Improved Confidence Bounds for the Linear Logistic Model and Applications to Bandits

Kwang-Sung Jun [1] **Lalit Jain** [2] **Blake Mason** [3] **Houssam Nassif** [4]

Abstract

We propose improved fixed-design confidence bounds for the linear logistic model. Our bounds significantly improve upon the state-of-the-art bound by Li et al. (2017) via recent developments of the self-concordant analysis of the logistic loss (Faury et al., 2020). Specifically, our confidence bound avoids a direct dependence on $1/\kappa$, where κ is the minimal variance over all arms' reward distributions. In general, $1/\kappa$ scales exponentially with the norm of the unknown linear parameter θ^*. Instead of relying on this worst case quantity, our confidence bound for the reward of any given arm depends directly on the variance of that arm's reward distribution. We present two applications of our novel bounds to pure exploration and regret minimization logistic bandits, improving upon state-of-the-art performance guarantees. For pure exploration we also provide a lower bound highlighting a dependence on $1/\kappa$ for a family of instances.

1. Introduction

Multi-armed bandits algorithms offer a principled approach to solve sequential decision problems under limited feedback (Thompson, 1933). In bandit problems, at each time step, an agent chooses an arm to pull from an available pool of arms and receives an associated reward. Under this setting, two major objectives arise: *pure exploration* (aka best-arm identification) where the goal is to identify the arm with the highest average reward; and *regret minimization* where the goal is to maximize the total rewards gained. Bandit algorithms are widely deployed in industry, with applications spanning news recommendation (Li et al., 2010), ads (Sawant et al., 2018; Nabi et al., 2021), online retail (Teo et al., 2016), and drug discovery (Kazerouni & Wein, 2019). In such applications, the agent often has access to a feature vector for each arm. A common assumption is that the reward is a noisy linear measurement of the underlying feature vector of the arm being pulled. In other words, the binary reward received from a pull of the arm $x \in \mathbb{R}^d$ is $y_t = x^\top \theta^* + \epsilon$, where θ^* is a latent parameter vector, and ϵ is subGaussian noise. In this case, there are several algorithms that are near-optimal and/or practical for pure exploration (Xu et al., 2018; Fiez et al., 2019) and for regret minimization (Auer, 2002; Chu et al., 2011; Dani et al., 2008; Russo & Van Roy, 2014).

Unfortunately, in abundant real-world use-cases the linear reward model is not realistic and instead rewards are binary. For example, the prevalent form of data arising from user interactions is binary click/no-click feedback in the web and e-commerce domains (Geng et al., 2020). Another example is the problem of learning the best candidate from binary pairwise comparisons, used in matching recommender systems (Biswas et al., 2019). In this setting, the agent has access to a set of items (e.g., shoes), and repeatedly chooses a pair of items to present to the user to choose from. The goal of the agent is to infer the user's favorite shoe. In this paper, we use the linear logistic model for binary feedback. In other words, the binary reward received from a pull of the arm $x \in \mathbb{R}^d$ is $y_t \sim \mathsf{Bernoulli}(\mu(x^\top \theta^*))$, where $\mu(z) := (1 + e^{-z})^{-1}$ is the logistic link function.

Existing effective bandit algorithms in the linear feedback setting attempt to estimate θ^* to drive sampling. To do so, they require tight confidence intervals on the estimated mean reward $x^\top \theta^*$ of arm x. To adapt these algorithms to the logistic model, we require confidence intervals that account for the non-linearity introduced by the link function μ. However, there is a lack of tight confidence intervals in this setting. Our work builds on previous attempts in this area to a) provide tight confidence intervals, b) adapt existing linear bandit algorithms to the logistic setting. We now detail our contributions.

The first variance-dependent fixed design confidence interval for the linear logistic model. We first consider the *fixed design* setting. Assume we have access to data $(x_s, y_s) \subset \mathbb{R}^d \times \{0, 1\}$ for $1 \in [t] := \{1, \dots, t\}$ where the reward y_s is generated according to the logistic model. In addition we assume y_s is conditionally independent from

[1]University of Arizona [2]University of Washington [3]University of Wisconsin [4]Amazon Inc. Correspondence to: Kwang-Sung Jun <kjun@cs.arizona.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

$\{x_i\}_{i=1}^t \setminus \{x_s\}$ given x_s. Let $\hat{\theta}_t$ be the maximum likelihood estimator (MLE) of θ^*. We propose the first fixed design concentration inequalities such that the width: $i)$ scales with the actual variance instead of the worst-case variance κ^{-1} that scales exponentially with $\|\theta^*\|$, and $ii)$ is independent of d. Our bound takes the form of

$$\mathbb{P}\left(|x^\top(\hat{\theta}_t - \theta^*)| \le O(\|x\|_{H_t^{-1}(\theta^*)}\sqrt{\log(t/\delta)})\right) \ge 1-\delta,$$

where $H_t(\theta^*)$ is the Fisher information matrix at θ^* matching the asymptotic bound for the MLE.[1] By contrast, the bounds by Li et al. (2017) take a significantly looser form of $O(\kappa^{-1}\|x\|_{V_t^{-1}}\sqrt{\log(1/\delta)})$ where V_t satisfies $\kappa V_t \preceq H_t(\theta^*)$. Our improvements in fixed design confidence bounds parallel that of Faury et al. (2020) for adaptive sampling, but reduce a $\sqrt{d}$ factor required by adaptive bounds. Our confidence bound is a fundamental result in statistical learning. It tightly quantifies the amount of information learned from the training set $\{x_s, y_s\}_{s=1}^t$ that transfers to a test point x, in a data-dependent non-asymptotic manner and without distributional assumptions on $\{x_s\}_{s=1}^t$. We present the full theorem and provide detailed comparisons in Section 2.

Improved pure exploration algorithms. In Section 3 we propose RAGE-GLM, a new algorithm for pure exploration in transductive linear logistic bandits, which is a novel extension of RAGE by Fiez et al. (2019). RAGE-GLM significantly improves both theoretical and empirical performance over the state-of-the-art algorithm by Kazerouni & Wein (2019), reducing the sample complexity by a multiplicative factor of κ^{-1}. We perform empirical evaluations on a pairwise comparison problem.

Novel fundamental limits for pure exploration. While the sample complexity of RAGE-GLM does not have κ^{-1} in the leading term of $\log(1/\delta)$ where δ is the target failure rate, it has an *additive* dependence on κ^{-1}. In Section 4, we show that such an additive dependence is necessary via a novel *moderate confidence* lower bound that captures the non-asymptotic complexity of learning and is independent of transportation inequality techniques (Kaufmann et al., 2016). Our results also imply that there are settings where $O(e^d)$ samples are necessary even when gaps are large, a phenomena that does not exist for linear rewards.

Improved K-armed contextual bandits. We employ our confidence bounds to develop improved algorithms for contextual logistic bandits. The proposed algorithm SupLogistic makes nontrivial extensions over the state-of-the-art algorithm SupCB-GLM by Li et al. (2017). The main challenge is to $i)$ handle the confidence width that depends on the unknown θ^*, and $ii)$ design a novel sample bucket scheme to fix an issue of SupCB-GLM that invalidates its regret bound. We show that SupLogistic enjoys a regret bound of $\tilde{O}(\sqrt{dT\log(K)})$ (ignoring $o(\sqrt{T})$ terms), which is a significant improvement over SupCB-GLM that has an extra κ^{-1} factor, along with improvements in the lower-order terms. Such an improvement parallels that of Faury et al. (2020) over UCB-GLM of Li et al. (2017) where they achieve a regret bound $\tilde{O}(d\sqrt{T})$ that shaves of the factor κ^{-1} from UCB-GLM. We discuss our improved bounds and provide more detailed comparisons in Section 5.

