boldsymbol{\kappa} \cdot d^2 \cdot \log(T)^2\right)$	Logistic model

Table 1: Comparison of frequentist regret guarantees for the logistic bandit with respect to κ, d and T. κ is problem-dependent, and can be prohibitively large even for reasonable problem instances.

Uniform vs local control over $\dot{\mu}$. The presence of κ in existing regret bounds is inherited from the *learning* difficulties that arise from the logistic regression. Namely, when θ_* is large, repeatedly playing actions that are closely aligned with θ_* (a region where $\dot{\mu}$ is close to 0) will almost always lead to the same reward. This makes the estimation of θ_* in this direction *hard*. However, this should not impact the regret, as in this region the reward function is *flat*. Previous analyses ignore this fact, as they don't study the reward function *locally* but globally. More precisely, they use both uniform upper (L) and lower bounds (κ^{-1}) for the derivative of the sigmoid $\dot{\mu}$. Because they are not attained at the same point, at least one of them is loose. Alleviating the dependency in κ thus calls for an analysis and for algorithms that better handle the non-linearity of the sigmoid, switching from a uniform to a local analysis. As mentioned in Section 1.2, a thorough control on the *prediction error* Δ^{pred} is key to the tight design of an optimistic algorithm. The challenge therefore resides in finely handling the locality when controlling the prediction error.

On Filippi et al. (2010)'s conjecture. In their seminal work, Filippi et al. (2010) provided a prediction bound scaling as κ, directly impacting the size of the bonus. They however hint, by using an asymptotic argument, that this dependency could be reduced to a $\sqrt{\kappa}$. This suggests that a first limitation resides in their concentration tools. To this end, we introduce a novel Bernstein-like self-normalized martingale tail-inequality (Theorem 1) of potential independent interest. Coupled with a generalized self-concordant analysis, we give a formal proof of Filippi's asymptotic argument in the finite-time, adaptive-design case (Lemma 2). We leverage this refined prediction bound to introduce Logistic-UCB-1. We show that it suffers at most a regret in $\tilde{\mathcal{O}}(d\sqrt{\kappa T})$ (Theorem 2), improving previous guarantees by $\sqrt{\kappa}$. Our novel Bernstein inequality, together with the generalized self-concordance property of the log-loss are key ingredients of our local analysis, which allows to compare the derivatives of the sigmoid function at two different points without using L and κ^{-1}.

Dropping the κ dependency. Further challenge is to get rid of the remaining $\sqrt{\kappa}$ factor from the regret. This requires to eliminate it from the bonus of the algorithm. We show that this can be done by pushing κ to a second order term in the prediction bound (Lemma 3). Coupled with careful algorithmic design, this yields Logistic-UCB-2, for which we show a $\tilde{\mathcal{O}}(d\sqrt{T} + \kappa \log T)$ regret bound (Theorem 3), where the dependency in κ is removed from the leading term.

Outline of the following sections. Section 3 focuses on exhibiting improved upper-bound on prediction errors. We describe our algorithms and their regret bound in Section 4. Finally, we discuss our results and their implications in Section 5.

3. Improved Prediction Guarantees

This section focuses on the first challenge of the logistic bandit analysis, and aims to provide tighter prediction bounds for the logistic model. As explained earlier, bounding the prediction error relies on building tight confidence sets for θ_*, and our first contribution is to provide better adapted concentration tools to this end. Our new tail-inequality for self-normalized martingales allows to construct confidence sets with better dependencies with respect to κ.

3.1. Tail-Inequality for Self-Normalized Martingales

We present here a new, Bernstein-like tail inequality for self-normalized vectorial martingales. This inequality ex-

[2] Li et al. (2017) use a definition for κ which slightly differs from ours. However, it exhibits the same scaling in $\max |x^\top \theta_*|$. We keep this notation to ease discussions.

tends known results on self-normalized martingales (de la Pena et al., 2004; Abbasi-Yadkori et al., 2011). Compared to the concentration inequality from Theorem 1 of (Abbasi-Yadkori et al., 2011), its main novelty resides in considering martingale increments that satisfy a Bernstein-like condition instead of a sub-Gaussian condition. This allows to derive tail-inequalities for martingales "re-normalized" by their quadratic variation.

