Provably Optimal Algorithms for Generalized Linear Contextual Bandits

Lihong Li [1] **Yu Lu** [2] **Dengyong Zhou** [1]

Abstract

Contextual bandits are widely used in Internet services from news recommendation to advertising, and to Web search. Generalized linear models (logistical regression in particular) have demonstrated stronger performance than linear models in many applications where rewards are binary. However, most theoretical analyses on contextual bandits so far are on linear bandits. In this work, we propose an upper confidence bound based algorithm for generalized linear contextual bandits, which achieves an $\tilde{O}(\sqrt{dT})$ regret over T rounds with d dimensional feature vectors. This regret matches the minimax lower bound, up to logarithmic terms, and improves on the best previous result by a $\sqrt{d}$ factor, assuming the number of arms is fixed. A key component in our analysis is to establish a new, sharp finite-sample confidence bound for maximum-likelihood estimates in generalized linear models, which may be of independent interest. We also analyze a simpler upper confidence bound algorithm, which is useful in practice, and prove it to have optimal regret for certain cases.

1. Introduction

Contextual bandit problems are originally motivated by applications in clinical trials (Woodroofe, 1979). When a standard treatment and a new treatment are available for a certain disease, the doctor needs to decide, in a sequetial manner, which of them to use based on the patient's profiles such as age, general physical status or medicine history. With the development of modern technologies, contextual bandit problems have more applications, especially in web-based recommendation, advertising and search (Agarwal

[1]Microsoft Research, Redmond, WA 98052 [2]Department of Statistics, Yale University, New Haven, CT, USA. Correspondence to: Lihong Li <lihongli@microsoft.com>, Yu Lu <yu.lu@yale.edu>, Dengyong Zhou <denzho@microsoft.com>.

et al., 2009; Li et al., 2010; 2012). In the problem of personalized news recommendation, the website must recommend news articles that are most interesting to users that visit the website. The problem is especially challenging for breaking news, as little data are available to make good prediction about user interest. A trade-off naturally occurs in this kind of sequential decision making problems. One needs to balance *exploitation*—choosing actions that performed well in the past—and *exploration*—choosing actions that may potentially give better outcomes.

In this paper, we study the following stochastic, K-armed contextual bandit problem. Suppose at each of the T rounds, an agent is presented with a set of K actions, each of which is associated with a context (a d-dimensional feature vector). By choosing an action based on the rewards obtained from previous rounds and on the contexts, the agent will receive a stochastic reward generated from some unknown distribution conditioned on the context and the chosen action. The goal of the agent is to maximize the expected cumulative rewards over T rounds. The most studied model in contextual bandits literature is the linear model (Auer, 2002; Dani et al., 2008; Rusmevichientong & Tsitsiklis, 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011), in which the expected rewards at each round is a linear combination of features in the context vector. The linear model is theoretically convenient to work with. However, in practice, we usually have binary rewards (click or not, treatment working or not). Logistic regression model based algorithms have been shown to have substantial improvements over linear models (Li et al., 2012). We therefore consider generalized linear models (GLM) in the contextual bandit setting, in which linear, logistic and probit regression serve as three important special cases.

The celebrated work of Lai & Robbins (1985) first introduces the upper confidence bound (UCB) approach to efficient exploration. Later, the idea of confidence bound has been successfully applied to many stochastic bandits problems, from K-arm bandits problems (Auer et al., 2002a; Bubeck & Cesa-Bianchi, 2012) to linear bandits (Auer, 2002; Abbasi-Yadkori et al., 2011). UCB-type algorithms are both efficient and provable optimal in K-arm bandits and K-armed linear bandits. However, most study are limited to the linear case. While some UCB-type algorithms using GLMs perform well empirically (Li et al., 2012),

there is little theoretical study of them. A natural question arises: can we find an efficient algorithm to achieve the optimal convergence rate for generalized linear bandits?

Our Contributions In this paper, we propose a GLM version of the UCB algorithm called SupCB-GLM that achieves a regret over T rounds of order $\tilde{O}(\sqrt{dT})$. This rate improves the state-of-the-art results of Filippi et al. (2010) by a $\sqrt{d}$ factor, assuming the number of actions is fixed. Moreover, it matches the GLM bandits problem's minimax lower bound indicated by the linear bandits problem and thus is optimal. SupCB-GLM is inspired by the seminal work of Auer (2002), which introduced a technique to construct independence samples in linear contextual bandits. A key observation in proving this result is that the ℓ_2 confidence ball of the unknown parameter is insufficient to calculate a sharp upper confidence bound, yet what we need is the confidence interval in *all* directions. Thus, we prove a finite sample normality type confidence bound for the maximum likelihood estimator of GLM. To the best of our knowledge, this is the first non-asymptotic normality type result for the GLM and might be of its own theoretical value. We also analyze a simple version of UCB algorithm called UCB-GLM that is widely used in practice. We prove it also achieves the optimal regret bound under a reasonable assumption. These results shed light on explaining the good empirical performance of GLM bandits in practice.

Related Work The study of GLM bandits problem goes back at least to Sarkar (1991), who considered *discounted* regrets rather than cumulative regerts. They prove that a myopic rule without exploration is *asymptotically* optimal. Recently, Filippi et al. (2010) study the same stochastic GLM bandit problem considered here. They propose the GLM-UCB algorithm, similar to our Algorithm 1, which achieves a regret of $\tilde{O}(d\sqrt{T})$ after T rounds. However, as we believe the optimal regret for stochastic GLM bandits should be the same as linear case when the number of actions is small, their rates misses a $\sqrt{d}$ term than the optimal rates.

Another line of research focuses on using EXP-type algorithms, which can be applied to almost any model classes (Auer et al., 2002b). These algorithms, which choose actions using a carefully randomized policy, use importance sampling to reduce a bandit problem to its full-information analogue. Later variants of the EXP4 algorithm (Beygelzimer et al., 2010; Agarwal et al., 2014) give an $\tilde{O}(\sqrt{dKT})$ regret that is near-optimal with respect to T. However, these regret bounds have a $\sqrt{K}$ dependence. Moreover, these algorithms can be expensive to run: they either have a computational complexity exponential in d for our GLM case, or need to make a large number of calls to a nontrivial optimization oracle.