[1] While θ^* appears on the RHS as well, our full theorem shows that $\hat{\theta}_t$ can be used in place of θ^* with a slightly larger constant factor, although this is not useful in bandit analysis.

2. Improved Confidence Intervals for the Linear Logistic MLE

In this section we consider the fixed design setting We assume that we have a fixed $\theta^* \in \mathbb{R}^d$, a set of measurements $\{(x_s, y_s)\}_{s=1}^t \subset \mathbb{R}^d \times \mathbb{R}$ where each $y_s \in \{0,1\}$, and

$$P(y_s = 1) = \mu(x_s^\top\theta^*) = \frac{1}{1+e^{-x^\top\theta^*}}.$$

Let $\eta_s = y_s - \mu(x_s^\top\theta^*)$. Denote by $\dot{\mu}(z)$ the first order derivative of $\mu(z)$. Define $\kappa = \min_{x:\,\|x\|\le 1}\dot{\mu}(x^\top\theta^*)$.

The maximum likelihood estimate (MLE) is given by:

$$\hat{\theta} = \arg\max_{\theta\in\mathbb{R}^d}\sum_{s=1}^t y_s\log\mu(x_s^\top\theta) + (1-y_s)\log(1-\mu(x_s^\top\theta)). \tag{1}$$

We also define the Fisher information matrix at θ as

$$H_t(\theta) = \sum_{s=1}^t \dot{\mu}(x_s^\top\theta)x_s x_s^\top. \tag{2}$$

We now introduce our improved confidence interval for the linear logistic model under this fixed design setting.

Theorem 1. *Let $\delta \le e^{-1}$. Let $\hat{\theta}_t$ be the solution of Eq. (1) where, for every $s \in [t]$, y_s is conditionally independent from $\{x_i\}_{i=1}^t \setminus \{x_s\}$ given x_s (i.e. the x_s's are a fixed design). Fix $x \in \mathbb{R}^d$ with $\|x\| \le 1$. Let t_{eff} be the number of distinct vectors in $\{x_s\}_{s=1}^t$. Define $\gamma(d) = 64(d\log(6) + \log((2+t_{\mathsf{eff}})/\delta))$. Define the event $\mathcal{E}_{\mathsf{var}} = \{\forall x', \frac{1}{\sqrt{2.2}}\|x'\|_{H_t(\theta^*)^{-1}} \le \|x'\|_{H_t(\hat{\theta}_t)^{-1}} \le \sqrt{2.2}\|x'\|_{H_t(\theta^*)^{-1}}\}$. If $\xi_t^2 := \max_{s\in[t]}\|x_s\|^2_{H_t(\theta^*)^{-1}} \le \frac{1}{\gamma(d)}$, then*

$$\mathbb{P}\left(|x^\top(\hat{\theta}_t - \theta^*)| > 2.4\|x\|_{H_t(\theta^*)^{-1}}\sqrt{\log\tfrac{2(2+t_{\mathsf{eff}})}{\delta}}, \mathcal{E}_{\mathsf{var}}\right) \le \delta$$

Remark 1. *One can see that Theorem 1 implies empirical concentration inequality $|x^\top(\hat{\theta}_t - \theta^*)| \le 3.6\|x\|_{H_t(\hat{\theta}_t)^{-1}}\sqrt{\log\frac{2(2+t_{\mathsf{eff}})}{\delta}}$. This seemingly useful bound is in fact never used in our bandit analysis for technical reasons. Specifically, bandits select arms adaptively, which breaks the fixed design assumption of Theorem 1, so care is needed for algorithmic design.*

Asymptotically, under some conditions we expect for any

$x \in \mathbb{R}^d$, $x^\top(\hat{\theta} - \theta) \to N(0, \|x\|^2_{H(\theta^*)^{-1}})$ (Lehmann & Casella, 2006). Our bound matches this asymptotic rate up to constant factors.

Comparison to previous work. Our theorem is a significant improvement upon Li et al. (2017). Their bound depends on $\frac{1}{\kappa}\|x\|_{V_t^{-1}}$, with $V_t := \sum_{s=1}^t x_s x_s^\top$. In general since $\kappa V_t \preceq H_t(\theta^*)$, our bound is tighter and depends on the asymptotic variance. For the bound in (Li et al., 2017) to hold, they require that $\lambda_{\min}(V_t) \geq \Omega(d^2\kappa^{-4})$, which we call the *burn-in* condition. Recall that $\kappa^{-1} = \Theta(\exp(\|\theta^*\|))$ can be large even for moderate θ^*. In contrast, our bound's burn-in condition does not directly depend on κ^{-1}, and more importantly, it is possible to satisfy it without κ^{-1} samples in certain cases. For example, we show in Appendix B a case where a sample size of $\mathsf{polynomial}(\|\theta^*\|) \ll \exp(\|\theta^*\|)$ can satisfy the burn-in condition. For the sake of comparison, we can use the bound $\xi_t^2 \leq \kappa^{-1}\lambda_{\min}(V_t)$ to derive an inferior burn-in condition of $\lambda_{\min}(V_t) \geq \Omega(d\kappa^{-1})$. This is still a strict improvement over Li et al. (2017), saving a factor of d and a cubic factor in κ^{-1} as well as shaving off their large constant factor of 512.

We now compare our bound with that of Faury et al. (2020). The proof of Lemma 3 of Faury et al. (2020), under the assumption that $\|\theta^*\| \leq S_*$ and with a proper choice of regularization constant, implies the following confidence bound: w.p. at least $1-\delta$, $\forall t \geq 1, \forall x \in \mathbb{R}^d : \|x\| \leq 1$,

$$x^\top(\hat{\theta}_t - \theta^*) \leq \Theta\left(\|x\|_{H_t(\theta^*)^{-1}} S_*^{3/2}\sqrt{(d + \log(t/\delta))}\right) .$$

While their bound also does not directly depend on κ, it is an anytime confidence bound that holds for all $x \in \mathbb{R}^d$ simultaneously, and in addition allows for an adaptively chosen sequence of measurements. As a result, their bound suffers an additional factor of $\sqrt{d}$. Furthermore, their bound introduces a factor $S_*^{3/2}$ and requires the knowledge of both S_* and κ.[2] In contrast, our bound does not directly depend on the confidence width nor require the knowledge of S_* or κ, though these quantities may be needed to satisfy the burn-in condition.

Tightness of our bound. Empirically, we have observed that $\xi_t^2 \leq O(1)$ is necessary to control the confidence width as a function of $\|x\|_{H_t(\theta^*)^{-1}}$, but have not found a case where ξ_t^2 must be smaller than $O(1/d)$; studying the optimal burn-in condition is left as future work. We believe one can improve $\log(t_{\mathsf{eff}})$ to the metric entropy of the measurements $\{x_s\}_{s=1}^t$. Note that it is possible to remove the burn-in condition if we derive a *regularized* MLE version of Theorem 1, but this comes with an extra factor of $\sqrt{d}$ in the confidence width, which is not better than the confidence bound of Faury et al. (2020).

[2] We believe the factor $S_*^{3/2}$ can be removed by imposing an assumption on ξ_t like ours.