Theorem 1. *Let $\{\mathcal{F}_t\}_{t=1}^{\infty}$ be a filtration. Let $\{x_t\}_{t=1}^{\infty}$ be a stochastic process in $\mathcal{B}_2(d)$ such that x_t is $\mathcal{F}_t$ measurable. Let $\{\varepsilon_t\}_{t=2}^{\infty}$ be a martingale difference sequence such that ε_{t+1} is $\mathcal{F}_{t+1}$ measurable. Furthermore, assume that conditionally on $\mathcal{F}_t$ we have $|\varepsilon_{t+1}| \leq 1$ almost surely, and note $\sigma_t^2 := \mathbb{E}\left[\varepsilon_{t+1}^2 | \mathcal{F}_t\right]$. Let $\lambda > 0$ and for any $t \geq 1$ define:*

$$\mathbf{H}_t := \sum_{s=1}^{t-1} \sigma_s^2 x_s x_s^T + \lambda \mathbf{I}_d, \qquad S_t := \sum_{s=1}^{t-1} \varepsilon_{s+1} x_s.$$

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

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

Proof. The proof is deferred to Section A in the supplementary materials. It follows the steps of the pseudo-maximization principle introduced in (de la Pena et al., 2004), used by Abbasi-Yadkori et al. (2011) for the linear bandit and thoroughly detailed in Chapter 20 of (Lattimore and Szepesvári, 2018). The main difference in our analysis comes from the fact that we consider another super-martingale, which adds complexity to the analysis. □

Comparison to prior work. The closest inequality of this type was derived by Abbasi-Yadkori et al. (2011) to be used for the linear bandit setting. Namely, introducing $\omega := \inf_s \sigma_s^2$, it can be extracted from their Theorem 1 that that with probability at least $1 - \delta$ for all $t \geq 1$:

$$\|S_t\|_{\mathbf{V}_t^{-1}} \leq \sqrt{2d \log\left(1 + \frac{\omega t}{\lambda d}\right)}, \tag{6}$$

where $\mathbf{V}_t = \sum_{s=1}^{t-1} x_s x_s^T + (\lambda/\omega)\mathbf{I}_d$. Note that this result can be used to derive another high-probability bound on $\|S_t\|_{\mathbf{H}_t^{-1}}$. Indeed notice that $\mathbf{H}_t \succeq \omega \mathbf{V}_t$, which yields that with probability at least $1 - \delta$:

$$\|S_t\|_{\mathbf{H}_t^{-1}} \leq \frac{1}{\sqrt{\omega}} \sqrt{2d \log\left(1 + \frac{\omega t}{\lambda d}\right)}. \tag{7}$$

In contrast the bound given by Theorem 1 gives that with high-probability $\|S_t\|_{\mathbf{H}_t^{-1}} = \mathcal{O}\left(d \log(t)\right)$ which is independent of ω. This saves up the multiplicative factor $1/\sqrt{\omega}$, which is potentially very large if some ε_s have small conditional variance. However, it is lagging by a $\sqrt{d \log(t)}$ factor behind the bound provided in (7). This issue can be fixed by simply adjusting the regularization parameter. More precisely, for a given horizon T, Theorem 1 ensure that choosing a regularization parameter $\lambda = d \log(T)$ yields that on a high-probability event, for all $t \leq T$:

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

In this case, our inequality is a *strict* improvement over previous ones, which involved the scalar ω.

3.2. A New Confidence Set

We now use our new concentration inequality (Theorem 1) to derive a confidence set for θ_* that in time will lead us to upper bounds on the prediction error. We introduce:

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

with g_t defined in (3), $\mathbf{H}_t$ in (4), and where

$$\gamma_t(\delta) := \sqrt{\lambda}(S + \frac{1}{2}) + \frac{2}{\sqrt{\lambda}} \log\left(\frac{2^d}{\delta}\left(1 + \frac{Lt}{d\lambda}\right)^{\frac{d}{2}}\right).$$

A straight-forward application of Theorem 1 proves that the sets $\mathcal{C}_t(\delta)$ are confidence sets for θ_*.

