An Experimental Design Approach for Regret Minimization in Logistic Bandits

Blake Mason[1], Kwang-Sung Jun[2], and Lalit Jain[3]

[1]Rice University
[2]University of Arizona
[3]University of Washington
bm63@rice.edu, kjun@cs.arizona.edu, lalitj@wu.edu

Abstract

In this work we consider the problem of regret minimization for logistic bandits. The main challenge of logistic bandits is reducing the dependence on a potentially large problem dependent constant κ that can at worst scale exponentially with the norm of the unknown parameter θ_*. Prior works have applied self-concordance of the logistic function to remove this worst-case dependence providing regret guarantees like $O(d\log^2(\kappa)\sqrt{\dot{\mu}T}\log(|\mathcal{X}|))$ where d is the dimensionality, T is the time horizon, and $\dot{\mu}$ is the variance of the best-arm. This work improves upon this bound in the fixed arm setting by employing an experimental design procedure that achieves a minimax regret of $O(\sqrt{d\dot{\mu}T\log(|\mathcal{X}|)})$. Our regret bound in fact takes a tighter instance (i.e., gap) dependent regret bound for the first time in logistic bandits. We also propose a new warmup sampling algorithm that can dramatically reduce the lower order term in the regret in general and prove that it can replace the lower order term dependency on κ to $\log^2(\kappa)$ for some instances. Finally, we discuss the impact of the bias of the MLE on the logistic bandit problem, providing an example where d^2 lower order regret (cf., it is d for linear bandits) may not be improved as long as the MLE is used and how bias-corrected estimators may be used to make it closer to d.

1 Introduction

Linear bandits, which have gained popularity since their success in online news recommendation (Li et al. 2010), solve sequential decision problems under limited feedback when each action (or arm) to be taken has a known feature vector deemed to predict the reward. Specifically, at each time step t, the learner chooses an arm $x_t \in \mathbb{R}^d$ from an available pool of arms $\mathcal{X}$, and then receives a reward $y_t = x_t^\top \theta_* + \eta_t$, where θ_* is unknown and η is usually assumed to be zero-mean subGaussian noise. The goal of the learner is to maximize the total cumulative rewards over the time horizon T by judiciously balancing between efficiently learning θ_* (exploration) and using the learned knowledge on θ_* to accumulate large rewards (exploitation). Since the pioneering studies by Abe and Long (1999) and Auer (2002), there have been significant developments in both theory (Dani, Hayes, and Kakade 2008; Abbasi-Yadkori, Pal, and Szepesvari 2011; Foster and Rakhlin 2020) and applications (Li et al. 2010; Sawant et al. 2018; Teo et al. 2016).

Many real-world applications, however, have binary rewards and are not captured by the additive noise setting. For example, the seminal work by Li et al. (2010) for contextual bandits considers a binary reward of click/no-click, yet they apply bandit algorithms based on linear models – this is comparable to applying linear regressions to a binary classification task. For binary rewards, the logistic linear model is natural when rewards are assumed to follow $y_t \sim \text{Bernoulli}(\mu(x_t^\top \theta_*))$ where $\mu(z) = 1/(1+\exp(-z))$ is the logistic function. While the link function μ can be changed, the logistic function is worth more attention for two reasons: (i) it is extensively used in practice, even in state-of-the-art deep architectures whose last layer is the negative log likelihood of the logistic linear model, and (ii) if we were to use a trained network to compute the input to the last layer and take it as the feature vector for bandit tasks, the features are likely optimized to excel with the logistic model.

The first work on logistic bandits due to Filippi et al. (2010) showed a regret bound of $\tilde{O}(d\kappa\sqrt{T})$ where $\kappa = \max_{\|x\|_2\le 1}\dot{\mu}(x^\top\theta_*)^{-1}$ and $\tilde{O}$ ignores polylogarithmic factors for all variables except for κ. Since κ can be exponential in $\|\theta_*\|_2$, the key challenge in developing bandit algorithms in the logistic setting both theoretically and practically is to overcome this worst-case scaling. In the last few years, there has been a flurry of activity on this problem that exploits the *self-concordance* of the logistic loss with the seminal work of Faury et al. (2020). Recently, Abeille, Faury, and Calauzènes (2021) proposed a UCB style algorithm called `OFULog`, establishing a regret bound with the leading term of $\tilde{O}(d\log^2(\kappa)\sqrt{\dot{\mu}(x_*^\top\theta_*)T}+\kappa d^2\wedge(d^2+|\mathcal{X}_-|))$ where $\mathcal{X}_- \subset \mathcal{X}$ is a set of *detrimental arms*. In the finite armed contextual bandit setting, Jun et al. (2021) propose an improved fixed design confidence interval and adapted a SupLinRel style algorithm (Auer, Cesa-Bianchi, and Fischer 2002) called `SupLogistic` to establish a regret scaling like $\tilde{O}(\sqrt{dT})$. `SupLogistic` achieves a better dependence on d and κ. However, it has a worse dependence with $\dot{\mu}(x_*^\top\theta_*)$ due to the changing arm set setting and makes a strong assumptions of stochastic contexts and bounded minimum eigenvalues. The regret bound of `OFULog` is free of these assumptions, but the leading term is suboptimal. We discuss key related work throughout the paper and postpone detailed reviews to our supplementary.

Motivated by the gaps in the regret bounds, we make the following contributions.

Improved Logistic Bandit Algorithm (Section 2): We take an experimental design approach to propose a new bandit algorithm called HOMER (H-Optimal MEthod for Regret) that achieves the best of the two state-of-the-art algorithms above: $\tilde{O}(\sqrt{d\dot{\mu}(x_*^\top\theta_*)T\log(|\mathcal{X}|)}+d^2\kappa)$ in the fixed-arm setting where the lower order term matches the state-of-the-art OFULog in the worst case. In fact, we prove an even tighter instance-dependent (i.e., gap-dependent) regret bound of $O(\frac{d\dot{\mu}(x_*^\top\theta_*)\log(|\mathcal{X}|T)}{\Delta}+d^2\kappa)$ for the first time to our knowledge where Δ is the reward gap between the best arm among the suboptimal arms and the overall best arm.

Novel Warmup Algorithm (Section 3): While HOMER achieves the best worst-case regret guarantee, it must be equipped with a warmup sampling procedure. Using a naive sampling procedure, HOMER will incur $d^2\kappa$ regret during the warmup. This stems from having to use fixed design confidence bounds (Jun et al. 2021, Theorem 1)– without them there is no known ways to achieve the factor $\sqrt{d}$ in the leading term of the regret bound when there are finitely many arms – that require the observed arms and their rewards to satisfy so-called "warmup" condition (see (2) below). In order to improve its practical performance, we propose a novel adaptive warmup algorithm called WAR (Warmup by Accepts and Rejects), which performs the warmup with much fewer samples than κd^2 in general. We prove its correctness guarantee, show that for 1d it never spends more samples than the naive warmup, and present an arm-set dependent optimality guarantee.