Proof Sketch of Theorem 1. The novelty of our argument is to exploit the variance term without introducing κ explicitly in the confidence width. We follow the main decomposition of Li et al. (2017, Theorem 1) but deviate from their proof by: i) employing Bernstein's inequality rather than Hoeffding's to obtain $H_t(\theta^*)$ in the bound (as opposed to $\kappa^{-1}V_t$); and ii) deriving a novel implicit inequality on $\max_{s\in[t]} |x_s^\top(\hat{\theta}_t - \theta^*)|$. The latter leads to the significant improvements in both d and κ^{-1} in the condition on ξ_t.

Let $\alpha_s(\hat{\theta}_t, \theta^*) := \frac{\mu(x_s^\top\hat{\theta}_t) - \mu(x_s^\top\theta^*)}{x_s^\top(\hat{\theta}_t - \theta^*)}$, $z := \sum_s^t \eta_s x_s$, and $G := \sum_s^t \alpha_s(\hat{\theta}_t, \theta^*) x_s x_s^\top$. By the optimality condition of $\hat{\theta}_t$, we have:

$$z = G(\hat{\theta}_t - \theta^*). \tag{3}$$

We use the shorthand $H := H_t(\theta^*)$ and define $E := G - H$. The main decomposition is

$$\begin{aligned} |x^\top(\hat{\theta}_t - \theta^*)| &= |x^\top(H+E)^{-1}z| \\ &\leq |x^\top H^{-1}z| + |x^\top H^{-1}E(H+E)^{-1}z| . \end{aligned} \tag{4}$$

We bound $x^\top H^{-1}z$ by $O(\|x\|_{H^{-1}}\sqrt{\log(1/\delta)})$ which uses Bernstein's inequality and the assumption on ξ_t^2. We control the second term by:

$$\begin{aligned} &|x^\top H^{-1}E(H+E)^{-1}z| \\ &\leq \|x\|_{H^{-1}}\|H^{-1/2}EH^{-1/2}\|\|G^{-1}z\|_H \\ &\overset{(a)}{\leq} \|x\|_{H^{-1}}\|H^{-1/2}EH^{-1/2}\|(1+D)\|z\|_{H^{-1}} \\ &\overset{(b)}{\leq} \|x\|_{H^{-1}}\|H^{-1/2}EH^{-1/2}\|(1+D)\cdot O(\sqrt{d+\log(1/\delta)}) \\ &\overset{(c)}{\leq} \|x\|_{H^{-1}} D \cdot O(\sqrt{d+\log(1/\delta)}), \end{aligned} \tag{5}$$

where (a) is by introducing $D := \max_{s\in[t]} |x_s^\top(\hat{\theta} - \theta^*)|$ and using the self-concordance property of the logistic loss (Faury et al., 2020) i.e. $|\ddot{\mu}| \leq \dot{\mu}$, (b) is Bernstein's inequality with a covering argument along with the assumption on ξ_t, and (c) is by a novel result which again employs self-concordance and the assumption to show that E can be bounded by $D \cdot H$. Our key observation is to apply Eq. (4) for every distinct vector x in $\{x_s\}_{s\in[t]}$ and employ $\|x\|_{H^{-1}} \leq \xi_t$ to see:

$$\begin{aligned} D &= \max_{s\in[t]} |x_s^\top(\hat{\theta} - \theta^*)| \\ &\leq O(\xi_t\sqrt{\log(1/\delta)}) + \xi_t D \cdot O(\sqrt{d+\log(1/\delta)}) . \end{aligned}$$

The assumption on ξ_t implies $\xi_t \cdot O(\sqrt{d+\log(1/\delta)}) \leq 1$. We solve for D to obtain the implicit equation

$$D = O(\xi_t\sqrt{\log(1/\delta)}) .$$

Plugging this back into Eq. (5) gives $\|x^\top(\hat{\theta} - \theta^*)\| \leq O(\|x\|_{H^{-1}}\sqrt{\log(1/\delta)})$, which holds with probability at least $1 - \Theta(t_{\mathsf{eff}}\delta)$, as we use the concentration inequality $\Theta(t_{\mathsf{eff}})$ times. To turn this into a statement that holds w.p. at least $1-\delta$, we substitute δ with $\Theta(\delta/t_{\mathsf{eff}})$, concluding the proof. See Appendix A for the statement on $\mathcal{E}_{\mathsf{var}}$. □

3. Transductive Pure Exploration

We now consider the linear logistic pure exploration problem. Specifically, the learner is given a confidence level $\delta \in (0,1)$ and has access to finite arm subsets $\mathcal{X}, \mathcal{Z} \subset \mathbb{R}^d$ but not the problem parameter $\theta^* \in \mathbb{R}^d$. The goal is to discover $z^* = \arg\max_{z\in\mathcal{Z}} z^\top\theta^*$ with probability at least $1-\delta$ using as few measurements from $\mathcal{X}$ as possible. This is a generalization of the standard linear pure exploration (Soare et al., 2014) and is referred to as the transductive setting (Fiez et al., 2019). In each time step t, the learner chooses an arm x_t, which is measurable with respect to the history $\mathcal{F}_{t-1} = \{(x_s, y_s)_{s<t}\}$, and observes a reward $y_t \in \{0,1\}$. It stops at a random stopping time τ and recommends $\hat{z} \in \mathcal{Z}$, where τ and $\hat{z}$ are both measurable w.r.t. the history $\mathcal{F}_\tau$. We assume that $\mathbb{P}(y_t = 1|x_t, \mathcal{F}_{t-1}) = \mu(x_t^\top\theta^*)$. Let $\mathbb{P}_\theta, \mathbb{E}_\theta$ denote the probability and expectation induced by the actions and rewards when the true parameter is θ. Formally, we define a δ-PAC algorithm as follows: *An algorithm for the logistic-transductive-bandit problem is δ-PAC for $(\mathcal{X}, \mathcal{Z})$ if* $\mathbb{P}_\theta(z_\tau \neq z^*) \leq \delta\ , \forall\theta \in \mathbb{R}^d$.

Example: Pairwise Judgements. As a concrete and natural example, consider an e-commerce application where the goal is to recommend an item from a set of items based on relative judgements by the user. For example, the user may be repeatedly shown two items to compare, and must choose one. The use of relative judgements is natural when trying to build a user preference profile (Jain et al., 2016) or in adaptive interactive search (Biswas et al., 2019). In each round, the system chooses two items $z, z' \in \mathcal{Z}$, and observes the binary user preference of item z or item z'. A natural model is to give each item $z \in \mathcal{Z}$ a score $z^\top\theta^*$, with the goal of finding $z^* = \max_{z\in\mathcal{Z}} z^\top\theta^*$. The probability the user prefers item z over z' is modeled by $\mathbb{P}(z > z') = \mu((z-z')^\top\theta^*)$. Hence the set of measurement vectors is naturally $\mathcal{X} = \{z - z' : z, z' \in \mathcal{Z}\}$. Note that our setting is a natural generalization of the dueling bandit setting (Yue et al., 2012). As far as we are aware, this is the first work to propose the dueling bandit problem as a natural extension of the transductive linear bandit setting under a logistic noise model.