Organization Section 2 introduces the generalized linear bandit problem. Section 3 gives a brief review of the statistical properties of generalized linear model, and gives a sharp non-asymptotic normality-type result for GLM parameter estimation which can be of independent value. With this tool, Section 4 presents our algorithms and the main theoretical results. Section 5 concludes the paper with further discussions, including several open problems. All proofs are given in the supplementary materials.

Notations For a vector $x \in \mathbb{R}^d$, we use $\|x\|$ to denote its ℓ_2- norm and x' its transpose. $\mathbb{B}^d := \{x \in \mathbb{R}^d : \|x\| \leq 1\}$ is the unit ball centered at the origin. The weighted ℓ_2-norm associated with a positive-definite matrix A is defined by $\|x\|_A := \sqrt{x'Ax}$. The minimum and maximum singular values of a matrix A are written as $\lambda_{\min}(A)$ and $\|A\|$, respectively. The trace of a matrix A is $\operatorname{tr}(A)$. For two symmetric matrices A and B of the same dimensions, $A \succeq B$ means that $A - B$ is positive semi-definite. For a real-valued function f, we use $\dot{f}$ and $\ddot{f}$ to denote its first and second derivatives. Finally, $[n] := \{1, 2, \ldots, n\}$.

2. Problem Setting

We consider the stochastic K-armed contextual bandit problem. Let T be the number of total rounds. At round t, the agent observes a context consisting of a set of K feature vectors, $\{x_{t,a} \mid a \in [K]\} \subset \mathbb{R}^d$, which is drawn IID from an unknown distribution ν, with $\|x_{t,a}\| \leq 1$. Each feature vector $x_{t,a}$ is associated with an unknown stochastic reward $y_{t,a} \in [0, 1]$. The agent selects one action, denoted a_t, and observes the corresponding reward y_{t,a_t}. Finally, we make a regularity assumption about the distribution ν: there exists a constant $\sigma_0 > 0$ such that $\lambda_{\min}(\mathbb{E}[\frac{1}{K} \sum_{a \in [K]} x_{t,a} x'_{t,a}]) \geq \sigma_0^2$ for all t.

In this paper, we are concerned with the generalized linear model, or GLM, in which there is an unknown $\theta^* \in \mathbb{R}^d$ and a fixed, strictly increasing *link function* $\mu : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}[Y \mid X] = \mu(X'\theta^*)$, where X is the chosen action's feature and Y the corresponding reward. One can verify that linear and logistic models are special cases of GLM with $\mu(x) = x$ and $\mu(x) = 1/(1 + e^{-x})$, respectively.

The agent's goal is to maximize the cumulative expected rewards over T rounds. Suppose the agent takes action a_t at round t. Then the agent's strategy can be evaluated by comparing its expected reward to the best expected reward. To do so, define the optimal action at round t by $a_t^* = \operatorname{argmax}_{a \in [K]} \mu(x'_{t,a}\theta^*)$. Then, the agent's total regret of following strategy π can be expressed as follows

$$R_T(\pi) := \sum_{t=1}^{T} \left(\mu(x'_{t,a_t^*}\theta^*) - \mu(x'_{t,a_t}\theta^*) \right).$$

Note that $R_T(\pi)$ is in general a random variable due to the

possible randomness in π. Denote by $X_t = x_{t,a_t}$, $Y_t = y_{t,a_t}$, and our model can be written as

$$Y_t = \mu(X_t'\theta^*) + \epsilon_t, \tag{1}$$

where $\{\epsilon_t, t \in [T]\}$ are independent zero-mean noise. Here, X_t is a random variable because the agent chooses current action based on previous rewards. Formally, we assume there is an increasing sequence of sigma fields $\{\mathcal{F}_n\}$ such that ϵ_t is $\mathcal{F}_t$-measurable with $\mathbb{E}[\epsilon_t \mid \mathcal{F}_{t-1}] = 0$. An example of $\mathcal{F}_n$ will be the sigma-field generated by $\{X_1, Y_1, \ldots, X_n, Y_n\}$. Also, we assume the noise ϵ_t is sub-Gaussian with parameter σ, where σ is some positive, universal constant; that is, for all t,

$$\mathbb{E}\left[e^{\lambda\epsilon_t} \mid \mathcal{F}_{t-1}\right] \le e^{\lambda^2\sigma^2/2}. \tag{2}$$

In practice, when we have bounded reward $Y_t \in [0,1]$, the noise ϵ_t is also bounded and hence satisfies (2) with some appropriate σ value. In addition to the boundedness assumption on the rewards and feature vectors, we also need the following assumption on the link function μ.

Assumption 1. $\kappa := \inf_{\{\|x\| \le 1, \; \|\theta - \theta^*\| \le 1\}} \dot{\mu}(x'\theta) > 0$.

As we shall see in Section 3, the asymptotic normality of maximum-likelihood estimates implies the necessity of this assumption. Note that this assumption is weaker than Assumption 1 in Filippi et al. (2010), as it only requires to control the *local* behavior of $\dot{\mu}(x'\theta)$ near θ^*.

Assumption 2. μ is twice differentiable. Its first and second order derivatives are upper-bounded by L_μ and M_μ, respectively.

It can be verified that Assumption 2 holds for the logistic link function, where we may choose $L_\mu = M_\mu = 1/4$.

3. Generalized Linear Models

To motivate the algorithms proposed in this paper, we first briefly review the classical likelihood theory of generalized linear models. In the canonical generalized linear model (McCullagh & Nelder, 1989), the conditional distribution of Y given X is from the exponential family, and its density, parameterized by $\theta \in \Theta$, can be written as

$$\mathbb{P}(Y \mid X) = \exp\left\{\frac{YX'\theta^* - m(X'\theta^*)}{g(\eta)} + h(Y, \eta)\right\}. \tag{3}$$

Here, $\eta \in \mathbb{R}^+$ is a known scale parameter; m, g and h are three normalization functions mapping from $\mathbb{R}$ to $\mathbb{R}$. The exponential family (3) is a very broad family of distributions including the Gaussian, binomial, Poisson, gamma and inverse-Gaussian distributions. It follows from standard properties of exponential families (Brown, 1986) that m is infinitely differentiable satisfying $\dot{m}(X'\theta^*) = \mathbb{E}[Y \mid$

$X] = \mu(X'\theta^*)$ and $\ddot{m}(X'\theta^*) = \mathbb{V}(Y \mid X)$. It can be checked that the data generated from (3) automatically satisfies the sub-Gaussian condition (2).