Conjectures on the Dimension Dependence of the Fixed Design Inequalities (Section 4): Omitting the κ dependence and logarithmic factors in this paragraph, all existing regret bounds for logistic bandits have an d^2 dependence in the lower order term, which is in stark contrast to an d dependence in linear bandits (Auer 2002). The consequence is that logistic bandit algorithms suffer a linear regret until $T = d^2$ in the worst case. Does this mean logistic bandits are fundamentally more difficult than linear bandits in high dimensional settings? While we do not have a complete answer yet, we provide a sketch of an argument that, when the MLE is used, such a d^2 dependence might be unimprovable. The argument starts from the classical fact that the MLE of generalized linear models (GLMs) are biased, e.g., (Bartlett 1953). Based on this fact, we observe that in order to obtain tight fixed design confidence bounds we need to perform oversampling of arms as a function of d until the squared bias gets smaller than the variance. Furthermore, based on the known fact that in 1d the KT estimator (Krichevsky and Trofimov 1981) is much less biased than the MLE (Cox and Snell 2018), which we verify numerically as well, we propose a new estimator and conjecture that it may lead to the lower order regret term of $O(d^{4/3})$.

2 Regret Minimization for Logistic Bandits

Let us formally define the problem. We assume access to a finite and fixed set of arms $\mathcal{X} \subset \{x \in \mathbb{R}^d : \|x\|_2 \le 1\}$. At each time $t \ge 1$ the learner chooses an arm $x_t \in \mathcal{X}$ and observes a Bernoulli reward $y_t \in \{0, 1\}$ with

$$\mathbb{E}[y_t|x_t] = \mu(x_t^\top\theta_*) \; .$$

Let $x_* = \arg\max_{x\in\mathcal{X}} x^\top\theta_*$. For ease of notation, we assume that x_* is unique though this condition can be relaxed. The goal of the learner is to minimize the *cumulative (pseudo-)regret* up to time T: $R_T := \sum_{t=1}^T \mu(x_*^\top\theta_*) - \mu(x_t^\top\theta_*)$.

Notations Let $\mathcal{F}_t$ be the sigma algebra generated by the set of rewards and actions of the learner up to time t, i.e., $\sigma(x_1, y_1, \dots, x_{t-1}, y_{t-1})$. We assume that the learner has knowledge of an upper bound S on $\|\theta_*\|$. Define $\kappa := \max_{\|x\|\le 1} \dot{\mu}(x^\top\theta_*)^{-1}$ and $\kappa_0 := \max_{x\in\mathcal{X}} \dot{\mu}(x^\top\theta_*)^{-1}$, the inverse of the smallest derivative of the link function among elements of $\mathcal{X}$. Denote by $\triangle_{\mathcal{A}}$ the set of probability distributions over the set $\mathcal{A}$. Let $\mathsf{Supp}(\lambda)$ with $\lambda \in \triangle_{\mathcal{A}}$ be the subset of $\mathcal{A}$ for which λ assigns a nonzero probability. We use $A \lesssim B$ to denote A is bounded by B up to absolute constant factors.

Logistic Regression We review logistic regression. Assume that we have chosen measurements $x_1, \dots, x_t$ to obtain rewards $y_1, \dots, y_t$. The *maximum likelihood estimate* (MLE), $\hat{\theta}$ of θ_* is given by,

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

The Fisher information of the MLE estimator is $H_t(\theta_*) := \sum_{s=1}^t \dot{\mu}(x_s^\top\theta_*)x_s x_s^\top$. Obtaining (near-)optimal regret hinges on the availability of tight confidence bounds on the *means* of each arm. For technical reasons, tight confidence bounds (i.e., without extra factors such as $\sqrt{d}$ in the confidence bound like (Faury et al. 2020)) require the data to observe the *fixed design* setting: for each $s \in [t]$, y_s is conditionally independent of $\{x_s\}_{s=1}^t \setminus \{x_s\}$ given x_s. Recent work by Jun et al. (2021) provide a tight finite-time fixed design confidence interval on the natural parameter of $x^\top\theta_*$ for any $x \in \mathcal{X}$. For regret minimization, we instead require estimates of the mean parameter $\mu(x^\top\theta_*)$ each arm $x \in \mathcal{X}$. Define $\gamma(d) := \max\{d + \log(6(2+t_{\mathsf{eff}})/\delta), 6.1^2\log(6(2+t_{\mathsf{eff}})/\delta)\}$ where t_{eff} is the number of distinct vectors in $\{x_s\}_{s=1}^t$. We refer to the following assumption on our samples as the *warmup condition*:

$$\xi_t^2 := \max_{1\le s\le t} \|x_s\|^2_{H_t(\theta_*)^{-1}} \le \frac{1}{\gamma(d)} \; . \tag{2}$$

Lemma 1. *Fix $\delta \le e^{-1}$. Let $\hat{\theta}_t$ denote the MLE estimate for a fixed design $\{x_1, \dots, x_t\} \subset \mathcal{X}$. Under the warmup condition* (2)*, with probability $1-\delta$, we have, $\forall x \in \mathcal{X}$,*

$$\begin{aligned} &|x^\top(\hat{\theta}_t - \theta_*)| \le 1 \text{ and } |\mu(x^\top\hat{\theta}_t) - \mu(x^\top\theta_*)| \\ &\quad \le 4.8\dot{\mu}(x^\top\theta_*)\|x\|_{H_t(\theta_*)^{-1}}\sqrt{\log(2(2+t_{\mathit{eff}})|\mathcal{X}|/\delta)}. \end{aligned} \tag{3}$$

The proof relies on Theorem 1 of (Jun et al. 2021) and the (generalized) self-concordance of the logistic loss (Faury et al. (2020, Lemma 9)). While similar results were used in the proof of SupLogistic in Jun et al. (2021), for our bandit algorithm below it is crucial to guarantee $\forall x \in \mathcal{X}, |x^\top(\hat{\theta}_t - \theta_*)| \le 1$ as well. As far as we know, this is the first non-trivial confidence bound on the mean parameter for logistic models. To contextualize this result, consider that via the delta-method, $\sqrt{t}(\mu(x^\top\hat{\theta}) - \mu(x^\top\theta_*)) \xrightarrow{D} \mathcal{N}\left(0, \dot{\mu}(x^\top\theta_*)^2\|x\|^2_{H_t(\theta_*)^{-1}}\right)$. Hence, Lemma 1 guarantees an asymptotically tight normal-type tail bound up to constant factors provided the warmup condition holds.