Related Work. Soare et al. (2014) first proposed the problem of pure exploration in linear bandits with Gaussian noise when $\mathcal{X} = \mathcal{Z}$. Fiez et al. (2019) introduced the general transductive setting, provided the RAGE elimination based method which is the main inspiration for our algorithm. RAGE achieves the lower bound up to logarithmic factors with excellent empirical performance. Other works include Degenne et al. (2020); Karnin et al. (2013), which achieve the lower bound asymptotically. Finally we mention Katz-Samuels et al. (2020), which follows a similar approach to Fiez et al. (2019) but uses empirical process theory to replace the union bound over the number of arms with a Gaussian width.

Despite its importance in abundant real-life settings, pure-exploration for logisitic bandits has received little attention. The only work we are aware of is Kazerouni & Wein (2019) which defines the problem and provides an algorithm motivated by Xu et al. (2018). However, their theoretical and empirical performance are both far behind our proposed algorithm RAGE-GLM as we elaborate more below.

Notation. Define $\kappa_0 = \min_{x\in\mathcal{X}} \dot{\mu}(x^\top\theta^*)$, the smallest derivative of the link function among elements in $\mathcal{X}$ (differing slightly from the previous section). Let $\Delta_\mathcal{X} = \{\lambda \in \mathbb{R}^{|\mathcal{X}|}, \lambda \geq 0, \sum_{x\in\mathcal{X}} \lambda_x = 1\}$ be the probability simplex over $\mathcal{X}$. Given a design $\lambda \in \Delta_\mathcal{X}$, define:

$$H(\lambda,\theta) := \sum_{x\in\mathcal{X}} \lambda_x \dot{\mu}(x^\top\theta) x x^\top,\ A(\lambda) := \sum_{x\in\mathcal{X}} \lambda_x x x^\top.$$

3.1. Algorithm

Algorithm 1 RAGE-GLM

Input: ϵ, δ, $\mathcal{X}$, $\mathcal{Z}$, κ_0, effective rounding procedure **round**(n, ϵ, λ)

initialize $k = 1, \mathcal{Z}_1 = \mathcal{Z}, r(\epsilon) = d^2/\epsilon$
$\hat{\theta}_0 \leftarrow$ **BurnIn**$(\mathcal{X}, \kappa_0)$ ▷ Burn-in phase
while $|\mathcal{Z}_k| > 1$ **do** ▷ Elimination phase
  $\delta_k = \delta/(2k^2 \max\{|\mathcal{Z}|, |\mathcal{X}|\}(2 + |\mathcal{X}|))$
  $f(\lambda) := \max\Big[\gamma(d) \max_{x\in\mathcal{X}} \|x\|^2_{H(\lambda,\hat{\theta}_{k-1})^{-1}},$
    $2^{2k} \cdot (2.4)^2 \max_{z,z'\in\mathcal{Z}_k} \|z - z'\|^2_{H(\lambda,\hat{\theta}_{k-1})^{-1}}\Big]$
  $\lambda_k = \arg\min_{\lambda\in\Delta_\mathcal{X}} f(\lambda)$
  $n_k = \max\{3(1+\epsilon) f(\lambda_k) \log(1/\delta_k), r(\epsilon)\}$
  $x_1, \cdots, x_{n_k} \leftarrow$ **round**(n, ϵ, λ)
  Observe rewards $y_1, \cdots, y_{n_k} \in \{0,1\}$.
  Compute the MLE $\hat{\theta}_k$ with $\{(x_i, y_i)\}_{i=1}^{n_k}$. ▷ Eq (1)
  $\hat{z}_k = \arg\max_{z\in\mathcal{Z}_k} \hat{\theta}_k^\top z$
  $\mathcal{Z}_{k+1} \leftarrow \mathcal{Z}_k \setminus \left\{z \in \mathcal{Z}_k : \hat{\theta}_k^\top(\hat{z}_k - z) \geq 2^{-k}\right\}$
  $k \leftarrow k + 1$
return $\hat{z}_k$

Our RAGE-GLM algorithm (Alg. 1) proceeds in rounds. In each round k, it maintains a set of active arms $\mathcal{Z}_k$. It computes an experimental design that is dependent on $\hat{\theta}_{k-1}$, the estimate of θ^* from the previous round, and uses this experimental design to draw n_k samples. Then, the algorithm eliminates any arms verified to be sub-optimal using $\hat{\theta}_k$. We now go into the algorithmic details.

Rounding. In the k-th round, we have $H_k(\theta) := \sum_{s=1}^{n_k} \dot{\mu}(x_s^\top\theta) x_s x_s^\top$.[3] The algorithm utilizes an efficient rounding procedure to ensure that $H_k(\hat{\theta}_{k-1})$ is within a constant factor of $n_k \cdot H(\lambda_k, \hat{\theta}_{k-1})$. Given distribution λ_k, tolerance ϵ, and a number of samples $n_k \geq r(\epsilon)$, the pro-

[3] We abuse notation in this section and use the subscript k to denote the round, not the number of samples as in section 2.

Algorithm 2 BurnIn

Input: $\mathcal{X}, \kappa_0$

1: **initialize** $\lambda_0 = \arg\min_{\lambda \in \Delta_\mathcal{X}} \max_{x \in \mathcal{X}} \|x\|^2_{A(\lambda)^{-1}}$
2: $n_0 = 3(1+\epsilon)\kappa_0^{-1} d\gamma(d) \log(2|\mathcal{X}|(2+|\mathcal{X}|)/\delta)$
3: $x_1, \cdots, x_{n_0} \leftarrow$ **round**$(n_0, \lambda_0, \epsilon)$
4: Observe associated rewards $y_1, \cdots, y_{n_0}$.
5: **return** MLE $\hat{\theta}_0$ ▷ Use Eq (1)

cedure **round**(n_k, ϵ, λ) returns an allocation $\{x_s\}_{s=1}^{n_k}$ such that for any $\theta \in \mathbb{R}^d$, $H_k(\theta) \geq \frac{n_k}{1+\epsilon} H(\lambda, \theta)$. Efficient rounding procedures with $r(\epsilon) = d^2/\epsilon$ are described in (Fiez et al., 2019); see Appendix C for more details.

Burn-In Phase. The burn-in phase computes $\hat{\theta}_0$, an estimate of θ^* to be used in the first round. To do so, we need to guarantee that θ^* is well-estimated in all directions $\mathcal{X}$, i.e., $|x^\top(\hat{\theta} - \theta^*)| < 1, \forall x \in \mathcal{X}$. Ensuring this requires that we can employ the confidence interval in Theorem 1. Thus, burn-in Algorithm 2 must ensure that $\max_{x \in \mathcal{X}} \|x\|^2_{(\sum_{s=1}^{n_0} \dot{\mu}(x_s^\top \theta^*) x_s x_s^\top)^{-1}} \leq 1/\gamma(d)$. As we yet lack information on θ^*, we take the naive approximation:

$$\sum_{s=1}^{n_0} \dot{\mu}(\theta^{*\top} x_s) x_s x_s^\top \geq \frac{n_0}{1+\epsilon} H(\lambda, \theta) \quad \text{(from rounding)}$$
$$\geq \frac{n_0}{1+\epsilon} \kappa_0 A(\lambda),$$

and instead consider the optimization problem $\min_{\lambda \in \mathcal{X}} \max_{x \in \mathcal{X}} \|x\|^2_{A(\lambda)^{-1}}$. This is a G-optimal experimental design, and has a value of d by the Kiefer-Wolfowitz theorem (Soare et al., 2014). For the burn-in phase we assume we have access to an upper bound on κ_0^{-1}, which is equivalent to knowing an upper bound on $\|\theta^*\|$.