Suppose we have independent samples of $Y_1, Y_2, \ldots, Y_n$ condition on $X_1, X_2, \ldots, X_n$. The log-likelihood function of θ under model (3) is

$$
\begin{aligned}
\log \ell(\theta) &= \sum_{t=1}^{n}\left[\frac{Y_t X_t'\theta - m(X_t'\theta)}{v(\eta)} + c(Y_t, \eta)\right] \\
&= \frac{1}{v(\eta)}\sum_{t=1}^{n}[Y_t X_t'\theta - m(X_t'\theta)] + \text{constant}.
\end{aligned}
$$

Consequently, the maximum likelihood estimate (MLE) may be defined by

$$\hat{\theta}_n \in \operatorname*{argmax}_{\theta \in \Theta} \sum_{t=1}^{n}[Y_t X_t'\theta - m(X_t'\theta)].$$

From classical likelihood theory (Lehmann & Casella, 1998), we know that when the sample size n goes to infinity, the MLE $\hat{\theta}_n$ is asymptotically normal, that is, $\hat{\theta}_n - \theta^* \to \mathcal{N}(0, \mathcal{I}_{\theta^*}^{-1})$, where $\mathcal{I}_\theta = \sum_{t=1}^{n} \dot{\mu}(X_t'\theta)X_t X_t'$ is the Fisher Information Matrix. Note that if $\dot{\mu}(X_t'\theta^*) \to 0$, the asymptotic variance of $x'\hat{\theta}$ can go to infinity for some x. This suggests the necessity of Assumption 1.

As we will see later, the normality result is crucial in our regret analysis of GLM bandits. However, to the best of our knowledge, there is no non-asymptotic normality results of the MLE for GLM. In the following, we present a finite-sample version of the classical asymptotic normality results, which can be of independent interest.

Theorem 1. *Define* $V_n = \sum_{t=1}^{n} X_t X_t'$, *and let* $\delta > 0$ *be given. Furthermore, assume that*

$$\lambda_{\min}(V_n) \ge \frac{512 M_\mu^2 \sigma^2}{\kappa^4}\left(d^2 + \log\frac{1}{\delta}\right). \tag{4}$$

Then, with probability at least $1 - 3\delta$, *the maximum-likelihood estimator satisfies, for any* $x \in \mathbb{R}^d$, *that*

$$|x'(\hat{\theta}_n - \theta^*)| \le \frac{3\sigma}{\kappa}\sqrt{\log(1/\delta)}\,\|x\|_{V_n^{-1}}. \tag{5}$$

This theorem characterizes the behavior of MLE on *every* direction. It implies that $x'(\hat{\theta}_n - \theta^*)$ has a sub-Gaussian tail bound for any $x \in \mathbb{R}^d$. It also provides a rigorous justification of the asymptotic upper confidence bound derived heuristically by Filippi et al. (2010, Section 4.2).

The proof of the theorem is given in the appendix. It consists of two main steps, as is typical for proving normality-type results of MLEs (Van der Vaart, 2000). We first show the $n^{-1/2}$-consistency of $\hat{\theta}$ to θ^*. Then, by using a second-order Taylor expansion or Newton-step, we can prove the desired normality of $\hat{\theta}$.

The condition (4) on $\lambda_{\min}(V_n)$ is necessary for the consistency of estimating linear models (Lai & Wei, 1982; Bickel et al., 2009) and generalized linear models (Fahrmeir & Kaufmann, 1985; Chen et al., 1999). It can be satisfied under mild conditions such as the proposition below, which will be useful for our analysis.

Proposition 1. *Define* $V_n = \sum_{t=1}^n X_t X_t'$, *where* X_t *is drawn iid from some distribution* ν *with support in the unit ball,* $\mathbb{B}^d$. *Furthermore, let* $\Sigma := \mathbb{E}[X_t X_t']$ *be the second moment matrix, and* B *and* $\delta > 0$ *be two positive constants. Then, there exist positive, universal constants* C_1 *and* C_2 *such that* $\lambda_{\min}(V_n) \geq B$ *with probability at least* $1 - \delta$, *as long as*

$$ n \; \geq \; \left(\frac{C_1 \sqrt{d} + C_2 \sqrt{\log(1/\delta)}}{\lambda_{\min}(\Sigma)} \right)^2 + \frac{2B}{\lambda_{\min}(\Sigma)} . $$

Proof Sketch. We give a proof sketch here, and the full proof is found in the appendix. In the following, for simplicity, we will drop the subscript n when there is no ambiguity. Therefore, V_n is denoted V and so on. We will need a technical lemma, which is an existing result in random matrix theory. The version we presented here is adapted from Equation (5.23) of Theorem 5.39 from Vershynin (2012).

Lemma 1. *Let* $A \in \mathbb{R}^{n \times d}$ *be a matrix whose rows* A_i *are independent sub-Gaussian isotropic random vectors in* $\mathbb{R}^d$ *with parameter* σ, *namely,* $\mathbb{E} \exp(x'(A_i - \mathbb{E}A_i)) \leq \exp(\sigma^2 \|x\|^2 / 2)$ *for any* $x \in \mathbb{R}^d$. *Then, there exist positive, universal constants* C_1 *and* C_2 *such that, for every* $t \geq 0$, *the following holds with probability at least* $1 - 2\exp(-C_2 t^2)$, *where* $\varepsilon = \sigma^2(C_1\sqrt{d/n} + t/\sqrt{n})$: $\left\| \frac{1}{n} A'A - \mathbf{I}_d \right\| \leq \max\{\varepsilon, \varepsilon^2\}$.

Let X be a random vector drawn from the distribution ν. Define $Z := \Sigma^{-1/2} X$. Then Z is isotropic, namely, $\mathbb{E}[ZZ'] = \mathbf{I}_d$. Define $U = \sum_{t=1}^n Z_t Z_t' = \Sigma^{-1/2} V \Sigma^{-1/2}$. From Lemma 1, we have that, for any t, with probability at least $1 - 2\exp(-C_2 t^2)$, $\lambda_{\min}(U) \geq n - C_1 \sigma^2 \sqrt{nd} - \sigma^2 t\sqrt{n}$, where σ is the sub-Gaussian parameter of Z, which is upper-bounded by $\|\Sigma^{-1/2}\| = \lambda_{\min}^{-1/2}(\Sigma)$ (see, e.g., Vershynin (2012)). We thus can rewrite the above inequality (which holds with probability $1 - \delta$ as

$$ \lambda_{\min}(U) \geq n - \lambda_{\min}^{-1}(\Sigma) \left(C_1 \sigma^2 \sqrt{nd} + t\sqrt{n} \right) . $$

This implies the following lower bound:

$$ \lambda_{\min}(V) \geq \lambda_{\min}(\Sigma) n - C_1 \sqrt{nd} - C_2 \sqrt{n \log(1/\delta)} . $$

Finally, simple calculations show that the last expression is no less than B as long as n is no smaller than the expression stated in the proposition, finishing the proof. □

4. Algorithms and Main Results

In this section, we are going to present two algorithms. While the first algorithm is computationally more efficient, the second algorithm has a provable optimal regret bound.

