Randomized Exploration in Generalized Linear Bandits

Branislav Kveton
Google Research

Manzil Zaheer
Google Research

Csaba Szepesvári
DeepMind / University of Alberta

Lihong Li
Google Research

Mohammad Ghavamzadeh
Facebook AI Research

Craig Boutilier
Google Research

Abstract

We study two randomized algorithms for generalized linear bandits. The first, GLM-TSL, samples a generalized linear model (GLM) from the *Laplace approximation* to the posterior distribution. The second, GLM-FPL, fits a GLM to a *randomly perturbed history* of past rewards. We analyze both algorithms and derive $\tilde{O}(d\sqrt{n \log K})$ upper bounds on their n-round regret, where d is the number of features and K is the number of arms. The former improves on prior work while the latter is the first for Gaussian noise perturbations in non-linear models. We empirically evaluate both GLM-TSL and GLM-FPL in logistic bandits, and apply GLM-FPL to neural network bandits. Our work showcases the role of randomization, beyond posterior sampling, in exploration.

1 Introduction

A *multi-armed bandit* [Lai and Robbins, 1985, Auer et al., 2002, Lattimore and Szepesvari, 2019] is an online learning problem where actions of the *learning agent* are represented by *arms*. The arms can be treatments in a clinical trial or ads on a website. After an arm is *pulled*, the agent receives a *stochastic reward*. The agent aims to maximize its expected cumulative reward. Since the agent does not know the mean rewards of the arms in advance, it faces the *exploration-exploitation dilemma*: *explore*, and learn more about the reward distributions of the arms; or *exploit*, and pull the arm with the highest estimated reward thus far.

A *generalized linear bandit* [Filippi et al., 2010, Zhang et al., 2016, Li et al., 2017, Jun et al., 2017] is a variant of the multi-armed bandit where the expected rewards of arms are modeled using a *generalized linear model (GLM)* [McCullagh and Nelder, 1989]. Specifically, the expected reward is a known function μ, such as a sigmoid, of the dot product of a known feature vector and an unknown parameter vector. In the earlier clinical example, the feature and parameter vectors could be treatment indicators and effects of individual treatments, respectively.

Most existing algorithms for generalized linear bandits are based on *upper confidence bounds (UCBs)*. Motivated by the superior performance of randomized GLM algorithms [Chapelle and Li, 2012, Russo et al., 2018], we study two randomized algorithms for this class of problems, GLM-TSL and GLM-FPL. GLM-TSL samples a GLM from the Laplace approximation to the posterior distribution. GLM-FPL fits a GLM to a *randomly perturbed history* of past rewards.

We analyze GLM-TSL and GLM-FPL, and prove that their n-round regret is $\tilde{O}(d\sqrt{n \log K})$, where d is the number of features and K is the number of arms. The regret bound of GLM-TSL improves on the best prior regret bound [Abeille and Lazaric, 2017] by a multiplicative factor of $\sqrt{d / \log K}$ in the finite arm setting and matches it in the infinite arm setting. The regret bound of GLM-FPL is the first for Gaussian noise perturbations in non-linear models, although we derive it under an additional assumption on arm features.

We also evaluate GLM-TSL and GLM-FPL empirically. Both have a state-of-the-art performance in logistic bandits, the most important practical use case of GLM bandits. Just as importantly, the perturbation scheme in GLM-FPL generalizes straightforwardly to complex reward models, such as a neural network. To demonstrate this, we apply GLM-FPL to high-dimensional classification problems and show that it can learn complex neural network mappings from features to rewards. The simplicity of GLM-FPL suggests that it may find broad application in the future.