Experimental Design. In each round, line 5 of Algorithm 1 optimizes a convex experimental design that minimizes two objectives simultaneously. The main objective is

$$2^{2k} \min_{\lambda \in \Delta_\mathcal{X}} \max_{z, z' \in \mathcal{Z}_k} \|z - z'\|^2_{H(\lambda, \hat{\theta}_{k-1})^{-1}}. \tag{6}$$

which ensures the the gap $\theta^\top(z^* - z)$ is estimated to an error of 2^{-k} for each z. This allows us to eliminate arms whose gaps are significantly larger than 2^{-k} in each round, guaranteeing that $\mathcal{Z}_k \subset \mathcal{S}_k := \{z : (z^* - z)^\top \theta^* \leq 2 \cdot 2^{-k}\}$.

The other component of line 5 minimizes $\max_{x \in \mathcal{X}} \|x\|^2_{H(\lambda, \hat{\theta}_{k-1})^{-1}}$ similarly to the burn-in phase. This guarantees that we satisfy the conditions of Theorem 1. It additionally guarantees that the estimate $\hat{\theta}_k$ is sufficiently close to θ^* for all directions in $\mathcal{X}$. Combining this with self-concordance, $|\ddot{\mu}| \leq \dot{\mu}$, we show that $H(\lambda_k, \theta^*)$ is within a constant factor of $H(\lambda_k, \hat{\theta}_k)$ (see the Appendix C). We stop when $|\mathcal{Z}_k| = 1$ and return the remaining arm.

Theorem 2 (Sample Complexity). *Take $\delta < 1/e$ and assume $\|z\| \leq 1/2$ for all $z \in \mathcal{Z}$. Define*

$$\beta_k = \min_{\lambda \in \Delta_\mathcal{X}} \max \left[2^{2k} \max_{z, z' \in \mathcal{S}_k} \|z - z'\|^2_{H(\lambda, \theta^*)^{-1}}, \gamma(d) \max_x \|x\|^2_{H(\lambda, \theta^*)^{-1}} \right]$$

Algorithm 1 returns z^* with probability greater than $1 - 3\delta$ in a number of samples no more than

$$O\Bigg(\sum_{k=1}^{\lceil \log_2(2/\Delta_{\min}) \rceil} \beta_k \log(\max\{|\mathcal{Z}|, |\mathcal{X}|\} k^2/\delta)$$
$$+ d(1+\epsilon)\gamma(d)\kappa_0^{-1} \log(|\mathcal{X}|/\delta) + r(\epsilon) \log(\Delta_{\min}^{-1}) \Bigg)$$

where $\Delta_{\min} = \min_{z \neq z^* \in \mathcal{Z}} \langle \theta^*, z^* - z \rangle$.

Interpreting the Upper Bound. Before comparing our bound with prior work, we show concrete examples that show the strength of our sample complexity bound.

Example 1. Consider a simple setting where $\mathcal{Z} = \mathcal{X} = \{e_1, e_2\} \subset \mathbb{R}^2$, and $\theta^* = (r, r - \epsilon)$, for $r \geq 0$. In this case, $\kappa_0^{-1} = \max_{i \in \{1,2\}} \dot{\mu}(z_i^\top \theta^*)^{-1} \leq e^r$. Thus in the burn-in phase, we take roughly $\tilde{O}(e^r)$ samples. Now, for small ϵ, the minimizer of $\min_{\lambda \in \Delta_\mathcal{X}} \|e_1 - e_2\|^2_{H(\theta^*)^{-1}}$ places roughly equal mass on e_1 and e_2, giving an objective value that is roughly bounded by e^r. Thus the sample complexity of Algorithm 1 is $O(\sum_{k=1}^{\log_2(1/\epsilon)} 2^{2k} e^r \log(1/\delta)) \approx \frac{e^r}{\epsilon^2}$.

Note this problem is equivalent to a standard best-arm identification algorithm with two Bernoulli arms (Kaufmann et al., 2015). Standard results in Pure Exploration show that a lower bound on this problem is given by the KL-divergence $\mathsf{KL}(\mathsf{Bernoulli}(\mu(\theta^\top z_1)), \mathsf{Bernoulli}(\mu(\theta^\top z_2)))^{-1} \approx \frac{e^r}{2\epsilon^2}$ for sufficiently small ϵ. This shows that our bound is at least no worse than the well-studied unstructured case.

Example 2. We extend the above setting and consider $\mathcal{X} = \{e_1, e_2, e_1 - e_2\}$, $\mathcal{Z} = \{e_1, e_2\}$ and the same θ^*. As above, the burn-in phase requires $\kappa^{-1} \approx e^r$ samples. Starting from the first round, our computed experimental design will place all of its samples on the third arm. In this case, $\min_\lambda \|e_1 - e_2\|^2_{H(\theta^*)^{-1}} = 1/\dot{\mu}(\epsilon) \leq C$, for small ϵ.[4] The main term of the sample complexity becomes

$$O\left(\sum_{k=1}^{\log_2(1/\Delta_{\min})} 2^{2k} \log \frac{1}{\delta} \right) \leq O\left(\frac{1}{\epsilon^2} \log \frac{1}{\delta} \right).$$

Hence ignoring burn-in or the additional samples we take in each round to guarantee the confidence interval, the total sample complexity would be $\tilde{O}(\frac{1}{\epsilon^2})$. This is exponentially smaller than in Example 1 and demonstrates the power of an informative arm in reducing the sample complexity.

On the other hand, the burn-in phase, common to all logis-

[4] With $H(\theta^*)^{-1}$ interpreted as a pseudo-inverse.

tic bandit algorithms based on the MLE, may potentially take a number of samples exponential in r. This example demonstrates the need for further work on understanding the precise dependence of κ in pure exploration. In Section 4, we take a first step towards this by showing κ^{-1} burn-in is unavoidable in some cases.

Comparison to past work. As β_k grows exponentially each round, the first element in the maximum for β_k dominates our sample complexity. Focusing on this term while ignoring logarithmic terms and the burn-in samples, the sample complexity in Theorem 2 scales as

$$\rho* := \sum_{k=1}^{\log_2(2/\Delta_{\min})} 2^{2k} \min_{\lambda \in \Delta_{\mathcal{X}}} \max_{z,z' \in \mathcal{S}_k} \|z - z'\|^2_{H(\theta^*)} .$$

Importantly, this depends on $H(\theta^*)$ instead of a loose bound based on κ^{-1}. In Appendix E.1 we show

$$\rho* \leq \log\left(\frac{2}{\Delta_{\min}}\right) \min_{\lambda \in \Delta_{\mathcal{X}}} \max_{z \in \mathcal{Z} \setminus z^*} \frac{\|z^* - z\|^2_{H(\lambda,\theta^*)^{-1}}}{\langle \theta^*, z^* - z \rangle^2}.$$

This is reminiscent of a similar quantity that is within a $\log(\cdot)$ factor of being optimal for pure exploration linear bandits (Fiez et al., 2019). We provide a close connection between our upper bound and information theoretic lower bounds in Appendix E.1, although they do not match exactly. We also prove a novel lower bound in moderate confidence regimes, which we elaborate more in Section 4.