4.1. Algorithm UCB-GLM

The idea of upper confidence bounds (UCB) is highly effective in dealing with the exploration and exploitation trade-off in many parametric bandit problems, including K-arm bandits (Auer et al., 2002a) and linear bandits (Abbasi-Yadkori et al., 2011; Auer, 2002; Chu et al., 2011; Dani et al., 2008). For the generalized linear model considered here, since μ is a strictly increasing function, our goal is equivalent to choosing $a \in [K]$ to maximize $x_{t,a}' \theta^*$ at round t. Suppose $\hat{\theta}_t$ is our current estimator of θ^* after round t. An exploitation action is to take the action that maximizes the estimated mean value, while an exploration action is to choose the one that has the largest variance. Thus, to balance exploitation and exploration, we can simply choose the action that maximizes the sum of estimated mean and variance, which can be interpreted as an upper confidence bound of $x_{t,a}' \hat{\theta}_t$. This leads to the algorithm UCB-GLM (Algorithm 1).

Algorithm 1 UCB-GLM

Input: the total rounds T, tuning parameter τ and α.
Initialization: randomly choose $a_t \in [K]$ for $t \in [\tau]$, set $V_{\tau+1} = \sum_{i=1}^\tau X_t X_t'$
For $t = \tau + 1, \tau + 2, \ldots, T$ **do**
  1. Calculate the maximum-likelihood estimator $\hat{\theta}_t$ by solving the equation
  $$ \sum_{i=1}^{t-1} (Y_i - \mu(X_i'\theta)) X_i = 0 \qquad (6) $$
  2. Choose $a_t = \text{argmax}_{a \in [K]} \left(X_{t,a}' \hat{\theta}_t + \alpha \|X_{t,a}\|_{V_t^{-1}} \right)$
  3. Observe Y_t, let $X_t \leftarrow X_{t,a_t}$, $V_{t+1} \leftarrow V_t + X_t X_t'$
End For

UCB-GLM take two parameters. At the initialization stage, we randomly choose actions to ensure a unique solution of (6). The choice of τ in the theorem statement follows from Proposition 1 with $B = 1$. It should be noted that the IID assumption about contextual (i.e., the distribution ν) is only needed to ensure $V_{\tau+1}$ is invertable (similar to the first phase in the algorithm of Filippi et al. (2010)); the rest of our analysis does not depend on this stochastic assumption. The same may be achieved by using regularization (see, e.g., Abbasi-Yadkori et al. (2011)). Another tuning parameter α is used to control the amount of exploration. The larger the α is, the more exploration will be used.

As mentioned earlier, the feature vectors X_t depend on the

previous rewards. Consequently, the rewards $\{Y_i, i \in [t]\}$ may not be independent given $\{X_i, i \in [t]\}$. We instead use results on self-normalized martingales (Abbasi-Yadkori et al., 2011), together with a finite-time normality result like Theorem 1, to prove the next theorem.

Theorem 2. *Fix any $\delta > 0$. There exists a universal constant $C > 0$, such that if we run UCB-GLM with $\alpha = \frac{\sigma}{\kappa}\sqrt{\frac{d}{2}\log(1 + 2T/d) + \log(1/\delta)}$ and $\tau = C\sigma_0^{-2}(d + \log(1/\delta))$, then, with probability at least $1 - 2\delta$, the regret of the algorithm is upper bounded by*

$$R_T \leq \tau + \frac{2L_\mu \sigma d}{\kappa}\log\left(\frac{T}{d\delta}\right)\sqrt{T}.$$

The theorem shows an $\tilde{O}(d\sqrt{T})$ regret bound that is independent of K. Indeed, this rate matches the minimax lower bound up to logarithm factor for the infinite actions contextual bandit problems (Dani et al., 2008). By choosing $\delta = 1/T$ and using the fact that $R_T \leq T$, this high-probability result implies a bound on the *expected* regret: $\mathbb{E}[R_T] = \tilde{O}(d\sqrt{T})$. Our result improves the previous regret bound of Filippi et al. (2010) by a $\sqrt{\log T}$ factor. Moreover, the algorithm proposed in Filippi et al. (2010) involves a projection step, which is computationally more expensive comparing to UCB-GLM. Finally, this algorithm works well in practice. We give a heuristic argument for its strong performance in Section 5, under a specific condition that sometimes are satisfied.

Proof of Theorem 2. We first bound the one-step regret. To do so, fix t and let $X_t^* = x_{t,a_t^*}$ and $\Delta_t = \hat{\theta}_t - \theta^*$, where $a_t^* \in \arg\max_{a \in [K]} \mu(x_{t,a}'\theta^*)$ is an optimal action at round t. The selection of a_t in UCB-GLM implies

$$\langle X_t^*, \hat{\theta}_t \rangle + \alpha \|X_t^*\|_{V_t^{-1}} \leq \langle X_t, \hat{\theta}_t \rangle + \alpha \|X_t\|_{V_t^{-1}}. \quad (7)$$

Then, we have

$$
\begin{aligned}
\langle X_t^*, \theta^* \rangle - \langle X_t, \theta^* \rangle &= \langle X_t^* - X_t, \hat{\theta}_t \rangle - \langle X_t^* - X_t, \hat{\theta}_t - \theta^* \rangle \\
&\leq \alpha\|X_t\|_{V_t^{-1}} - \alpha\|X_t^*\|_{V_t^{-1}} - \langle X_t^* - X_t, \Delta_t \rangle \\
&\leq \alpha\big(\|X_t\|_{V_t^{-1}} - \|X_t^*\|_{V_t^{-1}}\big) + \|X_t^* - X_t\|_{V_t^{-1}}\|\Delta_t\|_{V_t},
\end{aligned}
$$

where the last inequality is due to Cauchy-Schwartz inequality. We have the following two lemmas to bound $\|\Delta_t\|_{V_t}$ and $\|X_t\|_{V_t^{-1}}$, respectively. Their proofs are deferred to the appendix.