Proceedings of the 23$^{\text{rd}}$International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

Algorithm 1 General randomized exploration in generalized linear bandits.

```
Inputs: Number of exploration rounds $\tau$

for $t = 1, \dots, n$ do
    if $t > \tau$ then
        $\tilde{\theta}_t \leftarrow$ Randomized MLE on $\{(X_\ell, Y_\ell)\}_{\ell=1}^{t-1}$
        $I_t \leftarrow \arg\max_{i \in [K]} x_i^\top \tilde{\theta}_t$
    else
        Choose $I_t$ based on $\{X_\ell\}_{\ell=1}^{t-1}$
    Pull arm $I_t$ and get reward $Y_{I_t,t}$
    $X_t \leftarrow x_{I_t}$, $Y_t \leftarrow Y_{I_t,t}$
```

2 Setting

We adopt the following notation. The set $\{1, \dots, n\}$ is denoted by $[n]$. All vectors are column vectors. For any *positive semi-definite (PSD)* matrix M, $\lambda_{\min}(M) \geq 0$ is the minimum eigenvalue of M. For any $n \times n$ PSD matrices M_1 and M_2, $M_1 \preceq M_2$ if and only if $x^\top M_1 x \leq x^\top M_2 x$ for all $x \in \mathbb{R}^d$. We let $\|x\|_M = \sqrt{x^\top M x}$ and $\mathrm{Ber}(p)$ be the Bernoulli distribution with mean p. The indicator that event E occurs is $\mathbb{1}\{E\}$. We use $\tilde{O}$ for the big-O notation up to logarithmic factors in horizon n.

A *generalized linear model (GLM)* is a probabilistic model where observation Y given feature vector $x \in \mathbb{R}^d$ has an exponential-family distribution with mean $\mu(x^\top \theta)$, where μ is the *mean function* and $\theta \in \mathbb{R}^d$ are model parameters [McCullagh and Nelder, 1989]. Let $\mathcal{D} = \{(x_\ell, y_\ell)\}_{\ell=1}^n$ be a set of n observations, where $x_\ell \in \mathbb{R}^d$ and $y_\ell \in \mathbb{R}$. The *negative log likelihood* of $\mathcal{D}$ under model parameters θ is

$$L(\mathcal{D}; \theta) = \sum_{\ell=1}^{|\mathcal{D}|} b(x_\ell^\top \theta) - y_\ell x_\ell^\top \theta - c(y_\ell)\,,$$

where c is a real function, and b is twice continuously differentiable and its derivative is the mean function, $\dot{b} = \mu$. The *gradient* and *Hessian* of $L(\mathcal{D}; \theta)$ with respect to θ are

$$\nabla L(\mathcal{D}; \theta) = \sum_{\ell=1}^{|\mathcal{D}|} (\mu(x_\ell^\top \theta) - y_\ell) x_\ell\,, \qquad (1)$$

$$\nabla^2 L(\mathcal{D}; \theta) = \sum_{\ell=1}^{|\mathcal{D}|} \dot{\mu}(x_\ell^\top \theta) x_\ell x_\ell^\top\,, \qquad (2)$$

where $\dot{\mu}$ denotes the derivative of μ. The mean function μ is increasing and therefore its derivative $\dot{\mu}$ is positive. The *maximum likelihood estimate (MLE)* of model parameters is a vector $\theta \in \mathbb{R}^d$ such that $\nabla L(\mathcal{D}; \theta) = \mathbf{0}$.

A *stochastic GLM bandit* [Filippi et al., 2010] is an online learning problem where the rewards of arms are generated by some underlying GLM. Let K be the number of arms, $x_i \in \mathbb{R}^d$ be the *feature vector* of arm $i \in [K]$, and $\theta_* \in \mathbb{R}^d$ be an unknown *parameter vector*. Then the *reward* of arm i in round $t \in [n]$, $Y_{i,t}$, is drawn i.i.d. from a distribution with mean $\mu_i = \mu(x_i^\top \theta_*)$. We assume that $\eta_{i,t} = Y_{i,t} - \mu(x_i^\top \theta_*)$ is σ^2-sub-Gaussian. That is,

$$\mathbb{E}\left[\exp[\lambda \eta_{i,t}]\right] \leq \exp[\lambda^2 \sigma^2 / 2]$$

holds for all arms i, rounds t, and $\lambda \geq 0$. In round t, the agent *pulls* arm $I_t \in [K]$ and observes its reward $Y_{I_t,t}$. The goal of the agent is to maximize its *expected cumulative reward* in n rounds. To simplify notation, we denote the feature vector of arm I_t by $X_t = x_{I_t}$ and its stochastic reward by $Y_t = Y_{I_t,t}$.