Lemma 1. *Let $\delta \in (0, 1]$ and*

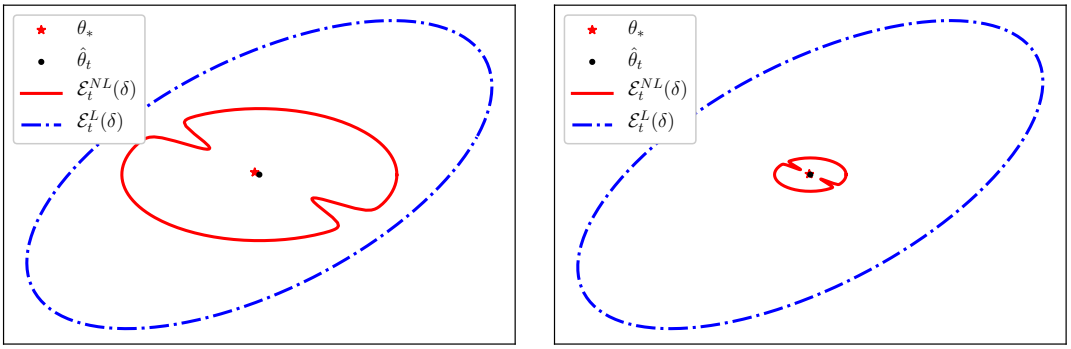
$$E_\delta := \{\forall t \geq 1,\ \theta_* \in \mathcal{C}_t(\delta)\}.$$

Then $\mathbb{P}\left(E_\delta\right) \geq 1 - \delta$.

Sketch of proof. We show $\left\|g_t(\theta_*) - g_t(\hat{\theta}_t)\right\|_{\mathbf{H}_t^{-1}(\theta_*)} \leq \gamma_t(\delta)$ with probability at least $1 - \delta$. As $\nabla \mathcal{L}_t^{\lambda}(\hat{\theta}_t) = 0$, we have

$$g_t(\hat{\theta}_t) - g_t(\theta_*) = \sum_{s=1}^{t-1} \underbrace{(r_{s+1} - \mu(\theta_*^{\mathsf{T}} x_s))}_{:=\varepsilon_{s+1}} x_s - \lambda \theta_*.$$

This equality is obtained by using the characterization of $\hat{\theta}_t$ given by the log-loss. By (1), $\{\varepsilon_{s+1}\}_{s=1}^{\infty}$ are centered Bernoulli variables with parameter $\mu(x_s^{\mathsf{T}}\theta_*)$, and variance $\sigma_s^2 = \mu(x_s^{\mathsf{T}}\theta_*)(1 - \mu(x_s^{\mathsf{T}}\theta_*)) = \dot{\mu}(x_s^T\theta_*)$. Theorem 1 leads to the claimed result up to some simple upper-bounding. The formal proof is deferred to Section B.1 in the supplementary materials. □



(a) $\kappa = 50$, $\delta = 0.05$ (b) $\kappa = 500$, $\delta = 0.05$

Figure 2: Visualization of $\mathcal{E}_t^{\mathrm{L}}(\delta)$ and $\mathcal{E}_t^{\mathrm{NL}}(\delta)$ for different values of κ. On both figures, a direction is over-sampled to highlight the non-linear nature of $\mathcal{E}_t^{\mathrm{NL}}(\delta)$. As κ grows, the difference in diameter between $\mathcal{E}_t^{\mathrm{L}}(\delta)$ and $\mathcal{E}_t^{\mathrm{NL}}(\delta)$ increases.

Illustration of confidence sets. We provide here some intuition on how this confidence set helps us improve the prediction error upper-bound. To do so, and for the ease of exposition, we will consider for the remaining of this subsection the case when $\hat{\theta}_t \in \Theta$. We back our intuition on a slightly degraded but more comprehensible version of the upper-bound on the prediction error:

$$\Delta^{\mathrm{pred}}(x, \theta) \leq L \|x\|_{\mathbf{H}_t^{-1}(\theta)} \|\theta - \theta_*\|_{\mathbf{H}_t(\theta)}.$$

The regret guarantees of our algorithms can still be recovered from this cruder upper-bound, up to some multiplicative constants (for the sake of completeness, technical details are deferred to Section B.3 in the appendix). The natural counterpart of $\mathcal{C}_t$ that allows for controlling the second part of this decomposition is a marginally inflated confidence set,

$$\mathcal{E}_t^{\mathrm{NL}}(\delta) := \left\{ \theta \in \Theta, \left\| \theta - \hat{\theta}_t \right\|_{\mathbf{H}_t(\theta)} \leq (1 + 2S)\gamma_t(\delta) \right\} .$$

It is important to notice (see Figure 2) that $\mathcal{E}_t^{\mathrm{NL}}(\delta)$ effectively handles the local curvature of the sigmoid function, as the metric $\mathbf{H}_t(\theta)$ is *local* and depends on θ. This results in a confidence set that is not an ellipsoid, and that does not penalize all estimators in the same ways.