We now compare to the result of Kazerouni & Wein (2019). Using a variant of the UGapE algorithm for linear bandits (Xu et al., 2018), they demonstrate a sample complexity $\tilde{O}(\frac{d|\mathcal{X}|}{\kappa^2 \Delta^2_{\min}})$ in the setting where $\mathcal{X} = \mathcal{Z}$. This sample complexity, unlike ours, scales linearly with the number of arms, and only captures a dependency on the smallest gap. We note that one can improve on their sample complexity by using a naive passive algorithm that uses a fixed G-optimal design, along with the trivial bound $H(\lambda, \theta^*) \geq \kappa_0 A(\lambda)$, resulting in $\tilde{O}(d/(\kappa_0 \Delta^2_{\min}))$ (Soare et al., 2014).[5] In contrast, the bound of Theorem 2 only depends on the number of arms logarithmically, captures a local dependence on θ^*, and has a better gap dependence.

Extra samples. Algorithm 1 potentially samples in each round to ensure the confidence interval is valid (i.e., the first argument of the max in line 5). In Appendix D, we propose RAGE-GLM-2 that removes these samples needed in each round (but not the burn-in samples) by employing the confidence interval of Faury et al. (2020). This algorithm has a better asymptotic behavior as $\delta \to 0$, but could perform worse with large d or S_* due to an additional factor of $\sqrt{d}$ and a factor of S^3_*.

[5]This is equivalent to computing the allocation from Algorithm 2, and sampling until all arms are eliminated.

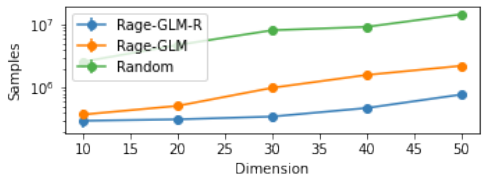

Figure 1. Standard Baseline Example

	open	pointy	sporty	comfort
$\lvert\mathcal{Z}\rvert$	3327	2932	3219	3374
RAGE-GLM-R	**9.17e+07**	**2.38e+05**	**2.29e+05**	**3.34e+06**
RAGE-GLM	**9.17e+07**	**2.38e+05**	**2.29e+05**	4.55e+06
Passive	2.69e+08	**2.38e+05**	**2.29e+05**	8.54e+06

Table 1. Zappos pairwise comparison data, bold indicates a win.

3.2. Experiments

This section evaluates the empirical performance of **RAGE-GLM**, alongside two additional algorithms:

- **RAGE-GLM-R**: This is a heuristic version of RAGE-GLM that makes two changes. First, it does not do a burn-in in each round and samples from $\lambda_k = \min_{z,z'} \|z - z'\|^2_{H(\lambda,\hat{\theta}_{k-1})^{-1}}$ directly. Second, to compute the estimate $\hat{\theta}_k$, it uses all samples up to round k.
- **Passive Baseline**: This baseline runs the burn-in procedure and then computes the static design $\lambda = \min_{z,z' \in \mathcal{Z}} \|z - z'\|^2_{H(\lambda,\hat{\theta}_0)^{-1}}$. It then proceeds in rounds, drawing 2^k samples in round k, terminating when we are able to verify that each arm is sub-optimal using the fixed design confidence interval (see Appendix H for details). As in the heuristic, we recycle samples over rounds.

Remark. We also implemented the algorithm of (Kazerouni & Wein, 2019). However a) the algorithm was extremely slow to run since an MLE and a convex optimization had to be run each round, b) the confidence bounds do not exploit the true variance. As a result, the algorithm did not terminate on any of our examples.

Our first experiment (Fig. 1) is a common baseline in the linear bandits literature. We consider $d + 1$ arms in d dimensions with arm $i \in [1, n]$ being the i-th basis vector, and arm $i + 1$ as $\cos(.1)e_1 + \sin(.1)e_2$. We use $d \in \{10, 20, 30, 40, 50\}$ and 10 repetitions for each value of d. In all instances, all algorithms found the best arm correctly. RAGE-GLM was competitive against the heuristic RAGE-GLM-R and took roughly a factor of 10 less samples compared to Random.

Our next experiment is based on the Zappos pairwise comparison dataset (Yu & Grauman, 2014; 2017). This dataset consists of pairwise comparisons on 50k images of shoes and 960 dimensional vision-based feature vectors for each shoe. Given a pair of shoes, participants were asked to compare them on the the attributes of "open", "pointy", "sporty" and "comfort" obtaining several thousand queries. For each

one of these categories, we fit a logistic model to the set of pairwise comparisons after PCA-ing the features down to 25 dimensions (for computational tractability) and used the underlying weights as θ^*. We then set $\mathcal{Z}$ to be the set of shoes that were considered in that category and $\mathcal{X}$ to be 5000 random pairs. Table 1 shows the result. For the "open" and "comfort" category, RAGE-GLM took about a factor of 3 less samples compared to Random. The large sample complexity for "open" is due to an extremely small minimum gap of $O(10^{-4})$. For "pointy" and "sporty" the empirical gaps were large and all three algorithms terminated after the burn-in phase. Finally, κ for all instances was on the order of 0.1. See Appendix H for a deeper discussion, alongside pictures of winning and informative shoes.

4. $1/\kappa_0$ Is Necessary

In this section, we turn to lower bounds on the sample complexity of pure exploration problems.

Theorem 3. *Fix $\delta_1 < 1/16$, $d \geq 4$, and $\epsilon \in (0, 1/2]$ such that $d\varepsilon^2 \geq 12.2$. Let $\mathcal{Z}$ denote the action set and Θ denote a family of possible parameter vectors. There exists instances satisfying the following properties simultaneously*

1. *$|\mathcal{Z}| = |\Theta| = e^{\epsilon^2 d/4}$ and $\|z\| = 1$ for all $z \in \mathcal{Z}$.*
2. *$S = \|\theta_*\| = O(\epsilon^2 d)$*
3. *Any algorithm that succeeds with probability at least $1 - \delta_1$ satisfies*

$$\exists \theta \in \Theta \text{ s.t. } \mathbb{E}_\theta[T_{\delta_1}] > \Omega\left(e^{\epsilon^2 d/4}\right) = c\left(\frac{1}{\kappa_0}\right)^{\frac{1-\epsilon}{1+3\epsilon}}$$

where T_{δ_1} is the random variable of the number of samples drawn by an algorithm and c is an absolute constant.

The implications of this bound are two-fold. Firstly, it shows a family of instances where the dependence on $1/\kappa_0$ in the sample complexity of Algorithm 1 is necessary. Secondly, this bound demonstrates a particular phenomenon of the logistic bandit problem: there are settings where $\kappa_0^{-1} \approx e^d$ samples are needed despite $\theta^* \in \mathbb{R}^d$. By contrast, for linear bandits, $O(d/\Delta_{\min}^2)$ samples are always sufficient (Soare et al., 2014). For the instances in the theorem, $\Delta_{\min} \geq \Omega(1 - e^{-d})$. In Appendix E.2, we state a second lower bound that captures the asymptotic sample complexity as $\delta \to 0$, but show that this bound would suggest that only $O(\log(1/\delta))$ samples are necessary, which is vacuous for $\delta > e^{-e^d}$. Instead, the above *moderate confidence* bound reflects the true sample complexity of the problem for values of δ seen in practice, e.g. $\delta \approx .05$. This dichotomy highlights that there are important challenges to logistic bandit problems that are not captured by their asymptotic sample complexity. In particular, this demonstrates that there exist instances of pure exploration logistic bandits that are *exponentially harder* than their linear counterparts. The proof is in Appendix E.2, inspired by a construction from Dong et al. (2019).