Without loss of generality, we assume that arm 1 is the *unique optimal arm*, that is $\mu_1 > \max_{i>1} \mu_i$. Let $\Delta_i = \mu_1 - \mu_i$ be the *suboptimality gap* of arm i. Maximization of the expected cumulative reward over n rounds is equivalent to minimizing the *expected n-round regret*, which is defined as

$$R(n) = \sum_{i=2}^{K} \Delta_i \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{1}\{I_t = i\}\right]. \qquad (3)$$

3 Algorithms

Our GLM bandit algorithms follow the template in Algorithm 1. They *explore* initially in τ rounds, so that the estimated parameters in subsequent rounds have "good" properties. The exploration strategy is detailed in Section 4.5. After the initial exploration, they act greedily with respect to *randomized parameter vectors* $\tilde{\theta}_t$. Specifically, they pull arm $I_t = \arg\max_{i \in [K]} x_i^\top \tilde{\theta}_t$ in round t. If this maximum is not unique, any tie breaking can be used.

3.1 Algorithm `GLM-TSL`

We study two algorithms. The first algorithm, `GLM-TSL`, is a variant of *Thompson sampling* [Thompson, 1933] where the posterior of θ_* is approximated by its *Laplace approximation*. The randomized parameter vector is sampled from the Laplace approximation

$$\tilde{\theta}_t \sim \mathcal{N}(\bar{\theta}_t, a^2 H_t^{-1})\,, \qquad (4)$$

where

$$\begin{aligned} \bar{\theta}_t &= \arg\min_{\theta \in \mathbb{R}^d} L(\{(X_\ell, Y_\ell)\}_{\ell=1}^{t-1}; \theta)\,, \\ H_t &= \sum_{\ell=1}^{t-1} \dot{\mu}(X_\ell^\top \bar{\theta}_t) X_\ell X_\ell^\top\,, \end{aligned} \qquad (5)$$

and $a > 0$ is a tunable parameter. Chapelle and Li [2012] and Russo et al. [2018] evaluated `GLM-TSL` empirically. In addition, Abeille and Lazaric [2017] proved that `GLM-TSL` has $\tilde{O}(d^{\frac{3}{2}} \sqrt{n})$ regret in the infinite arm setting. We prove that `GLM-TSL` has $\tilde{O}(d \sqrt{n \log K})$ regret when the number of arms is K.

3.2 Algorithm GLM-FPL

We also propose a *follow-the-perturbed-leader (FPL)* algorithm, GLM-FPL. In GLM-FPL, the randomized parameter vector is the MLE from past $t-1$ rewards *perturbed with Gaussian noise*,

$$\tilde{\theta}_t = \arg\min_{\theta \in \mathbb{R}^d} L(\{(X_\ell, Y_\ell + Z_\ell)\}_{\ell=1}^{t-1}; \theta)\,, \qquad (6)$$

where $Z_\ell \sim \mathcal{N}(0, a^2)$ are normal random variables that are resampled in each round, independently of each other and the history, and $a > 0$ is a tunable parameter. Surprisingly, this perturbation does not change the parameter estimation problem. In particular, it only shifts the gradient of the log likelihood in (1) by $Z_\ell X_\ell$ and the Hessian in (2) remains positive semi-definite. In this work, we show that GLM-FPL has $\tilde{O}(d\sqrt{n \log K})$ regret when the number of arms is K, under an additional assumption on arm features.