Lemma 2. *Let $\{X_t\}_{t=1}^\infty$ be a sequence in $\mathbb{R}^d$ satisfying $\|X_t\| \leq 1$. Define $X_0 = \mathbf{0}$ and $V_t = \sum_{s=0}^{t-1} X_s X_s'$. Suppose there is an integer m such that $\lambda_{\min}(V_{m+1}) \geq 1$, then for all $n > 0$,*

$$\sum_{t=m+1}^{m+n} \|X_t\|_{V_t^{-1}} \leq \sqrt{2nd\log\left(\frac{n+m}{d}\right)}.$$

Lemma 3. *Suppose $\lambda_{\min}(V_{\tau+1}) \geq 1$. For any $\delta \in [1/T, 1)$, define event*

$$\mathcal{E}_\Delta := \left\{\|\Delta_t\|_{V_t} \leq \frac{\sigma}{\kappa}\sqrt{\frac{d}{2}\log(1 + 2t/d) + \log(1/\delta)}\right\}.$$

Then, event $\mathcal{E}_\Delta$ holds for all $t \geq \tau$ with probability at least $1 - \delta$.

We now choose $\alpha = \frac{\sigma}{\kappa}\sqrt{\frac{d}{2}\log(1 + 2T/d) + \log(1/\delta)}$. If event $\mathcal{E}_\Delta$ holds for all $t \geq \tau$, then,

$$
\begin{aligned}
\langle X_t^*, &\theta^* \rangle - \langle X_t, \theta^* \rangle \\
&\leq \alpha\left(\|X_t\|_{V_t^{-1}} - \|X_t^*\|_{V_t^{-1}} + \|X_t^* - X_t\|_{V_t^{-1}}\right) \\
&\leq 2\alpha\|X_t\|_{V_t^{-1}}.
\end{aligned}
$$

Combining the above with Lemma 2 yields

$$
\begin{aligned}
\sum_{t=\tau+1}^T \left(\langle X_t^*, \theta^* \rangle - \langle X_t, \theta^* \rangle\right) &\leq 2\alpha\sqrt{2Td\log\left(\frac{T}{d}\right)} \\
&\leq \frac{2d\sigma}{\kappa}\log\left(\frac{T}{d\delta}\right)\sqrt{T}. \quad (8)
\end{aligned}
$$

Note that μ is an increasing Lipschitz function with Lipschitz constant L_μ and the μ function is bounded between 0 and 1. The regret of algorithm UCB-GLM can be upper bounded as

$$
\begin{aligned}
R_T &= \sum_{t=1}^\tau \left(\mu(\langle X_t^*, \theta^* \rangle) - \mu(\langle X_t, \theta^* \rangle)\right) \\
&\quad + \sum_{t=\tau+1}^T \left(\mu(\langle X_t^*, \theta^* \rangle) - \mu(\langle X_t, \theta^* \rangle)\right) \\
&\leq \tau + L_\mu \sum_{t=\tau+1}^T \left(\langle X_t^*, \theta^* \rangle - \langle X_t, \theta^* \rangle\right).
\end{aligned}
$$

The proof can be finished by applying (8) and the specified value of τ to the bound above. $\qquad \square$

4.2. Algorithm SupCB-GLM

While the algorithm UCB-GLM performs sufficiently well in practice (Li et al., 2012), it is unclear whether it can achieve the optimal rates of $O(\sqrt{dT\log K})$, when K is fixed and small. As mentioned in Section 4.1, the key technical difficulty in analyzing UCB-GLM is the dependence between samples. Inspired by a technique developed by Auer (2002) to create independent samples for linear contextual bandits, we propose another algorithm SupCB-GLM (Algorithm 3), which uses algorithm CB-GLM (Algorithm 2) as a sub-routine.

Algorithm 2 CB-GLM

Input: parameter α, index set $\Psi(t)$, and candidate set A.

1. Let $\hat{\theta}_t$ be the solution of

$$\sum_{i \in \Psi(t)} [Y_i - \mu(X_i'\theta)] X_i = 0$$

2. $V_t = \sum_{i \in \Psi(t)} X_i X_i'$

3. **For** $a \in A$, **do**

$$w_{t,a} = \alpha \|x_{t,a}\|_{V_t^{-1}}, \quad m_{t,a} = \langle x_{t,a}, \hat{\theta}_t \rangle$$

End For

This algorithm also relies on the idea of confidence bound to do exploration. At round t, the algorithm screens the candidate actions based on the value of $w_{t,a}^{(s)}$ through S stages until an action is chosen. At stage s, we set the confidence level at stage s to be 2^{-s}. If $w_{t,a}^{(s)} > 2^{-s}$ for some a, we need to do more exploration on $x_{t,a}$ and thus we choose this action. Otherwise, the actions are filtered in step 2d such that the actions passed to the next stage are close enough to the optimal action. Since all the widths are smaller than 2^{-s}, if $m_{t,a}^{(s)} < m_{t,j}^{(s)} - 2 \cdot 2^{-s}$ for some $j \in A_s$, the action a can not be the optimal action. The filter process terminates when we have already got accurate estimate of all $x_{t,a}'\theta^*$ up to the $1/\sqrt{T}$ level and we do not need to do exploration. Thus in step 2c we just choose the action that maximizes the estimated mean value.

Our algorithm is different from the algorithm SupLinRel in Auer (2002) that we directly maximize the mean, rather than the upper confidence bound, in steps c and d. This modification leads to a simpler algorithm and a cleaner regret analysis. Also, we would like to point out that, unlike SpectralEliminator (Valko et al., 2014), the algorithm can easily handle a changing action set.

The following result, adapted from Auer (2002, lemma 14), shows how the algorithm SupCB-GLM will give us independent samples. For the sake of completeness, we also present its proof here.