Experimental Design We leverage our tight confidence widths on the means $\mu(x^\top\theta_*)$ given in Lemma 1 to develop a novel logistic bandit algorithm. Motivated by the form of the confidence width on the mean parameter, we consider the following experimental design problem:

$$h^* := \min_{\lambda\in\triangle_{\mathcal{X}}} \max_{x\in\mathcal{X}} \dot{\mu}(x^\top\theta_*)^2\|x\|^2_{H_\lambda(\theta_*)^{-1}}$$
$$\text{where} \quad H_\lambda(\theta) = \sum_{x\in\mathcal{X}} \lambda_x\dot{\mu}(x^\top\theta)xx^\top ,$$

which we refer to as an *H-optimal design*. That is, we want to find an allocation of samples over $\mathcal{X}$ minimizing the worst case confidence-width of Lemma 1 for any $x \in \mathcal{X}$. Note that h^* depends on the arm set $\mathcal{X}$, though we omit it for brevity. This experimental design is closely linked to the G-optimal design objective for the MLE in an exponential family. Indeed, in our setting, the G-optimal design is

$$g^* := \min_{\lambda\in\triangle_{\mathcal{X}}} \max_{x\in\mathcal{X}} \|x\|^2_{H_\lambda(\theta_*)^{-1}} \tag{4}$$

We point out that in the setting of linear bandits (i.e. the linear GLM), the celebrated Kiefer-Wolfowitz theorem states that the optimal value of this objective is just d for any choice of $\mathcal{X}$ (Kiefer and Wolfowitz 1960). Hence, for any $\theta \in \mathbb{R}^d$, letting $\mathcal{Y} = \{\sqrt{\dot{\mu}(x^\top\theta)}x, x \in \mathcal{X}\}$ we see that

$$h^* = \min_{\lambda\in\triangle_{\mathcal{X}}} \max_{x\in\mathcal{X}} \dot{\mu}(x^\top\theta)^2\|x\|^2_{H(\theta_*)^{-1}}$$
$$\le \frac{1}{4} \min_{\lambda\in\triangle_{\mathcal{Y}}} \max_{y\in\mathcal{Y}} \|y\|^2_{(\sum_{y\in\mathcal{Y}}\lambda_y yy^\top)^{-1}} \le \frac{d}{4}$$

where we use $\dot{\mu}(z) \le 1/4$ for the first inequality and the Kiefer-Wolfowitz theorem for the last inequality. Since $\dot{\mu}(x^\top\theta)$ decays exponentially fast in $|x^\top\theta|$, this bound is overly pessimistic and in practice the values of the H-optimal will be much smaller. In contrast, for the logistic setting, the G-optimal design objective may be large and we only have a naive bound $f_G(\mathcal{X}) \le \kappa_0 d$ obtained by naively lower bounding $H(\lambda) \ge \kappa_0\sum_{x\in\mathcal{X}}\lambda_x xx^\top$. In general these two criteria can produce extremely different designs. We provide an example where these designs are very different in our supplementary.

Though experimental design for logistic models is an important and abundant topic, e.g., social science applications where tight estimates on entries of θ are required for causal interpretation (Erlander 2005), as far as we know, the design above has not previously been proposed. The closest that we are aware of is Russell (2018) that considers $\dot{\mu}(x^\top\theta)\|x\|^2_{H(\theta)^{-1}}$ rather than $\dot{\mu}(x^\top\theta)^2\|x\|^2_{H(\theta)^{-1}}$. Most existing studies on optimal design in nonlinear models study theoretical properties of the design problem assuming the knowledge of θ^* and then uses a plugin estimate. However, they hardly study under what conditions the plugin estimate must satisfy for the plug-in design problem to closely approximate the true design problem and how one can efficiently collect data for the plugin estimate. In contrast, we address these in our paper for the H- and G-optimal design problem for logistic models.

2.1 From Experimental Design to Regret Minimization

We now introduce our primary algorithm, `HOMER` (**H**-**O**ptimal **ME**thod for **R**egret) which is centered around the H-optimal design objective and the confidence bound on the mean parameter given in Lemma 1. To build an initial estimate $\hat{\theta}_0$ of θ_*, the algorithm begins by calling a warmup procedure `WarmUp` with the following guarantee:

Definition 1. A warmup algorithm $\mathcal{A}$ is said to be δ-valid if it returns an estimator $\hat{\theta}_0$ such that it is certified to have $\mathbb{P}(\forall x \in \mathcal{X} : |x^\top(\hat{\theta}_0 - \theta_*)| \le 1) \ge 1 - \delta$.

One natural attempt would the aforementioned experimental design approach of solving $g^* = \min_{\lambda\in\triangle_{\mathcal{X}}} \max_{x\in\mathcal{X}} \|x\|^2_{H_\lambda(\theta_*)^{-1}}$. We are guaranteed a solution λ^* with a support of at most $d(d+1)/2$ via Caratheodory's theorem; e.g., see Lattimore and Szepesvári (2020, Theorem 21.1). We can then pull arm x exactly $\lceil\lambda^*_x g^*\gamma(d)\rceil$ times to satisfy (2), which in turns makes the MLE $\hat{\theta}_0$ trained on these samples to perform a δ-valid warmup. However, we do not know θ_*. Fortunately, when an upper bound S on θ_* is known, we can we consider the following naive warm-up procedure:

$$g^{\mathsf{naive}} = \min_{\lambda\in\triangle_{\mathcal{X}}} \max_{x\in\mathcal{X}} \|x\|^2_{(H^{\mathsf{naive}}_\lambda)^{-1}}$$
$$\text{where} \quad H^{\mathsf{naive}}_\lambda = \sum_x \lambda_x\dot{\mu}(\|x\|S)xx^\top . \tag{5}$$

Let $\hat{\lambda}^{\mathsf{naive}}$ be the solution to this problem. Since $x^\top\theta_* \le \|x\|\|\theta_*\|$, we have $g^{\mathsf{naive}} \ge g^*$, so we can guarantee that the warmup condition (2) is satisfied when pulling each arm x exactly $\lceil\hat{\lambda}^{\mathsf{naive}}_x g^{\mathsf{naive}}\gamma(d)\rceil$ times. Computing the MLE $\hat{\theta}_0$ with these samples leads to a δ-valid warmup with the sample complexity of $O(g^{\mathsf{naive}}\gamma(d) + d^2)$ which in the worse case is $O(\kappa d^2)$. We discuss a more efficient warmup algorithm in Section 3; in this section, let us use this naive warmup procedure.