The design of GLM-FPL is motivated by the equivalence of posterior sampling and perturbations by Gaussian noise in linear models [Lu and Van Roy, 2017], when the prior of θ_* and rewards are Gaussian. In GLMs, these two are not equivalent. Thus GLM-TSL and GLM-FPL are different algorithms. GLM-FPL can be also viewed as an instance of randomized least-squares value iteration [Osband et al., 2016] applied to bandits. The specific instance in this work, additive Gaussian noise in a GLM, is novel. Finally, we note that the perturbation in (6) can be directly applied to more complex models, such as neural networks (Section 5). This is arguably its most attractive property.

3.3 Computationally-Efficient Implementations

The MLEs in (4) and (6) can be computed by *iteratively reweighted least squares (IRLS)* [Wolke and Schwetlick, 1988], which uses Newton's method. Roughly speaking, each step of IRLS multiplies the inverse of (2) and (1). If (2) and (1) can be expressed independently of round t, the computational cost of an IRLS step does not increase with t. This is viable for any set of feature vectors $\mathcal{X}$ using

$$\textstyle\sum_{x \in \mathcal{X}} (N_x \mu(x^T \theta) - Y_x) x\,, \quad \sum_{x \in \mathcal{X}} N_x \dot{\mu}(x^T \theta) x x^T\,,$$

where N_x is the number of times that x appears in history $\mathcal{D}$, and Y_x is the sum of its rewards. Both N_x and Y_x can be updated incrementally. Finally, adding $\mathcal{N}(0, a^2)$ noise to each reward in (6) is equivalent to adding $\mathcal{N}(0, N_x a^2)$ noise to each Y_x above.

The pulled arm in line 5 of Algorithm 1 can be computed efficiently even when the arm space is infinite, such as an intersection of half spaces. This is true of prior GLM bandit algorithms (Section 6). The MLE in line 4 cannot be computed efficiently in general, independently of round t, as in all prior algorithms except that of Jun et al. [2017]. We study one approximation empirically in Section 5.2.

4 Analysis

Our analysis is organized as follows. In Section 4.1, we review technical challenges that arise in analyzing GLM bandits and their solutions. In Section 4.2, we outline our analysis. In Sections 4.3 and 4.4, we prove regret bounds for GLM-TSL and GLM-FPL. We discuss them in Section 4.5.

4.1 Technical Challenges

One challenge in analyzing GLMs is that they do not have closed-form solutions. Nevertheless, their solutions can be expressed using the gradient and Hessian of the log likelihood (Section 2). This is the key idea in the analyses of GLM bandits [Filippi et al., 2010, Li et al., 2017] and we present it below.

Lemma 1. *Let $\mathcal{D}_1 = \{(x_\ell, y_{\ell,1})\}_{\ell=1}^n$ be a set of n observations and $\mathcal{D}_2 = \{(x_\ell, y_{\ell,2})\}_{\ell=1}^n$ have the same features as $\mathcal{D}_1$. Let θ_1 be the minimizer of $L(\mathcal{D}_1; \theta)$ and θ_2 be the minimizer of $L(\mathcal{D}_2; \theta)$. Then*

$$\sum_{\ell=1}^n (y_{\ell,2} - y_{\ell,1}) x_\ell = \nabla^2 L(\mathcal{D}_1; \theta')(\theta_2 - \theta_1)\,,$$

where $\theta' = \alpha \theta_1 + (1 - \alpha)\theta_2$ for some $\alpha \in [0, 1]$.