Lemma 4. *For all $s \in [S]$ and $t \in [T]$, given $\{x_{i,a_i}, i \in \Psi_s(t)\}$, the rewards $\{y_{i,a_i}, i \in \Psi_s(t)\}$ are independent random variables.*

Proof of Lemma 4. Since a trial t can only be added to $\Psi_s(t)$ in step 2b of algorithm SupCB-GLM, the event $\{t \in \Psi_s\}$ only depends on the results of trials $\tau \in \cup_{\sigma<s}\Psi_\sigma(t)$ and on $w_{t,a}^{(s)}$. From the definition of $w_{t,a}^{(s)}$, we know it only depends on the feature vectors $x_{i,a_i}, i \in \Psi_s(t)$ and on $x_{t,i}$. This implies the lemma. $\square$

Algorithm 3 SupCB-GLM

Input: tuning parameter α, τ, the number of trials T.

Initialization:

> **for** $t \in [\tau]$, randomly choose $a_t \in [K]$.
> Set $S = \lfloor \log_2 T \rfloor$, $\mathcal{F} = \{a1, \cdots, a\tau\}$ and $\Psi_0 = \Psi_1 = \cdots = \Psi_S = \varnothing$.

For $t = \tau+1, \tau+2, \cdots, T$ **do**

1. Initialize $A_1 = [K]$ and $s = 1$.

2. While $a_t =$ Null
   a. Run CB-GLM with α and $\Psi_s \cup \mathcal{F}$ to calculate $m_{t,a}^{(s)}$ and $w_{t,a}^{(s)}$ for all $a \in A_s$.
   b. If $w_{t,a}^{(s)} > 2^{-s}$ for some $a \in A_s$,

   $$\text{set } a_t = a, \quad \text{update } \Psi_s = \Psi_s \cup \{t\}$$

   c. Else if $w_{t,a}^{(s)} \le 1/\sqrt{T}$ for all $a \in A_s$,

   $$\text{set } a_t = \underset{a \in A_s}{\arg\max}\, m_{t,a}^{(s)}, \quad \text{update } \Psi_0 = \Psi_0 \cup \{t\}$$

   d. Else if $w_{t,a}^{(s)} \le 2^{-s}$ for all $a \in A_s$,

   $$A_{s+1} = \{a \in A_s, m_{t,a}^{(s)} \ge \max_{j \in A_s} m_{t,j}^{(s)} - 2 \cdot 2^{-s}\},$$

   $$s \leftarrow s+1.$$

End For

With Lemma 4, we are able to apply the non-asymptotic normality result (5) and thus to prove our regret bound of Algorithm SupCB-GLM.

Theorem 3. *For any $0 < \delta < 1$, if we run the SupCB-GLM algorithm with $\tau = \sqrt{dT}$ and $\alpha = \frac{3\sigma}{\kappa}\sqrt{2\log(TK/\delta)}$ for $T \ge T_0$ rounds, where*

$$T_0 = \Omega\left(\frac{\sigma^2}{\kappa^4}\max\left\{d^3, \frac{\log(TK/\delta)}{d}\right\}\right), \qquad (9)$$

the regret of the algorithm is bounded as

$$R_T \le 45(\sigma L_\mu/\kappa)\sqrt{\log T \log(TK/\delta) \log(T/d)}\sqrt{dT},$$

with probability at least $1 - \delta$. With $\delta = 1/T$, we obtain

$$\mathbb{E}[R_T] = O\left((\log T)^{1.5}\sqrt{dT \log K}\right).$$

The theorem demonstrates an $\tilde{O}(\sqrt{dT \log K})$ regret bound for the algorithm SupCB-GLM. It has been proved in Chu et al. (2011, Theorem 2) that $\sqrt{dT}$ is the minimax lower bound of the expected regret for K-armed linear bandits, a special of the GLM bandits considered here. Therefore,

the regret of our SupCB-GLM algorithm is optimal up to logarithm terms of T and K. To the best of our knowledge, this is the first algorithm which achieves the (near-)optimal rate of GLM bandits.

It is worthwhile to compare Theorem 3 with the result in Theorem 2. When $K = o(2^d)$ is small, the rate of SupCB-GLM is faster, and we improve the previous rates by a $\sqrt{d}$ factor. Here, we give a briefly illustration of how we get rid of the extra $\sqrt{d}$ factor. Both in Theorem 2 and in Filippi et al. (2010), $|x'(\hat{\theta}_n - \theta^*)|$ is upper bounded by using the Cauchy-Schwartz inequality,

$$|x'(\hat{\theta}_n - \theta^*)| \leq \|x\|_{V_n^{-1}} \left\|\hat{\theta}_n - \theta^*\right\|_{V_n}. \tag{10}$$

Lemma 3 in the supplementary material establishes that

$$\left\|\hat{\theta}_n - \theta^*\right\|_{V_n} \leq C_2\sqrt{d\log(T/\delta)}.$$

This will lead to an extra $\sqrt{d}$ factor compared to (5). By using the Cauchy-Schwartz inequality (10), we only make use of the fact that $\hat{\theta}_n$ is close to θ^* in the ℓ_2 sense. However, (5) tells us that actually $\hat{\theta}_n$ is close to θ^* in *every* direction. This is the reason why we are able to remove the extra $\sqrt{d}$ factor to achieve a near-optimal regret. It also explains why the bound in Theorem 2 is tight when K is large. As K goes large, it is likely there is a direction x for which (10) is tight.

Proof of Theorem 3. To facilitate our proof, we first present two technical lemmas. Lemma 5 follows from Lemma 4, Theorem 1, Theorem 5.39 of Vershynin (2012) and a union bound. The proof of Lemma 6 is deferred to the appendix.

Lemma 5. *Fix $\delta > 0$. Choose in SupCB-GLM $\tau = \sqrt{dT}$ and $\alpha = \frac{3\sigma}{\kappa}\sqrt{2\log(TK/\delta)}$. Suppose T satisfies condition (9). Define the following event:*

$$\mathcal{E}_X := \{|m_{t,a}^{(s)} - x_{t,a}'\theta^*| \leq w_{t,a}^{(s)}, \tag{11}$$
$$\forall t \in [\tau+1, T], s \in [S], a \in [K]\}$$

Then, event $\mathcal{E}_X$ holds with probability at least $1 - \delta$.