The pesudocode of `HOMER` can be found in Algorithm 1. In each round k, `HOMER` maintains an active set of arms $\mathcal{X}_k$ and computes two experimental design objectives over $\mathcal{X}_k$, each using the MLE estimate from the previous round $\hat{\theta}_{k-1}$. The main experimental design, denoted λ^H_k, is the H-optimal design in Line 5 which ensures the gap $\mu(x_*^\top\theta_*) - \mu(x^\top\theta_*)$ is estimated to precision 2^{-k}. This allows us to remove any arm x whose gap is significantly larger than 2^{-k}. The second optimal design, denoted λ^G_k, is a G-Optimal design given in Line 7. It is necessary to ensure that the warmup condition holds in each round as samples are not shared across rounds. In order to trade off these two, possibly competing design criteria, `HOMER` computes a mixture of λ^H_k and λ^G_k (denoted

Algorithm 1: HOMER: H Optimal MEthod for Regret

Require: $\epsilon, \delta, \mathcal{X}, \kappa_0$

$k = 1, \mathcal{X}_1 = \mathcal{X}$, $\gamma(d,n,\delta) := \max\{d + \log(6(2+n)/\delta), 6.1^2\log(6(2+n)/\delta)\}$

$\hat{\theta}_0 \leftarrow$ WarmUp($\mathcal{X}$)

while $|\mathcal{X}_k| > 1$ **do**

$\delta_k = \delta/(4(2+|\mathcal{X}|)|\mathcal{X}|k^2)$

$\lambda_k^H = \arg\min_{\lambda\in\triangle_{\mathcal{X}_k}} \hat{h}_k(\lambda)$

for $\hat{h}_k(\lambda) := \max_{x\in\mathcal{X}_k} \dot{\mu}(x^\top\hat{\theta}_{k-1})^2\|x\|^2_{H_\lambda(\hat{\theta}_{k-1})^{-1}}$

$\lambda_k^G = \arg\min_{\lambda\in\triangle_{\mathcal{X}_k}} \left(\hat{g}_k(\lambda) := \max_{x\in\mathcal{X}_k}\|x\|^2_{H_\lambda(\hat{\theta}_{k-1})^{-1}}\right)$

$n_k^H = \lceil 6(1+\epsilon)6.1^2 3^3 2^{2k}\hat{h}_k(\lambda_k^H)\log(\delta_k^{-1})\rceil$

$n_k^G = \lceil 6(1+\epsilon)\gamma(d,|\mathcal{X}_k|,\delta_k)\hat{g}_k(\lambda_k^G)\rceil$

$\tilde{\lambda}_{k,i} = \max\left\{\frac{n_k^H}{n_k^H+n_k^G}\lambda_{k,i}^H, \frac{n_k^G}{n_k^G+n_k^H}\lambda_{k,i}^G\right\}, \forall i\in[n]$,

where $\lambda_{k,i}^H$ and $\lambda_{k,i}^G$ are the i-th entry of λ_k^H and λ_k^G

$\lambda_{k,i} = \tilde{\lambda}_{k,i}/\sum_{j=1}^n \tilde{\lambda}_{k,j}$ and $n_k = \max\{n_k^H + n_k^G, r(\epsilon)\}$

$x_1,\ldots,x_{n_k} \leftarrow$ Round(n_k, λ_k, ϵ)

Observe $y_1,\cdots,y_{n_k}$, compute MLE $\hat{\theta}_k$ with $\{(x_i,y_i)\}_{i=1}^{n_k}$.

$\mathcal{X}_{k+1} \leftarrow \mathcal{X}_k \setminus \left\{x\in\mathcal{X}_k : \max_{x'\in\mathcal{X}_k}\mu(x'^\top\hat{\theta}_k) - \mu(x^\top\hat{\theta}_k) \geq 2\cdot 2^{-k}\right\}$

$k \leftarrow k+1$

end while

Continue to play the unique arm in $\mathcal{X}_k$ for all time.

λ_k) with approximately $1 - 2^{-k}$ of the mass being given to the H-optimal design λ_k^H. Rather than sampling directly from this distribution HOMER relies on an efficient rounding procedure, Round. Given a distribution λ, tolerance ϵ, and number of samples n, Round returns an allocation $\{x_i\}_{i=1}^n$ such that $H_n(\theta)$ is within a factor of $(1+\epsilon)$ $H(\lambda,\theta)$ for any $\theta\in\mathbb{R}^d$ provided $n \geq r(\epsilon)$ for a minimum number of samples $r(\epsilon)$. Efficient rounding procedures are discussed in Fiez et al. (2019). Recent work (Camilleri, Katz-Samuels, and Jamieson 2021) has shown how to avoid rounding for linear bandits through employing robust mean estimators, but it remains an open question for logistic bandits. HOMER passes the mixed distribution to Round and samples according to the returned allocation to compute an MLE estimate $\hat{\theta}_k$. Finally it removes suboptimal arms using plug in estimates of their means $\mu(x^\top\hat{\theta}_k)$.

Theoretical Guarantees We now present theoretical guarantees of HOMER.

Theorem 2. *Fix $\delta \leq e^{-1}$ and suppose WarmUp draws at most T_B samples and incurs regret at most R_B. Define $T' := T - T_B$ and assume $T' > 0$. Choose a rounding procedure with $r(\varepsilon) = O(d/\varepsilon^2)$ (e.g., Fiez et al. (2019, Appendix B)) and set $\varepsilon = O(1)$. Let $\Delta := \min_{x\in\mathcal{X}\setminus\{x_*\}}\mu(x_*^\top\theta_*) - \mu(x^\top\theta_*)$ Then, with probability at least $1-2\delta$, HOMER obtains a regret within a doubly logarithmic factor of*

$$R_B + \min_{\nu\geq 0}\left(T\nu + \frac{d\dot{\mu}(x_*^\top\theta_*)}{\Delta\vee\nu}\log\left(\frac{|\mathcal{X}|}{\delta}\right) + d\log\left(\frac{1}{\Delta\vee\nu}\right)\right)$$
$$+ d\kappa_0\left(d + \log\left(\frac{|\mathcal{X}|}{\delta}\right)\right).$$

Remark 1. Using the naive warmup (5), the one can show that $R_B = O\left(d^2\kappa\log(|\mathcal{X}|/\delta)\right)$.

The OFULog algorithm of (Abeille, Faury, and Calauzènes 2021) achieves a regret bound of $\tilde{O}(d\log^2(\kappa)\sqrt{\dot{\mu}(x_*^\top\theta_*)T} + d^2\kappa)$ where the lower order term may be improved for the case where there are few sub-optimal arms so the lower order term scales with the number of arms without the factor κ. SupLogistic of Jun et al. (2021) follows SupLinRel-type sampling scheme to eliminates arms based on their natural parameter estimates and achieves a regret of $\tilde{O}(\sqrt{dT} + \kappa^2 d^3)$.[1] To compare the leading term of the regret bound, SupLogistic achieves a better dependence on d and κ. However, it has a worse dependence with $\dot{\mu}(x_*^\top\theta_*)$ due to the changing arm set setting and makes strong assumptions of stochastic contexts and bounded minimum eigenvalues. The regret bound of OFULog is free of these assumptions and have a better lower-order terms, but the leading term is suboptimal.

In the following Corollary, we further upper bound the result of Theorem 2 in order to compare to the results of (Abeille, Faury, and Calauzènes 2021) and (Jun et al. 2021) and show that HOMER enjoys state of the art dependence on d, $\dot{\mu}(x_*^\top\theta_*)$, and $\log(\kappa)$ simultaneously and also matches the state-of-the-art worst-case lower order terms $d^2\kappa$.