5. K-Armed Contextual Bandits

We now switch gears and consider the contextual bandit setting where at each time step t the environment presents the learner with an arm set $\mathcal{X}_t = \{x_{t,1}, \ldots, x_{t,K}\} \subset \mathbb{R}^d$ independently of the learner's history (Auer, 2002). The learner then chooses an arm index $a_t \in [K]$ and receives a reward $y_t \sim \mathsf{Bernoulli}(\mu(x_{t,a_t}^\top \theta^*))$, where parameter θ^* is unknown to the learner. Let $x_{t,a^*} = \arg\max_{x \in \mathcal{X}_t} \mu(x_{t,a}^\top \theta^*)$ be the best arm at time step t. The goal is to minimize the cumulative (pseudo-)regret over the time horizon T:

$$\mathsf{Reg}_T = \sum_{t=1}^{T} \mu(x_{t,a^*}^\top \theta^*) - \mu(x_{t,a_t}^\top \theta^*). \tag{7}$$

While the regret $\tilde{O}(d\sqrt{T} + d^2\kappa^{-1})$ is achievable by Faury et al. (2020)[6], one can aim to achieve an accelerated regret bound when $K = o(e^d)$. Specifically, Li et al. (2017) achieve the best-known bound of $\tilde{O}(\frac{1}{\kappa}\sqrt{dT\log(K)})$. However, the factor $1/\kappa$ is exponential w.r.t. $\|\theta^*\|$, which makes the regret impractically large. Leveraging our new confidence bound, we propose a new algorithm SupLogistic that removes $1/\kappa$ from the leading term: $\tilde{O}(\sqrt{dT\log(K)})$.

We assume that $\|x_{t,a}\| \leq 1, \forall t \in [T], a \in [K]$, and that $\|\theta^*\| \leq S_*$ where S_* is known to the learner. We follow Li et al. (2017) and assume that there exists σ_0^2 such that $\lambda_{\min}(\mathbb{E}[\frac{1}{K}\sum_{a\in[K]} x_{t,a}x_{t,a}^\top]) \geq \sigma_0^2$, which is used to characterize the length of the burn-in period in our theorem.

We describe SupLogistic in Alg. 3, which follows the standard mechanism for maintaining independent samples (Auer, 2002). As the confidence bound is not available until enough samples are accrued, we perform τ time steps of burn-in sampling and then spread the samples across the buckets $\Psi_1, \ldots, \Psi_S, \Phi$ equally. Note that our burn-in sampling is different from Li et al. (2017), we show in Appendix F that their approach is problematic.

After the burn-in, in each time step t, we loop through the buckets $s \in [S]$. In each loop, we compute $\theta_{t-1}^{(s)}$, the MLE given in Eq (1), using the samples in the bucket $\Psi_s(t-1)$. We compute θ_Φ in the same way using Φ. Let $X_t = x_{t,a_t}$. For any θ, define

$$H_t^{(s)}(\theta) := \sum_{u \in \Psi_s(t)} \dot{\mu}(X_u^\top \theta) X_u X_u^\top . \tag{8}$$

The algorithm computes the mean estimate and the confidence width of each arm $a \in [K]$ as follows:

$$m_{t,a}^{(s)} := \langle x_{t,a}, \theta_{t-1}^{(s)} \rangle, w_{t,a}^{(s)} := \alpha\sqrt{2.2}\|x_{t,a}\|_{H_{t-1}^{(s)}(\theta_\Phi)^{-1}}. \tag{9}$$

[6] $\tilde{O}$ hides poly-logarithmic factors in T.

Algorithm 3 SupLogistic

Input: time horizon T, burn-in length τ, and exploration rate α

initialize $S = \lfloor \log_2 T \rfloor$
initialize Buckets $\Psi_1 = \cdots = \Psi_S = \Psi_{S+1} = \varnothing$
for $t \in [\tau]$ **do**
  Choose $a_t \in [K]$ uniformly at random.
  Add a_t to the set $\Psi_{((t-1) \bmod S)+1}$.
initialize $\Psi_0 = \emptyset$, $\Phi = \Psi_{S+1}$
for $t = \tau+1, \tau+2, \cdots, T$ **do**
  initialize $A_1 = [K]$, $s = 1$, $a_t = \varnothing$.
  while $a_t = \varnothing$ **do**
    Compute $m_{t,a}$ and $w_{t,a}$ ▷ use Eq (9)
    if $w_{t,a}^{(s)} > 2^{-s}$ for some $a \in A_s$ **then**
      $a_t = a$ ▷ **(a)**
      $\Psi_s \leftarrow \Psi_s \cup \{t\}$
    else if $w_{t,a}^{(s)} \le 1/\sqrt{T}$, $\forall\, a \in A_s$ **then**
      $a_t = \arg\max_{a \in A_s} m_{t,a}^{(s)}$ ▷ **(b)**
      $\Psi_0 \leftarrow \Psi_0 \cup \{t\}$
    else if $w_{t,a}^{(s)} \le 2^{-s}$, $\forall\, a \in A_s$ **then**
      $A_{s+1} = \{a \in A_s :$ ▷ **(c)**
        $m_{t,a}^{(s)} \ge \max_{j \in A_s} m_{t,j}^{(s)} - 2 \cdot 2^{-s}\}$
      $s \leftarrow s+1$

For each $s \in [S]$, we check if there is an underexplored arm (step **(a)**) and pull it. Otherwise, we pull the arm with the highest empirical mean. Finally, we filter arms whose empirical means are sufficiently far from the highest empirical mean and go to the next iteration.

The bucketing is important to maintain the validity of the concentration inequality in the analysis, which requires that the data satisfies the fixed design assumption in Theorem 1. Our comment on Li et al. (2017) in Appendix F elaborates more on this issue.

The main challenge of the design of SupLogistic over SupCB-GLM (Li et al., 2017) is how we use our tight confidence bound, which requires the confidence width to depend on θ^* (see Theorem 1). Our solution is to use a separate bucket Φ dedicated for computing the width. If we do not use Φ and use the empirical version of Theorem 1, we would break the fixed design assumption as we collect samples as a function of the rewards from the same bucket.

We present our regret bound in the following theorem whose proof can be found in Appendix G.

Theorem 4. *Let* $T \ge d$, $\tau = \sqrt{dT}$, $\alpha = 2.4\sqrt{\log(\frac{2(2+\tau)\cdot 2STK}{\delta})}$ *with* $\delta = 1/T$, *and* $Z = \frac{1}{\sigma_0^4}\left(\frac{1}{\sigma_0^4} + \kappa^{-2}\right)\left(d + \frac{1}{d}\log^2(K)\right)$. *Then,*

$$\mathbb{E}[\mathsf{Reg}_T] \le 10\alpha\sqrt{dT\log(T/d)\log_2(T)} + O\left(\frac{\alpha^2 d}{\kappa}\log^2(T) + Z\log^4(Z)\right).$$

Our bound improves upon SupCB-GLM (Li et al., 2017) by removing the factor $1/\kappa = \Theta(\exp(S_*))$ in the leading term (i.e., $\sqrt{T}$ term). Such an improvement parallels that of Faury et al. (2020) with $\tilde{O}(d\sqrt{T})$ upon UCB-GLM (Li et al., 2017) with $\tilde{O}(\frac{1}{\kappa}d\sqrt{T})$. We remark that the constant σ_0^2 is at most $1/d$, and thus in the best case the lower-order term of the regret bound scales like $d^5 + d^3\kappa^{-2}$ ignoring K.