Proof of Lemma 5. By Lemma 4, we have independent samples now. Then to apply Theorem 1, the key is to lower bound the minimum eigenvalue of V_t. Note that we randomly select the feature vectors at the first $\tau = \sqrt{dT}$ rounds, that is, they are independent. Moreover, the feature vectors are bounded. Thus, $X_1, X_2, \ldots, X_\tau$ are independent sub-Gaussian with parameter 1. It follows from Proposition 1 that

$$\lambda_{\min}(V_t) \geq \lambda_{\min}(V_\tau) \geq c\sqrt{dT}$$

for some constant c with probability at least $1 - \exp(-\sqrt{dT})$. By Theorem 1 and union bound, we have the desired result under condition (9). $\square$

Lemma 6. *Suppose that event $\mathcal{E}_X$ holds, and that in round t, the action a_t is chosen at stage s_t. Then, $a_t^* \in A_s$ for all $s \leq s_t$. Furthermore, we have*

$$\mu(x_{t,a_t^*}'\theta^*) - \mu(x_{t,a_t}'\theta^*)$$
$$\leq \begin{cases} (8L_\mu)/2^{s_t} & \text{if } a_t \text{ is selected in step 2b} \\ (2L_\mu)/\sqrt{T} & \text{if } a_t \text{ is selected in step 2c}. \end{cases}$$

Define $V_{s,t} = \sum_{t\in\Psi_s(T)} X_t X_t'$, then by Lemma 2,

$$\sum_{t\in\Psi_s(T)} w_{t,a_t}^{(s)} = \sum_{t\in\Psi_s(T)} \alpha(\delta)\|x_{t,a_t}\|_{V_{s,t}^{-1}}$$
$$\leq \alpha(\delta)\sqrt{2d\log(T/d)|\Psi_s(n)|}.$$

On the other hand, by the step 2b of SupCB-GLM,

$$\sum_{t\in\Psi_s(T)} w_{t,a_t}^{(s)} \geq 2^{-s}|\Psi_s(T)|.$$

Combining the above two inequalities gives us

$$|\Psi_s(T)| \leq 2^s\alpha(\delta)\sqrt{2d\log(T/d)|\Psi_s(T)|}. \tag{12}$$

Let Ψ_0 be the collection of trials such that a_t is chosen in step 2c. Since we have chose $S = \log_2 T$, each $t \in [\tau+1, T]$ must be in one of Ψ_s and hence, $\{\tau, \tau+1, \ldots, T\} = \Psi_0 \cup (\cup_{s=1}^S \Psi_s(T))$. If we set $\tau = \sqrt{dT}$, we have

$$R_T = \sum_{t=1}^\tau \left(\mu(x_{t,a_t^*}'\theta^*) - \mu(x_{t,a_t}'\theta^*)\right)$$
$$+ \sum_{t=\tau+1}^T \left(\mu(x_{t,a_t^*}'\theta^*) - \mu(x_{t,a_t}'\theta^*)\right)$$
$$\leq \tau + \sum_{t\in\Psi_0} \left(\mu(x_{t,a_t^*}'\theta^*) - \mu(x_{t,a_t}'\theta^*)\right)$$
$$+ \sum_{s=1}^S \sum_{t\in\Psi_s(T)} \left(\mu(x_{t,a_t^*}'\theta^*) - \mu(x_{t,a_t}'\theta^*)\right)$$
$$\leq \sqrt{dT} + T \cdot \frac{2L_\mu}{\sqrt{T}} + \sum_{s=1}^S L_\mu \cdot 2^{3-s} \cdot |\Psi_s(T)|$$
$$\leq \sqrt{dT} + 2L_\mu\sqrt{T}$$
$$+ 8L_\mu\alpha(\delta)\sum_{s=1}^S \sqrt{2d\log\frac{T}{d}|\Psi_s(T)|}$$
$$\leq \sqrt{dT} + 2L_\mu\sqrt{T} + 8L_\mu\alpha(\delta)\sqrt{2d\log(T/d)}\sqrt{ST}$$
$$\leq 45(\sigma L_\mu/\kappa)\sqrt{\log T\log(TK/\delta)\log(T/d)}\sqrt{dT},$$

with probability at least $1 - \delta$. Here, the first inequality is due to the assumption that $0 \leq \mu \leq 1$. The second inequality is Lemma 6. The third inequality is the inequality (12)

and the fourth inequality is implied by Cauchy-Schwartz. This completes the proof of the high-probability result. $\square$

5. Discussions

In this paper, we propose two algorithms for K-armed bandits with generalized linear models. While the first algorithm, UCB-GLM, achieves the optimal rate for the case of infinite number of arms, the second algorithm SupCB-GLM is provable optimal for the case of finite number actions at each round. However, it remains open whether UCB-GLM can achieve the optimal rate for small K.

5.1. A better regret bound for UCB-GLM

A key quantity in determining the regret of UCB-GLM is the minimum eigenvalue of V_t. If we make an addition assumption on the minimum eigenvalue of V_t, we will be able to prove an $O(\sqrt{dT})$ regret bound for UCB-GLM.

Theorem 4. *We run algorithm UCB-GLM with $\tau = \frac{8\sigma^2}{\kappa^2}d\log T$ and $\alpha \leq L_\mu\sigma/\kappa$. For any $\delta \in [1/T, 1)$, suppose there is an universal constant c such that*

$$\sum_{t=\tau+1}^{T} \lambda_{\min}^{-1/2}(V_t) \leq c\sqrt{T}. \tag{13}$$

holds with probability at least $1 - \delta$, and

$$T = \Omega\left(\frac{\sigma R}{\kappa L_\mu}d\log^2 T\right). \tag{14}$$

Then, the regret of the algorithm is bounded by

$$R_T \leq \frac{CL_\mu\sigma}{\kappa}\sqrt{dT\log(T/\delta)}$$

with probability at least $1 - 2\delta$, where C is a positive, universal constant.