Corollary 3. *Suppose we run HOMER with $\delta = 1/T$ with the naive warmup (5) in the same setting as Theorem 2. Then, HOMER satisfies*

$$\mathbb{E}[R_T] = \hat{O}\left(\left(\sqrt{dT\dot{\mu}(x_*^\top\theta_*)\log(|\mathcal{X}|T)}\right.\right.$$
$$\left.\left.\wedge\frac{d\dot{\mu}(x_*^\top\theta_*)\log(|\mathcal{X}|T)}{\Delta}\right) + d^2\kappa\log(|\mathcal{X}|T)\right)$$

where $\hat{O}$ hides doubly logarithmic factors.

This highlights that HOMER simultaneously enjoys the optimal dependence on d exhibited by SupLogistic and the state of the art dependence on $\dot{\mu}(x_*^\top\theta_*)$ seen in OFULog. Furthermore, HOMER avoids a dependence on $\log(\kappa)$ in its leading term.

3 Warmup by Accepts and Rejects (WAR)

We now describe our novel and efficient warmup procedure that is δ-valid. The problem with the naive warmup (5) is that when arms in the support of the design have norm close to 1, then g^{naive} will scale with κd leading a to κd^2 regret lower order term. In this section, we propose a novel warmup algorithm called Warmup by Accepts and Rejects (WAR) that can significantly reduce the number of samples while provably being never worse than the naive warmup.

Inuition from 1d. Before describing the method, we provide some intuition from the case of $d = 1$ with $\mathcal{X} = [-1, 1]$.

[1] Jun et al. (2021) in fact have reported that the lower order term is $O(\kappa^2 d)$, but there is a hidden dependency on d. This is because they assume that the expected minimum eigenvalue is at least σ_0^2, but σ_0^2 is $1/d$ at best. Our reported rate is for the best case of $\sigma_0^2 = 1/d$.

In this case, the design problem is simplified because of the fact that λ^*, the solution to the G-optimal problem 4, is supported on only one arm. Thus, it suffices to find $x^\dagger$:

$$x^\dagger = \arg\max_{x\in\mathcal{X}} \dot{\mu}(x\theta_*)x^2. \tag{6}$$

Without loss of generality, assume that $|\theta_*| \geq \arg\max_{z\in\mathbb{R}} \dot{\mu}(z)z^2 = 2.399..$ (otherwise $\kappa_0 \leq 13.103$ is not too large and we can employ the naive warmup). Then,

$$\begin{aligned}\max_{x\in[-1,1]} \dot{\mu}(x\theta_*)x^2 &= \frac{1}{(\theta_*)^2} \max_{x\in[-1,1]} \dot{\mu}(x\theta_*)x^2(\theta_*)^2 \qquad (7)\\ &= \frac{1}{(\theta_*)^2} \max_{z\in[-|\theta_*|,|\theta_*|]} \dot{\mu}(z)z^2 \overset{(a)}{=} \frac{0.439..}{(\theta_*)^2}\end{aligned}$$

where (a) is by the assumption $|\theta_*| \geq 2.399..$ and numerical evaluation, and $x^\dagger = \arg\max_{x\in[-1,1]} \dot{\mu}(x\theta_*)x^2 = \pm\frac{2.399..}{|\theta_*|}$.

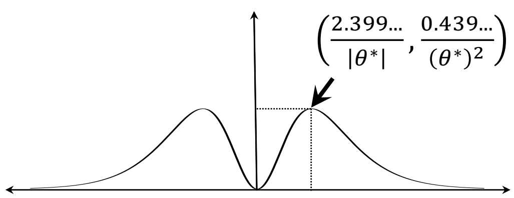

Figure 1: The objective function of (6) with $\mathcal{X} = [-1, 1]$.

We summarize the solution of the optimization problem above in Figure 1. We make two observations from this 1d example. Firstly, somewhat surprisingly the best design for the warmup does not always choose the arm with the largest magnitude unlike G-optimal design in the linear case. Secondly, in the best case, the number of samples needed to ensure the warmup is $O(\theta_*^2)\gamma(1)$. Thus, we speculate that, for $d > 1$, only $O(\|\theta_*\|^2)d\gamma(d)$ samples may be needed for the warmup, which is significantly smaller than $\kappa d\gamma(d)$.

The challenge is that we do not know θ^*. However, we can use a halving procedure to find a constant factor approximation of $|\theta^*|$ in $O(\log(|\theta_*|)$ sample complexity. The key idea is that by choosing an arm x s.t. $|x| \approx 1/|\theta_*|$, the rewards conditioned on x must have high variance, guaranteeing $|x\theta_*|$ is sufficiently small. Thus, starting from the arm $x = 1$, we can sample until verifying that the reward variance conditional on x is either small enough (e.g., $|x\theta_*| \geq 1$) or large enough (e.g., $|x\theta_*| \leq 2$), which can be done using confidence bounds such as empirical Bernstein's inequality. Once we verify that the variance is small enough, it means that $|x|$ is large enough, so we can then move on to the next arm $x = 1/2$ (i.e., halving). We repeat this process until we identify an arm whose variance is large enough, which means that we have approximately solved (6). It is easy to see that this procedure terminates in $\sim \log|\theta_*|$ iterations, spending total $\sim \log|\theta_*|$ samples.

Note that finding the arm $x \approx 1/|\theta_*|$ alone is not sufficient because we need to certify that the warmup condition (2) holds. For this, we realize that the series of accept/rejects form a confidence bound on θ_*. Specifically if $\hat{x}$ is the arm that is accepted in the last iteration we have that $\theta_* \in C := \{\theta \in \mathbb{R} : 1/(2|\hat{x}|) \leq |\theta_*| \leq 2/|\hat{x}|\}$ using the fact that $2\hat{x}$ was rejected. We can then solve the design $\min_\lambda \max_{x\in[-1,1],\theta\in C} \|x\|^2_{H_\lambda(\theta)^{-1}}$ and then sample to certify (2) with high probability.

Warmup by Accepts and Rejects (`WAR`) The remaining challenge is to extend the previous halving procedure to generic discrete arm sets and also to the multidimensional case. This leads to our algorithm `WAR` described in Algorithm 2. Before describing the details, we introduce key quantities. Assuming an arm x is pulled N times, let $\hat{\mu}_x$ be the empirical mean of the observed rewards. We now construct valid lower and upper confidence bounds L_x and U_x on $|x^\top\theta_*|$ where L_x can be 0 and U_x can be ∞. To do so, for each arm x we will use an anytime version of the empirical Bernstein inequality (Mnih, Szepesvári, and Audibert 2008) which has the following confidence width:

$$W_x := \sqrt{\frac{\hat{\mu}_x(1-\hat{\mu}_x)2\log(3/\delta_N)}{N}} + \frac{3\log(3/\delta_N)}{N}$$

where $\delta_N = |\mathcal{X}|N(N+1)/\delta$.