Using the same tools as for GLM-UCB, such as the concentration result reminded in (6), a similar reasoning leads to the confidence set

$$\mathcal{E}_t^{\mathrm{L}}(\delta) := \left\{ \theta \in \Theta, \left\| \theta - \hat{\theta}_t \right\|_{\mathbf{V}_t} \leq \kappa \beta_t(\delta) \right\} ,$$

where $\beta_t(\delta)$ is a slowly increasing function of t with similar scaling as γ_t. Using global bounds on $\dot{\mu}$ leads to the appearance of κ in $\mathcal{E}_t^{\mathrm{L}}(\delta)$, illustrated by the large difference of diameter between the blue and red sets in Figure 2. This highlights the fact that the local metric $\mathbf{H}_t(\theta)$ is much better-suited than $\mathbf{V}_t$ to measure distances between parameters. The intuition laid out in this section underlies the formal improvements on the prediction error bounds we provide in the following.

3.3. Prediction Error Bounds

We are now ready to derive our new prediction guarantees, inherited from Theorem 1.

We give a first prediction bound obtained by *degrading* the local information carried by estimators in $\mathcal{C}_t(\delta)$. This guarantee is conditioned on the good event E_δ (introduced in Lemma 1), which occurs with probability at least $1 - \delta$.

Lemma 2. *On the event E_δ, for all $t \geq 1$, any $\theta \in \mathcal{C}_t(\delta)$ and $x \in \mathcal{X}$:*

$$\Delta^{\mathrm{pred}}(x, \theta) \leq L\sqrt{4 + 8S}\sqrt{\kappa}\gamma_t(\delta) \|x\|_{\mathbf{V}_t^{-1}} .$$

In term of scaling with κ, note that Lemma 2 improves the prediction bounds of (Filippi et al., 2010) by a $\sqrt{\kappa}$. It therefore matches their asymptotic argument, providing its first rigorous proof in finite-time and for the adaptive-design case. The proof is deferred to Section B.4 in the supplementary materials.

A more careful treatment of $\mathcal{C}_t(\delta)$ naturally leads to better prediction guarantees, laying the foundations to build Logistic-UCB-2. This is detailed by the following Lemma.

Lemma 3. *On the event E_δ, for all $t \geq 1$, any $\theta \in \mathcal{C}_t(\delta)$ and any $x \in \mathcal{X}$:*

$$\Delta^{\mathrm{pred}}(x, \theta) \leq (2 + 4S)\dot{\mu}(x^\mathsf{T}\theta) \|x\|_{\mathbf{H}_t^{-1}(\theta)} \gamma_t(\delta) + (4 + 8S)M\kappa\gamma_t^2(\delta) \|x\|^2_{\mathbf{V}_t^{-1}} .$$

The proof is deferred to Section B.5. The strength of this result is that it displays a first-order term that con-

Algorithm 1 Logistic-UCB-1

Input: regularization parameter λ
for $t \geq 1$ **do**
  Compute $\theta_t^{(1)}$ (Equation (8))
  Observe the contexts-action feature set $\mathcal{X}_t$.
  Play $x_t = \arg\max_{x \in \mathcal{X}_t} \mu(x^\mathsf{T}\theta_t^{(1)}) + \epsilon_{t,1}(x)$
  Observe rewards r_{t+1}.
end for

tains only *local* information about the region of the sigmoid function at hand, through the quantities $\dot{\mu}(x^T\theta)$ and $\|x\|_{\mathbf{H}_t^{-1}(\theta)}$. Global information (measured through M and κ) are pushed into a second order term that vanishes quickly. Finally, we anticipate on the fact that the decomposition displayed in Lemma 3 is not innocent. In what follows, we will show that both terms cumulate at different rates, the term involving κ becoming an explicit second order term. However, this will require a careful choice of $\theta \in \mathcal{C}_t$, as the bound on Δ^{pred} (and thus the bonus of the algorithm) now depends on θ.