Proof. By the definition of the gradient in (1),

$$\nabla L(\mathcal{D}_1; \theta) - \nabla L(\mathcal{D}_2; \theta) = \sum_{\ell=1}^n (y_{\ell,2} - y_{\ell,1}) x_\ell$$

holds for any θ. Moreover, from the definitions of θ_1 and θ_2, $\nabla L(\mathcal{D}_1; \theta_1) = \nabla L(\mathcal{D}_2; \theta_2) = \mathbf{0}$. Now we apply these identities and obtain

$$\begin{aligned} \sum_{\ell=1}^n (y_{\ell,2} - y_{\ell,1}) x_\ell &= \nabla L(\mathcal{D}_1; \theta_2) - \nabla L(\mathcal{D}_2; \theta_2) \\ &= \nabla L(\mathcal{D}_1; \theta_2) - \nabla L(\mathcal{D}_1; \theta_1) \\ &= \nabla^2 L(\mathcal{D}_1; \theta')(\theta_2 - \theta_1)\,. \end{aligned}$$

where θ' is defined in the claim. □

Another challenge is $\dot{\mu}(x_\ell^\top \theta)$ in (2). To apply ideas from linear bandit analyses, it must be eliminated. We do so as follows. Let $G = \sum_{\ell=1}^{|\mathcal{D}|} x_\ell x_\ell^\top$ be an *unweighted Hessian* with the same features as (2). Let $c_{\min} \leq \dot{\mu}(x_\ell^\top \theta) \leq c_{\max}$ for some $c_{\min}$ and $c_{\max}$, and for all $\ell \in [|\mathcal{D}|]$. Then from the definition of (2), $c_{\min} G \preceq \nabla^2 L(\mathcal{D}; \theta) \preceq c_{\max} G$ and $c_{\min}^{-1} G^{-1} \succeq (\nabla^2 L(\mathcal{D}; \theta))^{-1} \succeq c_{\max}^{-1} G^{-1}$. Because of this, the derivatives of μ must be controlled.

To control the derivatives of μ at $\bar{\theta}_t$ and $\tilde{\theta}_t$ (Section 3), we initially explore so that $\bar{\theta}_t$ and $\tilde{\theta}_t$ are in the unit ball centered at θ_* with a high probability. This gives rise to

$$\dot{\mu}_{\min} = \min_{\|x\|_2 \leq 1,\, \|\theta - \theta_*\|_2 \leq 1} \dot{\mu}(x^\top \theta)$$

in our regret bounds, the *minimum derivative of* μ in the unit ball centered at θ_*. This trick [Li et al., 2017] requires that $\|x_i\|_2 \leq 1$ for all arms i, and we assume this in our analysis. We define the *maximum derivative of* μ as

$$\dot{\mu}_{\max} = \max_{\|x\|_2 \leq 1,\, \theta \in \mathbb{R}^d} \dot{\mu}(x^\top \theta)\,.$$

This factor is typically easy to control. In logistic regression, for instance, $\dot{\mu}_{\max} = 1/4$.

4.2 Outline of Our Analyses

Let θ_* be the unknown parameter vector, $\bar{\theta}_t$ be its MLE in round t, and $\tilde{\theta}_t$ be the randomized MLE in round t. At a high level, we bound the regret under assumptions that $\bar{\theta}_t \to \theta_*$, $\tilde{\theta}_t \to \bar{\theta}_t$, and $\tilde{\theta}_t$ is optimistic. We show that the corresponding favorable conditions hold with a high probability and define the corresponding events below.

Let $\mathcal{F}_t = \sigma(I_1, \dots, I_t, Y_1, \dots, Y_t)$ be the σ-algebra generated by the pulled arms and their rewards by the end of round $t \in [n]$. We let $\mathcal{F}_0 = \{\emptyset, \Omega\}$, where Ω is the sample space of the probability space that holds all random variables. Then $(\mathcal{F}_t)_t$ is a filtration. Let

$$\mathbb{P}_t\left(\cdot\right) = \mathbb{P}\left(\cdot \,|\, \mathcal{F}_{t-1}\right)\,, \quad \mathbb{E}_t\left[\cdot\right] = \mathbb{E}\left[\cdot \,|\, \mathcal{F}_{t-1}\right]\,,$$

be the conditional probability and expectation, given the history at the beginning of round t, $\mathcal{F}_{t-1}$, respectively. Let $G_t = \sum_{\ell=1}^{t-1} X_\ell X_\ell^\top$ be the *unweighted Hessian* in round t and $\Delta_{\max} = \max_{i \in [K]} \Delta_i$ be the maximum regret.