Theorem 4 (Mnih, Szepesvári, and Audibert (2008))**.** *Define* $\mathcal{E} = \left\{\forall x \in \mathcal{X}, \forall N \geq 1, \mu(x^\top\theta_*) \in [\hat{\mu}_x - W_x, \hat{\mu}_x + W_x\right\}$. *Then,* $\mathbb{P}(\mathcal{E}) \geq 1-\delta$.

We can then define $L_x = \mu^{-1}(0 \vee (\hat{\mu}_x - W_x))$ and $U_x = \mu^{-1}(1 \wedge (\hat{\mu}_x + W_x))$.

The pseudocode of `WAR` can be found in Algorithm 2. Let m be the stage index and define $\mathcal{H}_m$ be the set of arms that were pulled up to (and including) the stage m. Let S be a known upper bound on $\|\theta_*\|$ and let $\mathcal{B}_d(S)$ be the L2-ball of radius S. We define a confidence set on θ_*:

$$C_m = \left\{\theta \in \mathcal{B}_d(S) : |x^\top\theta| \in [L_x, U_x], \forall x \in \mathcal{H}_m\right\}$$

Let $\dot{\mu}_m^{\mathsf{opt}}(x)$ be the optimistic estimate and $\dot{\mu}_m^{\mathsf{pes}}(x)$ be the pessimistic estimate defined by

$$\dot{\mu}_m^{\mathsf{opt}}(x) = \max_{\theta\in C_m} \dot{\mu}(x^\top\theta) \quad \text{and} \quad \dot{\mu}_m^{\mathsf{pes}}(x) = \min_{\theta\in C_m} \dot{\mu}(x^\top\theta)$$

Define the accept and reject event on line 5 as

$$\mathsf{Accept}(x) := \{U_x < U\} \quad \text{and} \quad \mathsf{Reject}(x) := \{L_x > L\}$$

for some $0 < L < U$. `WAR` consists of two parts. The first part, which we call optimistic probing, is the halving idea above extended to higher dimensions. The difference from 1d is that we first find d arms that form a good basis, under the assumption that their variances are not small. We then perform accepts/rejects to filter out arms with small variances that would likely introduce κ_0 dependency in the planning. This filtering is done using $\dot{\mu}^{\mathsf{opt}}(x)$ because we do not want arms whose variances are small even in the best case. Note we use the threshold L/r rather than L in line 9 in order to get the halving effect. The second part called Pessimistic Planning simply computes the design based on the pessimistic estimate of the variances $\dot{\mu}^{\mathsf{pes}}(x)$, which allows the resulting sample assignments to certify the warmup condition 2. We provide a practical and detailed version of `WAR` in our supplementary.

Algorithm 2: Warmup by Accepts and Rejects (`WAR`)

Require: Arm set $\mathcal{X}$, parameters $L < R$, $r > 1$.

1: Set $\mathcal{S} = \mathcal{X}$.
2: **[Optimistic probing]**
3: **for** $m = 1, 2, \dots$ **do**
4: Solve $\min_{\lambda \in \Delta_\mathcal{S}} \max_{x \in \mathcal{S}} \|x\|^2_{V_\lambda^{-1}}$ where $V_\lambda = \sum_{x \in \mathcal{S}} \lambda_x x x^\top$ to obtain a 2-approximate solution supported on $O(d \log(\log(d)))$ points; see Remark 2 below. Call this solution $\hat{\lambda}^{(m)}$.
5: For every arm $x \in \mathsf{Supp}(\hat{\lambda}^{(m)})$, pull it until we either accept or reject (if it was pulled previously, skip sampling and reuse the accept/reject result).
6: **if** all the arms in $\mathsf{Supp}(\hat{\lambda}^{(m)})$ are accepted **then**
7: **break**
8: **end if**
9: $\mathcal{S} \leftarrow \mathcal{S} \setminus \left\{ x \in \mathcal{S} : \dot{\mu}_m^{\mathsf{opt}}(x) \le \dot{\mu}\left(\frac{L}{r}\right)\right\}$.
10: **end for**
11: **[Pessimistic planning]**
12: Let $\hat{\lambda}^{\mathtt{WAR}}$ and $g^{\mathtt{WAR}}$ be the solution and the objective of $\arg\min_{\lambda \in \Delta_\mathcal{X}} \max_{x \in \mathcal{X}} \|x\|^2_{(H_\lambda^{\mathsf{pes}})^{-1}}$ where $H_\lambda^{\mathsf{pes}} = \sum_{x \in \mathcal{X}} \lambda_x \dot{\mu}_m^{\mathsf{pes}}(x) x x^\top$ so that the support size of $\hat{\lambda}^{\mathtt{WAR}}$ is at most $d(d+1)/2$.
13: Pull arm x exactly $\lceil \hat{\lambda}_x^{\mathtt{WAR}} \cdot \gamma(d) g^{\mathtt{WAR}} \rceil$.
14: **Return** the MLE $\hat{\theta}$ computed from these samples.

Remark 2. Section 3 of Todd (2016) provides various algorithms for solving the G-optimal design (line 4 of Algorithm 2). For example, using the Kumar-Yildirim initialization (Kumar and Yildirim 2005) along with the Khachiyan first-order algorithm (Khachiyan 1996) results in a 2-approximate solution with the support size of $O(d \log(\log(d)))$ (see Lemma 3.7(ii) of Todd (2016)).

`WAR` enjoys the following correctness guarantee:

Theorem 5. *Assume $|\theta_*| \ge 2.399$. Suppose $U \le 2.399$. Then, with probability at least $1 - \delta$,*

(i) `WAR` is a δ-valid warmup algorithm.

(ii) The sample complexity of the Pessimistic planning phase of `WAR` is never worse than the naive warmup, i.e., $O(g^{\mathsf{naive}} \gamma(d))$; see (5).

(iii) For $d = 1$, if $\mathcal{X}_L := \{x \in \mathcal{X} : |x \theta_*| \le \frac{L}{r}\}$ is nonempty, then our algorithm finds a design λ whose multiplicative approximation ratio w.r.t. the optimal continuous design

$$\min_{\lambda \in \Delta_{[-1,1]}} \max_{x \in \mathcal{X}} \|x\|^2_{H_\lambda(\theta_*)^{-1}} = \frac{(\theta^*)^2}{0.439...} \cdot \max_{x \in \mathcal{X}} |x|$$

is $\frac{1}{0.41} \frac{0.439..}{\dot{\mu}(x_0 \theta_*)(x_0 \theta_*)^2}$ where $x_0 := \arg\max_{x \in \mathcal{X}_L} |x \theta_*|$.