Compared to Logistic-UCB-2 of Faury et al. (2020), our regret bound can be better or worse depending on the problem, which we summarize in three cases. First, ours has a factor of $\sqrt{d\log(K)}$ in the $\sqrt{T}$ term, which is a $\sqrt{d}$ factor better than theirs when $K = o(e^d)$. Secondly, our bound's lower order term scales like $1/\kappa^2$, which is worse than $1/\kappa$ of Logistic-UCB-2. Thirdly, while Logistic-UCB-2 manages to avoid an exponential dependence on S_* in the leading term, the regret still has a factor $S_*^{1.5}$ and requires the knowledge of S_*.[7] In contrast, our bound does not depend on S_* in the leading term nor requires the knowledge of S_*.

Remark 2. *A parallel work by Abeille et al. (2020) achieves a leading term of* $d\sqrt{\dot{\mu}(x_*^\top\theta^*)T}$ *in the regret bound where* x_* *is the best arm that is fixed throughout. This is possible since their setting assumes a fixed arm set. In contrast, our setting assumes that the arm set is changing, so the best arm can change in every time step. For this reason, we do not expect to achieve a factor like* $\sqrt{\dot{\mu}(x_*^\top\theta^*)}$ *in the leading term without further assumptions.*

6. Future work

Our confidence bound utilizes self-concordance and local analysis to significantly improve upon the existing state of the art results for the logistic MLE. We remove a direct dependence on κ^{-1} in the confidence width and relax significantly the requirement on the minimum sample size for the bound to be valid. To better leverage our knowledge burn-in condition, we hope to develop online procedures that adapt to θ^* instead of paying a worst case dependence in κ to satisfy the burn-in condition. Furthermore, understanding the optimal burn-in condition is an important open problem with practical implications.

Pure exploration for linear logistic models is largely underexplored, although its applications are abundant. Exploiting the local nature of the logistic loss and closely working with non-uniform variances that naturally arise from the model is crucial in sample-efficient design of experiments. Our work is an important first step on understanding the true sample complexity of this problem and determining the precise dependence of the sample complexity on κ_0^{-1} is an

[7] Which may be removed if their algorithm uses a burn-in phase.

exciting direction.

A major road block to developing practical contextual bandit algorithms is the fact that SupLogistic (and its ancestors like in Auer (2002)) have to maintain independent buckets, and cannot share samples across buckets. It would be interesting to develop new algorithms that do not waste samples, without increasing the regret bound. Foster & Rakhlin (2020) have proposed such an algorithm but its dependence on the number of arms is sub-optimal.

References

Abeille, M., Faury, L., and Calauzènes, C. Instance-Wise Minimax-Optimal Algorithms for Logistic Bandits, 2020.

Antos, A., Grover, V., and Szepesvári, C. Active learning in heteroscedastic noise. *Theoretical Computer Science*, 411(29-30):2712–2728, 2010.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *The Journal of Machine Learning Research*, 3:397–422, 2002.

Biswas, A., Pham, T. T., Vogelsong, M., Snyder, B., and Nassif, H. Seeker: Real-time interactive search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2867–2875, 2019.

Chu, W., Li, L., Reyzin, L., and Schapire, R. E. Contextual Bandits with Linear Payoff Functions. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 15, pp. 208–214, 2011.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic Linear Optimization under Bandit Feedback. In *Proceedings of the Conference on Learning Theory (COLT)*, pp. 355–366, 2008.

Degenne, R., Ménard, P., Shang, X., and Valko, M. Gamification of pure exploration for linear bandits. In *ICML 2020: 37th International Conference on Machine Learning*, 2020.

Dong, S., Ma, T., and Van Roy, B. On the performance of thompson sampling on logistic bandits. *arXiv preprint arXiv:1905.04654*, 2019.

Faury, L., Abeille, M., Calauzènes, C., and Fercoq, O. Improved optimistic algorithms for logistic bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 3052–3060, 2020.

Fiez, T., Jain, L., Jamieson, K. G., and Ratliff, L. Sequential experimental design for transductive linear bandits. In *Advances in Neural Information Processing Systems*, pp. 10667–10677, 2019.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pp. 586–594, 2010.

Foster, D. J. and Rakhlin, A. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

Geng, S., Nassif, H., Manzanares, C., Reppen, M., and Sircar, R. Deep pqr: Solving inverse reinforcement learning using anchor actions. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 3431–3441, 2020.

Jain, L., Jamieson, K. G., and Nowak, R. Finite sample prediction and recovery bounds for ordinal embedding. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pp. 2711–2719, Barcelona, Spain, 2016.

Jun, K.-S., Bhargava, A., Nowak, R., and Willett, R. Scalable generalized linear bandits: Online computation and hashing. In *Advances in Neural Information Processing Systems*, pp. 99–109, 2017.

Karnin, Z., Koren, T., and Somekh, O. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pp. 1238–1246, 2013.

Katz-Samuels, J., Jain, L., Karnin, Z., and Jamieson, K. An empirical process approach to the union bound: Practical algorithms for combinatorial and linear bandits. *arXiv preprint arXiv:2006.11685*, 2020.

Kaufmann, E., Cappé, O., and Garivier, A. On the complexity of a/b testing, 2015.

Kaufmann, E., Cappé, O., and Garivier, A. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17 (1):1–42, 2016.

Kazerouni, A. and Wein, L. M. Best arm identification in generalized linear bandits. *arXiv preprint arXiv:1905.08224*, 2019.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Lehmann, E. L. and Casella, G. *Theory of point estimation*. Springer Science & Business Media, 2006.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A Contextual-Bandit Approach to Personalized News Article Recommendation. *Proceedings of the International Conference on World Wide Web (WWW)*, pp. 661–670, 2010.

Li, L., Lu, Y., and Zhou, D. Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 2071–2080, 2017.

Li, Y., Wang, Y., and Zhou, Y. Nearly Minimax-Optimal Regret for Linearly Parameterized Bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, volume 99, pp. 2173–2174, 2019.

Nabi, S., Nassif, H., Hong, J., Mamani, H., and Imbens, G. Bayesian meta-prior learning using empirical bayes. *Management Science*, pp. forthcoming, 2021.

Plan, Y. and Vershynin, R. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59 (1):482–494, 2012.

Pollard, D. Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*. JSTOR, 1990.

Pukelsheim, F. *Optimal design of experiments*. SIAM, 2006.

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

Sawant, N., Namballa, C. B., Sadagopan, N., and Nassif, H. Contextual multi-armed bandits for causal marketing. In *Proceedings of the International Conference on Machine Learning (ICML'18) Workshops*, Stockholm, Sweden, 2018.

Soare, M., Lazaric, A., and Munos, R. Best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems*, pp. 828–836, 2014.

Teo, C. H., Nassif, H., Hill, D., Srinivasan, S., Goodman, M., Mohan, V., and Vishwanathan, S. Adaptive, personalized diversity for visual discovery. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys)*, pp. 35–38, Boston, MA, 2016.

Thompson, W. R. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25(3/4):285, 1933.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Xu, L., Honda, J., and Sugiyama, M. A fully adaptive algorithm for pure exploration in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 843–851. PMLR, 2018.

Yu, A. and Grauman, K. Fine-grained visual comparisons with local learning. In *Computer Vision and Pattern Recognition (CVPR)*, Jun 2014.

Yu, A. and Grauman, K. Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *International Conference on Computer Vision (ICCV)*, Oct 2017.

Yue, Y., Broder, J., Kleinberg, R., and Joachims, T. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78:1538–1556, 2012.