To argue that $\bar{\theta}_t \to \theta_*$, we define

$$E_{1,t} = \left\{\forall i \in [K] : \left|x_i^\top \bar{\theta}_t - x_i^\top \theta_*\right| \leq c_1 \|x_i\|_{G_t^{-1}}\right\}\,, \quad (7)$$

the event that $x_i^\top \bar{\theta}_t$ and $x_i^\top \theta_*$ are "close" for all arms i in round t, where $c_1 > 0$ is tuned later such that event $E_{1,t}$ is likely. Specifically, let $\bar{E}_{1,t}$ be the complement of $E_{1,t}$. Then we set c_1 such that $\mathbb{P}\left(\bar{E}_{1,t}\right) = O(1/n)$.

The upper bound on $\mathbb{P}\left(\bar{E}_{1,t}\right)$ is motivated by Lemma 3 in Li et al. [2017]. We reprove the lemma since it contains a subtle error. In particular, the proof that $\|\bar{\theta}_t - \theta_*\|_2 \leq 1$ holds with a high probability assumes that the agent does not act adaptively up to round t, which it clearly *does* for any $t > \tau$.

To argue that $\tilde{\theta}_t \to \bar{\theta}_t$, we define

$$E_{2,t} = \left\{\forall i \in [K] : \left|x_i^\top \tilde{\theta}_t - x_i^\top \bar{\theta}_t\right| \leq c_2 \|x_i\|_{G_t^{-1}}\right\}\,, \quad (8)$$

the event that $x_i^\top \tilde{\theta}_t$ and $x_i^\top \bar{\theta}_t$ are "close" for all arms i in round t, where $c_2 > 0$ is tuned later such that event $E_{2,t}$ is likely given any past. Specifically, let $\bar{E}_{2,t}$ be the complement of $E_{2,t}$. Then we set c_2 such that $\mathbb{P}_t\left(\bar{E}_{2,t}\right) = O(1/n)$. This part of the analysis relies on the properties of our perturbations and is novel.

Finally, to argue that $\tilde{\theta}_t$ is sufficiently optimistic given any past, we define event

$$E_{3,t} = \left\{x_1^\top \tilde{\theta}_t - x_1^\top \bar{\theta}_t > c_1 \|x_1\|_{G_t^{-1}}\right\}\,. \quad (9)$$

To obtain $\mathbb{P}_t\left(E_{3,t}\right) = O(1)$, we set parameter a in (4) and (6). This part of the analysis relies on the properties of our perturbations and is novel.

Our analysis is sufficiently general, so that it can be used to analyze different randomized algorithms. To show this, we use it to analyze both `GLM-TSL` and `GLM-FPL`. The central part of the analysis is an upper bound on the expected per-round regret of any randomized algorithm whose perturbed solution in round t is a function of its history. The corresponding lemma is stated below.

Lemma 2. *Let $p_2 \geq \mathbb{P}_t\left(\bar{E}_{2,t}\right)$, $p_3 \leq \mathbb{P}_t\left(E_{3,t}\right)$, and $p_3 > p_2$. Then on event $E_{1,t}$,*

$$\mathbb{E}_t\left[\Delta_{I_t}\right] \leq \dot{\mu}_{\max}(c_1 + c_2)\left(1 + \frac{2}{p_3 - p_2}\right) \times \mathbb{E}_t\left[\|x_{I_t}\|_{G_t^{-1}}\right] + \Delta_{\max}\, p_2\,.$$

The hardest part in the analyses of `GLM-TSL` and `GLM-FPL` is to bound p_2 and p_3 in Lemma 2.

4.3 Analysis of `GLM-TSL`

Now we are ready to analyze `GLM-TSL` and `GLM-FPL`. The regret bound of `GLM-TSL` is stated below.