4. Algorithms and Regret Bounds

4.1. Logistic-UCB-1

We introduce an algorithm leveraging Lemma 2, henceforth matching the heuristic regret bound conjectured in (Filippi et al., 2010, §4.2). We introduce the feasible estimator:

$$\theta_t^{(1)} = \arg\min_{\theta \in \Theta} \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\mathbf{H}_t^{-1}(\theta)} . \tag{8}$$

This projection step ensures us that $\theta_t^{(1)} \in \mathcal{C}_t(\delta)$ on the high-probability event E_δ. Further, we define the bonus:

$$\epsilon_{t,1}(x) = L\sqrt{4+8S}\sqrt{\kappa}\gamma_t(\delta)\,\|x\|_{\mathbf{V}_t^{-1}} .$$

We define Logistic-UCB-1 as the optimistic algorithm instantiated with $(\theta_t^{(1)}, \epsilon_{t,1}(x))$, detailed in Algorithm 1. Its regret guarantees are provided in Theorem 2, and improves previous results by $\sqrt{\kappa}$.

Theorem 2 (Regret of Logistic-UCB-1)**.** *With probability at least* $1-\delta$*:*

$$R_T^{(1)} \leq C_1 L\sqrt{\kappa}\gamma_T(\delta)\sqrt{T}$$

with $C_1 = \sqrt{32d(1+2S)\max(1, 1/(\kappa\lambda))\log\left(1+\frac{T}{\kappa\lambda d}\right)}$. *Furthermore, if* $\lambda = d\log(T)$ *then:*

$$R_T^{(1)} = \mathcal{O}\left(\sqrt{\kappa}\cdot d\cdot\sqrt{T}\log(T)\right).$$

Sketch of proof. Note that by Lemma 2 the bonus $\epsilon_{t,1}(x)$ upper-bounds $\Delta^{\text{pred}}(x, \theta_t^{(1)})$ on a high-probability event. This ensures that $R_T^{(1)} \leq 2\sum_{t=1}^{\mathsf{T}} \epsilon_{t,1}(x_t)$ with high-probability. A straight-forward application of the Elliptical Lemma (see e.g. (Abbasi-Yadkori et al., 2011), stated in Appendix D) ensures that the bonus cumulates sub-linearly and leads to the regret bound. The formal proof is deferred to Section C.2 in the supplementary material. □

Remark 1. *The projection step presented in Equation (8) is very similar to the one employed in (Filippi et al., 2010), to the difference that we use the metric* $\mathbf{H}_t(\theta)$ *instead of* $\mathbf{V}_t$*. While both lead to complex optimization programs (i.e non-convex), neither needs to be carried out when* $\hat{\theta}_t \in \Theta$*, which can be easily checked online and happens most frequently in practice.*

4.2. Logistic-UCB-2

To get rid of the last dependency in $\sqrt{\kappa}$ and improve Logistic-UCB-1, we use the improved prediction bound provided in Lemma 3. Namely, we define the bonus:

$$\epsilon_{t,2}(x,\theta) = (2+4S)\dot{\mu}(x^\mathsf{T}\theta)\,\|x\|_{\mathbf{H}_t^{-1}(\theta)}\,\gamma_t(\delta) + (4+8S)M\kappa\gamma_t^2(\delta)\,\|x\|^2_{\mathbf{V}_t^{-1}} .$$

However, as this bonus now depends on the chosen estimate θ, existing results (such as the Elliptical Lemma) do not guarantees that it sums sub-linearly. To obtain this property, we need to restrain $\mathcal{C}_t(\delta)$ to a set of admissible estimates that, intuitively, make the most of the past information already gathered. Formally, we define the *best-case log-odds* at round s by $\ell_s := \max_{\theta' \in \mathcal{C}_s(\delta)\cap\Theta} |x_s^\mathsf{T}\theta'|$, and the set of admissible log-odds at time t as:

$$\mathcal{W}_t = \left\{\theta \in \Theta \text{ s.t } |\theta^\mathsf{T}x_s| \leq \ell_s,\ \forall s \leq t-1\right\}.$$