This theorem provides some insights of why UCB-GLM performs well in practice. Although the condition in (13) is hard to check and may be violated in some cases, for example, in K-armed bandits, we provide a heuristic argument to justify this assumption in a range of problems. When t is large enough, our estimator $\hat{\theta}_t$ is very close to θ^*. If we assume there is a positive gap between $\langle x_{t,a_t^*}, \theta^*\rangle$ and $\langle x_{t,a}, \theta^*\rangle$ for all $a \neq a_t^*$, we will have $a_t = a_t^*$ after, for example, $\sqrt{T}$ steps. Since $\{x_{t,a}, a \in [K]\}$ are independent for $t \in [T]$, $\{x_{t,a_t^*}\}$ are also independent samples. Then V_t/t will be well-approximated by the covariance matrix of x_{t,a_t^*}, which we denote by Σ_0. In many problem in practice, especially when features are *dense*, it is unlikely the feature vector x_{t,a_t^*} lies in a low-dimensional subspace of $\mathbb{R}^d$. It implies that Σ_0 has full rank, and that we will have $\lambda_{\min}(V_t) = \Theta(t \cdot \lambda_{\min}(\Sigma_0))$ when t is large enough. It

follows that,

$$\sum_{t=\tau+1}^{T} \frac{1}{\sqrt{\lambda_{\min}(V_t)}} = \sum_{t=\tau+1}^{T} \Theta(t^{-1/2}) = O(\sqrt{T}).$$

It should be cautioned that, since we do not know the distribution of our feature vectors, we cannot assume the above gap exists. It is therefore challenging to make the above arguments rigorous. In fact, when studying the ARIMA model in time series, Lai & Wei (1982, Example 1) provide an example such that $\lambda_{\min}(V_t) = O(\log t)$.

5.2. Open Questions

Computational efficient algorithms. While UCB-GLM and SupCB-GLM enjoy good theoretical properties, they can be expensive in some applications. First, they require inverting a $d \times d$ matrix in every step, a costly operation when d is large. Second, at step t, the MLE is computed using $\Theta(t)$ samples, meaning that the per-step complexity grows at least linearly with t for a straightforward implementation of the algorithms. It is therefore interesting to investigate more scalable alternatives. It is possible to use a first-order, iterative optimization procedure to amortize the cost, analogous to the approach of Agarwal et al. (2014).

K-dependent lower bound. Currently, all the lower bound results on (generalized) linear bandits have no dependence on K, the number of arms. The minimax lower bound will be of particularly interest because all current lower bound results assume that $K \leq d$. Although it will at most be a logarithm dependence on K, it is still a theoretically interesting question.

Randomized algorithms with optimal regret rate. As opposed to the deterministic, UCB-style algorithms studied in this paper, randomized algorithms like EXP4 (Auer et al., 2002b) and Thompson Sampling (Thompson, 1933) have advantages in certain situations, for example, when reward observations are delayed (Chapelle & Li, 2012). Recently developed techniques for analyzing Bayes regret in BLM bandits (Russo & Van Roy, 2014) may be useful to analyze the cumulative regret considered here.

References

Abbasi-Yadkori, Yasin, Pál, Dávid, and Szepesvári, Csaba. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, pp. 2312–2320, 2011.

Agarwal, Alekh, Hsu, Daniel, Kale, Satyen, Langford, John, Li, Lihong, and Schapire, Robert E. Taming the monster: A fast and simple algorithm for contextual bandits. In *Proceedings of the 31th International Conference on Machine Learning (ICML)*, pp. 1638–1646, 2014.

Agarwal, Deepak, Chen, Bee-Chung, Elango, Pradheep, Motgi, Nitin, Park, Seung-Taek, Ramakrishnan, Raghu, Roy, Scott, and Zachariah, Joe. Online models for content optimization. In *Advances in Neural Information Processing Systems 21*, pp. 17–24, 2009.

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

Auer, Peter, Cesa-Bianchi, Nicolo, and Fischer, Paul. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002a.

Auer, Peter, Cesa-Bianchi, Nicolo, Freund, Yoav, and Schapire, Robert E. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

Beygelzimer, Alina, Langford, John, Li, Lihong, Reyzin, Lev, and Schapire, Robert E. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 19–26, 2010.

Bickel, Peter J, Ritov, Ya'acov, and Tsybakov, Alexandre B. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Brown, Lawrence D. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*, volume 9 of *Lecture Notes-Monograph Series*. Institute of Mathematical Statistics, 1986.

Bubeck, Sébastien and Cesa-Bianchi, Nicolo. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Chapelle, Olivier and Li, Lihong. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems 24*, pp. 2249–2257, 2012.

Chen, Kani, Hu, Inchi, and Ying, Zhiliang. Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *The Annals of Statistics*, 27(4):1155–1163, 1999.

Chu, Wei, Li, Lihong, Reyzin, Lev, and Schapire, Robert E. Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 208–214, 2011.

Dani, Varsha, Hayes, Thomas P, and Kakade, Sham M. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, pp. 355–366, 2008.

Fahrmeir, Ludwig and Kaufmann, Heinz. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13(1):342–368, 1985.

Filippi, Sarah, Cappe, Olivier, Garivier, Aurélien, and Szepesvári, Csaba. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems 23*, pp. 586–594, 2010.

Lai, Tze Leung and Robbins, Herbert. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Lai, Tze Leung and Wei, Ching Zong. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1):154–166, 1982.

Lehmann, Erich Leo and Casella, George. *Theory of Point Estimation*, volume 31 of *Springer Texts in Statistics*. Springer Science & Business Media, 1998.

Li, Lihong, Chu, Wei, Langford, John, and Schapire, Robert E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, pp. 661–670. ACM, 2010.

Li, Lihong, Chu, Wei, Langford, John, Moon, Taesup, and Wang, Xuanhui. An unbiased offline evaluation of contextual bandit algorithms with generalized linear models. *JMLR Workshop and Conference Proceedings*, 26: 19–36, 2012.

McCullagh, Peter and Nelder, John A. *Generalized Linear Models*, volume 37. CRC press, 1989.

Pollard, David. Empirical processes: Theory and applications. In *NSF-CBMS regional conference series in probability and statistics*, pp. i–86. JSTOR, 1990.

Rusmevichientong, Paat and Tsitsiklis, John N. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

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

Sarkar, Jyotirmoy. One-armed bandit problems with covariates. *The Annals of Statistics*, 19(4):1978–2002, 1991.

Thompson, William R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3–4):285–294, 1933.

Valko, Michal, Munos, Rémi, Kveton, Branislav, and Kocák, Tomáš. Spectral bandits for smooth graph functions. In *Proceedings of the 31th International Conference on Machine Learning (ICML)*, pp. 46–54, 2014.

Van der Vaart, Aad W. *Asymptotic Statistics*, volume 3. Cambridge university press, 2000.

Vershynin, Roman. Introduction to the non-asymptotic analysis of random matrices. In Eldar, Yonina C. and Kutyniok, Gitta (eds.), *Compressed Sensing: Theory and Applications*, pp. 210–268. Cambridge University Press, 2012.

Woodroofe, Michael. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74(368):799–806, 1979.