Theorem 5(iii) provides an interesting characterization in 1d of when we can guarantee that the warmup does not scale with κ but rather scale with $\|\theta^*\|^2 \approx \log^2(\kappa)$. When $L/r = 2$, the approximation ratio is ≈ 18.23 in the best case of $x_0 = L/r$, and in general it degrades as $|x_0 \theta_*|$ decreases. Theorem 5 reflects the importance of existence of arms with large variances which makes sense given that the concentration bound scales like $\|x\|_{H_t(\theta_*)^{-1}}$. Note that by reducing r or increasing L we can guarantee that $\mathcal{X}_L$ is nonempty, at the cost of increasing the sample complexity.

Theorem 6. *Let $\Delta_{\mathsf{w}} := \mu(U) - \mu(L)$. In `WAR`, Pessimistic planning assigns total $O(\gamma(d) g^{\mathtt{WAR}} + d^2)$ samples. Furthermore, under the same assumptions as Theorem 5 with probability at least $1 - 2\delta$, when $d = 1$ Optimistic probing takes no more than $2 + \log_r(|\theta_*|/L)$ iterations and each iteration of optimistic probing takes no more than $O(d \log(\log(d)) \cdot \Delta_{\mathsf{w}}^{-2} \log(\Delta_{\mathsf{w}}^{-2} |\mathcal{X}|/\delta))$ samples.*

Specifically, a smaller r prolongs the number of optimistic probing iterations, and a larger L increases Δ_{w}^{-2} so the per-iteration sample complexity increases as well. One can show that $g^{\mathtt{WAR}}$ is $O(d)$ ignoring the dependence on $\|\theta_*\|$, so the total number of samples assigned by pessimistic planning $O(d^2)$. Proving the overall sample complexity of `WAR` for multi-dimensional cases would likely require analyzing how the volume of the confidence set evolve over optimistic probing iterations; we leave this as future work.

Numerical Evaluation To verify the performance of WAR numerically, we have drawn 20 arms from the a three-dimensional unit sphere. The unknown θ^* was drawn the same way but scaled to have the norm $S \in \{2, 4, 8\}$. We have run the naive warmup (5), WAR (2), and the oracle warmup that solves $g^* = \min_{\lambda \in \Delta_\mathcal{X}} \max_{x \in \mathcal{X}} \|x\|^2_{H_\lambda(\theta_*)^{-1}}$. We then computed the total number of samples required to satisfy the warmup condition (2) from each method, ignoring the integer effect for simplicity. We repeat this process 5 times and report the result in Table 1 where WAR is significantly better than the naive warmup and not far from the oracle warmup.

4 On Dimension Dependence in Warmup

In this section we discuss the tightness of the warm-up condition in Theorem 1 of Jun et al. (2021) as well as Lemma 1. Both confidence widths are tight asymptotically as they match up to constant factors the Gaussian-like deviation observed from the delta method. What is not immediately clear is whether the condition that $\max_{s \in [t]} \|x_s\|^2_{H_t(\theta_*)^{-1}} \lesssim \frac{1}{d}$ is optimal (note we omit the factors involving δ for simplicity).[2]. To satisfy this, one needs to pull at least $\Omega(d^2)$ arms even in the best case where $\dot{\mu}(x_s^\top \theta_*) = \Omega(1)$. This is quite different from the standard linear model's fixed design inequality (e.g., Lattimore and Szepesvári (2020, Eq (20.2))) that requires us to pull d linearly independent arm. Requiring only $O(d)$ arm pulls corresponds to relaxing the warmup condition to the following conjectured one:

$$\max_{s \in [t]} \|x_s\|^2_{H_t(\theta_*)^{-1}} \lesssim 1 \ . \tag{8}$$

which is likely necessary. Note, if we have to pull d^2 arms before using the concentration inequality, we will have to pay the lower order term of d^2 in regret, implying that the regret bound is vacuous up to $T = O(d^2)$, the current best known rate for logistic linear bandits. Again this is in contrast with the linear setting where the regret is vacuous up to time $T = O(d)$ only.

[2]When this condition is not true, one can use the confidence bound of Faury et al. (2020), which comes with a factor of $\sqrt{d}S$ in the width (can also be tuned to $\sqrt{dS}$).

Warmup	$S=2$	$S=4$	$S=8$
Naive	8,377± 0.3	49,794± 0.1	2,623,477± 3.0
WAR	6,536±237.9	19,701± 805.0	122,405±30815.5
Oracle	4,970± 69.0	11,720±1094.2	50,258± 4052.5

Table 1: Numerical evaluation of the naive warmup, WAR, and the oracle warmup. Each cell contains the average amount of samples required to satisfy the warmup condition and the standard deviation.

We claim that the conjecture (8) cannot be true for the MLE. We provide a sketch of our counterexample. Let $\theta_* = (c, c, \dots, c), c \in \mathbb{R}$ be the true underlying parameter and assume that $\mathcal{X} = \{e_1, \dots, e_d\}$, the canonical basis in $\mathbb{R}^d$, and we sample each arm N times. Denote by $v(i)$ the i-th component of a vector v. In this setup, if $\hat{\theta}$ is the MLE, then $\hat{\theta}(i)$ is just the one-dimensional MLE considered for each dimension i independently. Let x be the target arm. We are interested in controlling the high probability deviation of $x^\top(\hat{\theta} - \theta_*)$, which we call the *prediction error*, by $O(\|x\|_{H_t(\theta_*)^{-1}})$ where H_t is computed by dN samples. The key observation is that the MLE is biased in GLM's except for special cases like linear models; quantifying and correcting the bias has been studied since Bartlett (1953), though the corrections often rely on a plug-in estimate or are asymptotic in nature and do not have precise mathematical guarantees for finite samples. We now show that the prediction error may be dominated by the bias, rather than the variance, and so we are forced to oversample to correct this bias. Specifically, consider the following decomposition of the prediction error in the setting above:

$$\underbrace{\sum_{i=1}^{d} x(i) \cdot (\hat{\theta}(i) - \theta_*(i))}_{=:\text{ (prediction error)}} = \sum_{i=1}^{d} x(i) \cdot \underbrace{(\mathbb{E}[\hat{\theta}(i)] - \theta_*(i))}_{=:\text{ (A)}} + \underbrace{\sum_{i=1}^{d} x(i) \cdot (\hat{\theta}(i) - \mathbb{E}[\hat{\theta}(i)])}_{=:\text{ (B)}} .$$

The bias term (A) is the bias that is incurred per coordinate. In this setting, critically, *the magnitude of the coordinate-wise bias clearly does not depend on the dimension.*[3] By choosing $x = (h_1/\sqrt{d}, \dots, h_d/\sqrt{d})$ with $h_i := \text{sign}(\mathbb{E}[\hat{\theta}(i)] - \theta_*(i))$, one can see that the bias term (A) will grow with d. Consequently, even if the deviation (B) is controlled, i.e. (B) is bounded by $\|x\|_{H_t(\theta_*)^{-1}} = \sqrt{\sum_{i=1}^{d} \frac{(h_i/\sqrt{d})^2}{\dot{\mu}(c)N}} = \sqrt{\frac{1}{\dot{\mu}(c)N}}$ (which does not grow with d), the bias will be the bottleneck in controlling the LHS. This means that, for large enough d, one cannot aim to control the prediction error by $O(\|x\|_{H_t(\theta_*)^{-1}})$ unless we choose the number of samples N as a function of d. This is effectively the role of the warmup condition (2) – it is requiring an oversampling w.r.t. d so the bias is controlled.