Note that $\mathcal{W}_t$ is made up of $\min(|\mathcal{X}|, t-1)$ convex constraints, and is trivially not empty when $0_d \in \Theta$. Thanks to this new feasible set, we now define the estimator:

$$\theta_t^{(2)} := \arg\min_{\theta \in \mathcal{W}_t} \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\mathbf{H}_t^{-1}(\theta)} . \tag{9}$$

We define Logistic-UCB-2 as the optimistic bandit instantiated with $(\theta_t^{(2)}, \epsilon_{t,2}(x, \theta_t^{(2)}))$ and detailed in Algorithm 2. We state its regret upper-bound in Theorem 3. This result shows that the dominating term (in $\sqrt{T}$) of the regret is *independent* of κ. A dependency still exists, but for a *second-order term* which grows only as $\log(T)^2$.

Theorem 3 (Regret of Logistic-UCB-2)**.** *With probability at least* $1-\delta$*:*

$$R_T^{(2)} \leq C_2\gamma_T(\delta)\sqrt{T} + C_3\gamma_T^2(\delta)\kappa$$

with

$$C_2 = (4+8S)\sqrt{2dL\max(1, L/\lambda)\log\left(1+\frac{LT}{d\lambda}\right)}$$

Algorithm 2 Logistic-UCB-2

Input: regularization parameter λ
Initialize the set of admissible log-odds $\mathcal{W}_0 = \Theta$
for $t \geq 1$ **do**
  Compute $\theta_t^{(2)}$ (Equation (9))
  Observe the contexts-action feature set $\mathcal{X}_t$.
  Play $x_t = \arg\max_{x \in \mathcal{X}_t} \mu(x^\mathsf{T}\theta_t^{(2)}) + \epsilon_{t,2}(x, \theta_t^{(2)})$.
  Observe rewards r_{t+1}.
  Compute the log-odds $\ell_t = \sup_{\theta' \in \mathcal{C}_t(\delta) \cap \Theta} x_t^\mathsf{T}\theta'$.
  Add the new constraint to the feasible set:

$$\mathcal{W}_{t+1} = \mathcal{W}_t \cap \{\theta : -\ell_t \leq \theta^\mathsf{T} x_t \leq \ell_t\}$$

end for

$$C_3 = Md\max(1, 1/(\kappa\lambda))\log\left(1+\frac{T}{\kappa d\lambda}\right)(8+16S) \cdot (2+2\sqrt{1+2S}).$$

Furthermore if $\lambda = d\log(T)$ *then:*

$$R_T^{(2)} = \mathcal{O}\left(d\cdot\sqrt{T}\log(T) + \kappa\cdot d^2\cdot\log(T)^2\right).$$

The formal proof is deferred to Section C.3 in the supplementary materials. It mostly relies on the following Lemma, which ensures that the first term of $\epsilon_{t,2}(x, \theta_t^{(2)})$ cumulates sub-linearly and independently of κ (up to a second order term that grows only as $\log(T)$).

Lemma 4. *Let* $T \geq 1$. *Under the event* E_δ*:*

$$\sum_{t=1}^{T} \dot{\mu}(x_t^\mathsf{T}\theta_t^{(2)}) \|x_t\|_{\mathbf{H}_t^{-1}(\theta_t^{(2)})} \leq C_4\sqrt{T} + C_5 M\kappa\gamma_T(\delta)$$

where C_4 *and* C_5 *are independent of* κ.

Sketch of proof. The proof relies on the fact that $\theta_t^{(2)} \in \mathcal{W}_t$. Intuitively, this allows us to lower-bound $\mathbf{H}_t(\theta_t^{(2)})$ by the matrix $\sum_{s=1}^{t-1} \min_{\theta \in \mathcal{C}_s(\delta)\cap\Theta} \dot{\mu}(\theta^\mathsf{T}x_s)x_s x_s^\mathsf{T} + \lambda\mathbf{I}_d$. Note that in this case, $\min_{\theta\in\mathcal{C}_s(\delta)\cap\Theta}\dot{\mu}(\theta^\mathsf{T}x_s)$ is no longer a function of t. This, coupled with a one-step Taylor expansion of $\dot{\mu}$ allows us to use the Elliptical Lemma on a well chosen quantity and obtain the announced rates. The formal proof is deferred to Section B.6 in the supplementary materials. □

5. Discussion

In this work, we studied the scaling of optimistic logistic bandit algorithms for a particular GLM: the logistic model. We explicitly showed that previous algorithms suffered from prohibitive scaling introduced by the quantity κ, because of their sub-optimal treatment of the non-linearities of the reward signal. Thanks to a novel non-linear approach, we proved that they can be improved by deriving tighter prediction bounds. By doing so, we gave a rigorous justification for an algorithm that resembles the heuristic algorithm empirically evaluated in (Filippi et al., 2010). This algorithm exhibits a regret bound that only suffers from a $\sqrt{\kappa}$ dependency, compared to κ for previous guarantees. Further, we showed that a more careful algorithmic design leads to yet better guarantees, where the leading term of the regret is independent of κ. This result bridges the gap between logistic bandits and linear bandits, up to polynomial terms in constants of interest (e.g S).

The theoretical value of the regret upper-bound of Logistic-UCB-2 can be highlighted by comparing it to the Bayesian regret lower bound provided by Dong et al. (2019). Namely, they show that for any logistic bandit algorithm, and for any polynomial form p and $\epsilon > 0$, there exist a problem instance such that the regret is at least $\Omega(p(d)T^{1-\epsilon})$. This does not contradict our bound, as for hard problem instance one can have $\kappa = T$ in which case the second term of Logistic-UCB-2 will scale as d^2T. Note that other corner-cases instances further highlight the theoretical value of our regret bounds. Namely, note that $\kappa = \sqrt{T}$ turns GLM-UCB's regret guarantee vacuous as it will scale linearly with T. On the other hand for this case the regret of Logistic-UCB-1 scales as $T^{3/4}$, and the regret of Logistic-UCB-2 continues to scale as $\sqrt{T}$.

Extension to other GLMs. An important avenue for future work consists in extending our results to other generalized linear models. This can be done naturally by extending our work. Indeed, the properties of the sigmoid that we leverage are rather weak, and might easily transfer to other inverse link functions. We first used the fact that $\dot{\mu}$ represents the variance of the reward in order to use Theorem 1. This is not a specificity of the logistic model, but is a common relationship observed for all exponential models and their related mean function (Filippi et al., 2010, §2). We also used the generalized self-concordance property of the logistic loss, which is a consequence of the fact that $|\ddot{\mu}| \leq \dot{\mu}$. This control is quite mild, and other mean functions might display similar properties (with other constants). This is namely the case of another generalized linear model: the Poisson regression.

Randomized algorithms. The lessons learned here for optimistic algorithms might be transferred to randomized algorithms (such as Thompson Sampling) that are often preferred in practical applications thanks to their superior empirical performances. Extending our approach to such algorithms would therefore be of significant practical importance.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320.

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

Bach, F. et al. (2010). Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414.

Dani, V., Hayes, T. P., and Kakade, S. M. (2008). Stochastic Linear Optimization under Bandit Feedback. In *COLT*.

de la Pena, V. H., Klass, M. J., and Lai, T. L. (2004). Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *Annals of probability*, pages 1902–1933.

Dong, S., Ma, T., and Roy, B. V. (2019). On the Performance of Thompson Sampling on Logistic Bandits. In *Conference on Learning Theory, COLT 2019*, pages 1158–1160.

Dong, S. and Van Roy, B. (2018). An Information-Theoretic Analysis for Thompson Sampling with Many Actions. In *Advances in Neural Information Processing Systems*, pages 4157–4165.

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

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. (2010). Parametric Bandits: The Generalized Linear Case. In *Advances in Neural Information Processing Systems*, pages 586–594.

Jun, K.-S., Bhargava, A., Nowak, R., and Willett, R. (2017). Scalable Generalized Linear Bandits: Online Computation and Hashing. In *Advances in Neural Information Processing Systems*, pages 99–109.

Lattimore, T. and Szepesvári, C. (2018). Bandit Algorithms. *preprint*.

Li, L., Lu, Y., and Zhou, D. (2017). 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.

Rusmevichientong, P. and Tsitsiklis, J. N. (2010). Linearly Parameterized Bandits. *Mathematics of Operations Research*, 35(2):395–411.

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

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

Valko, M., Korda, N., Munos, R., Flaounas, I., and Cristianini, N. (2013). Finite-time Analysis of Kernelised Contextual Bandits. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 654–663.