We conjecture that the warmup condition (2) is tight for a concentration inequality on the MLE. To explain our reasoning, consider the same setup as above. Suppose the deviation (B) behaves like $\|x\|_{H_t(\theta_*)^{-1}} = \sqrt{\frac{1}{\dot{\mu}(c)N}}$. Using the formula by Cordeiro and McCullagh (1991), the bias of order $1/N$ for each coordinate is $\frac{1}{2N} \cdot \frac{\mu(c)-\mu(-c)}{\dot{\mu}(c)}$ (through a second order Taylor series expansion); we also confirm it numerically in our supplementary. Thus, as long as c is bounded away from 0, we have that the first order bias is $\Theta\left(\frac{1}{N\dot{\mu}(c)}\right)$. Let us set $N = q\dot{\mu}(c)^{-1}$ for some $q \geq 1$. Then, the bias (A) is $d \cdot \frac{1}{\sqrt{d}} \cdot \frac{1}{q} = \frac{\sqrt{d}}{q}$ and the deviation term (B) is $\sqrt{\frac{1}{q}}$. To control the bias term to be below the deviation term, we must have $q \geq d$. This means that we need to sample at least $d \cdot N = dq\dot{\mu}(c)^{-1} = d^2\dot{\mu}(c)^{-1}$, which matches the warmup condition (2).

Note that the result above is specific to the MLE. For the special case of the canonical basis arm set, one can use an alternative estimator per coordinate such as the KT estimator $\hat{\theta}(i) = \log(\frac{H+1/2}{T+1/2})$ where H is the number of successes and $T = N - H$ (Krichevsky and Trofimov 1981), which is equivalent to the bias correction method by Cox and Snell (2018). The effect is that the bias now scales like $1/N^2$, which can potentially improve the warmup condition (2). In our supplementary, we empirically verify this and provide a conjecture that an extension of the KT estimator may admit $O(d^{4/3})$ sample complexity for the warmup.

Finally, we emphasize that the fixed design inequalities capture the fundamental aspects of the prediction error *without distributional assumptions on the covariates x* besides conditional independence. The warmup conditions for these inequalities indicating Gaussian tails with the asymptotic variance are closely related to the question of 'when the predictive distribution can be approximately Gaussian'. Yet, we do not know what the fundamental limits are for these warmup conditions beyond the standard linear case nor how to correct the bias of the MLE with precise mathematical guarantees. Just as our discussion above naturally motivated new estimators, research surrounding these inequalities is likely to impact not only bandits but also design of experiments, uncertainty quantification, and improving prediction accuracy in supervised learning.

[3] In fact, the bias of the MLE is not well-defined since the observations can be all 1s or all 0s from an arm. One can go around it by setting a suitable for those special cases; e.g., when observing all 1s (or all 0s), set $\hat{\theta}_i = \log(\frac{\hat{p}}{1-\hat{p}})$ where $\hat{p} = \frac{n-.5}{n}$.

References

Abbasi-Yadkori, Y.; Pal, D.; and Szepesvari, C. 2011. Improved Algorithms for Linear Stochastic Bandits. In

Advances in Neural Information Processing Systems (NeurIPS), 1–19.

Abe, N.; and Long, P. M. 1999. Associative reinforcement learning using linear probabilistic concepts. In *Proceedings of the International Conference on Machine Learning (ICML)*, 3–11.

Abeille, M.; Faury, L.; and Calauzènes, C. 2021. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, 3691–3699. PMLR.

Auer, P. 2002. Using Confidence Bounds for Exploitation-Exploration Trade-offs. *Journal of Machine Learning Research*, 3.

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47(2–3): 235–256.

Bartlett, M. S. 1953. Approximate confidence intervals. II. More than one unknown parameter. *Biometrika*, 40(3/4): 306–317.

Camilleri, R.; Katz-Samuels, J.; and Jamieson, K. 2021. High-Dimensional Experimental Design and Kernel Bandits. *arXiv preprint arXiv:2105.05806*.

Cordeiro, G. M.; and McCullagh, P. 1991. Bias correction in generalized linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3): 629–643.

Cox, D. R.; and Snell, E. J. 2018. *Analysis of binary data*. Routledge, 2nd edition.

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

Erlander, S. 2005. Welfare, freedom of choice and composite utility in the logit model. *Social Choice and Welfare*, 24(3): 509–525.

Faury, L.; Abeille, M.; Calauzènes, C.; and Fercoq, O. 2020. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, 3052–3060. PMLR.

Fiez, T.; Jain, L.; Jamieson, K. G.; and Ratliff, L. J. 2019. Sequential Experimental Design for Transductive Linear Bandits. In *NeurIPS*.

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

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

Jun, K.-S.; Jain, L.; Nassif, H.; and Mason, B. 2021. Improved Confidence Bounds for the Linear Logistic Model and Applications to Bandits. In *International Conference on Machine Learning*, 5148–5157. PMLR.

Khachiyan, L. G. 1996. Rounding of polytopes in the real number model of computation. *Mathematics of Operations Research*, 21(2): 307–320.

Kiefer, J.; and Wolfowitz, J. 1960. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12: 363–366.

Krichevsky, R. E.; and Trofimov, V. K. 1981. The performance of universal encoding. *IEEE Trans. Information Theory*, 27(2): 199–206.

Kumar, P.; and Yildirim, E. A. 2005. Minimum-volume enclosing ellipsoids and core sets. *Journal of Optimization Theory and applications*, 126(1): 1–21.

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

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

Mnih, V.; Szepesvári, C.; and Audibert, J.-Y. 2008. Empirical bernstein stopping. In *Proceedings of the International Conference on Machine Learning (ICML)*, 672–679.

Russell, K. G. 2018. *Design of Experiments for Generalized Linear Models*. CRC Press.

Sawant, N.; Namballa, C. B.; Sadagopan, N.; and Nassif, H. 2018. Contextual Multi-Armed Bandits for Causal Marketing. In *Proceedings of the International Conference on Machine Learning (ICML'18) Workshops*. Stockholm, Sweden.

Teo, C. H.; Nassif, H.; Hill, D.; Srinivasan, S.; Goodman, M.; Mohan, V.; and Vishwanathan, S. 2016. Adaptive, Personalized Diversity for Visual Discovery. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys)*, 35–38. Boston, MA.

Todd, M. J. 2016. *Minimum-volume ellipsoids: Theory and algorithms*. SIAM.