Theorem 3. *The n-round regret of* `GLM-TSL` *is bounded as*

$$R(n) \leq \dot{\mu}_{\max}(c_1 + c_2)\left(1 + \frac{2}{0.15 - 1/n}\right) \times \sqrt{2dn\log(2n/d)} + (\tau + 3)\Delta_{\max}\,,$$

where

$$a = c_1\sqrt{\dot{\mu}_{\max}}\,,$$
$$c_1 = \sigma\dot{\mu}_{\min}^{-1}\sqrt{d\log(n/d) + 2\log n}\,,$$
$$c_2 = c_1\sqrt{2\dot{\mu}_{\min}^{-1}\,\dot{\mu}_{\max}\log(Kn)}\,,$$

and the number of exploration rounds τ satisfies

$$\lambda_{\min}(G_\tau) \geq \max\left\{\sigma^2\dot{\mu}_{\min}^{-2}(d\log(n/d) + 2\log n),\, 1\right\}\,.$$

Proof. The claim is proved in Appendix A.

The proof has three key steps. First, we bound the probability of event $\bar{E}_{1,t}$ from above (Lemma 8 in Appendix B). Second, we choose parameter a such that the probabilities of events $\bar{E}_{2,t}$ and $E_{3,t}$ are bounded for any history $\mathcal{F}_{t-1}$ (Lemma 4). Finally, we set the number of initial exploration rounds τ such that $\|\bar{\theta}_t - \theta_*\|_2 \leq 1$ is likely in any round $t \geq \tau$ (Lemma 9 in Appendix B). □

The above regret bound is $\tilde{O}(d\sqrt{n \log K})$. We derive the key concentration and anti-concentration lemma below.

Lemma 4. *Let*

$$a = c_1 \sqrt{\dot{\mu}_{\max}}\,, \quad c_2 = c_1 \sqrt{2 \dot{\mu}_{\min}^{-1} \, \dot{\mu}_{\max} \log(Kn)}\,.$$

Let $E = \left\{\|\bar{\theta}_t - \theta_*\|_2 \leq 1\right\}$. Then $\mathbb{P}_t\left(\bar{E}_{2,t}\right) \leq 1/n$ holds on event E and $\mathbb{P}_t\left(E_{3,t}\right) \geq 0.15$.

Proof. By the design of `GLM-TSL` in (4),

$$x^\top \tilde{\theta}_t - x^\top \bar{\theta}_t \sim \mathcal{N}(0, a^2 \|x\|_{H_t^{-1}}^2)$$

for any vector $x \in \mathbb{R}^d$, where matrix H_t is defined in (5). Let $U = x^\top \tilde{\theta}_t - x^\top \bar{\theta}_t$. Because $U \sim \mathcal{N}(0, a^2 \|x\|_{H_t^{-1}}^2)$ is a normal random variable, we have that

$$\mathbb{P}_t\left(U \geq a \|x\|_{H_t^{-1}}\right) \geq 0.15\,,$$
$$\mathbb{P}_t\left(U \geq c \|x\|_{H_t^{-1}}\right) \leq \exp\left[-\frac{c^2}{2a^2}\right]\,,$$

for any $c > 0$.

Now note that $H_t \preceq \dot{\mu}_{\max} G_t$. As a result,

$$\begin{aligned} 0.15 &\leq \mathbb{P}_t\left(U \geq a \|x\|_{H_t^{-1}}\right) \\ &\leq \mathbb{P}_t\left(U \geq a \sqrt{\dot{\mu}_{\max}^{-1}} \|x\|_{G_t^{-1}}\right)\,. \end{aligned}$$

For $a = c_1 \sqrt{\dot{\mu}_{\max}}$ and $x = x_1$, we get that event $E_{3,t}$ in (9) occurs with probability at least 0.15.

Moreover, $H_t \succeq \dot{\mu}_{\min} G_t$ on event E, which yields

$$\begin{aligned} \exp\left[-\frac{c^2}{2a^2}\right] &\geq \mathbb{P}_t\left(U \geq c \|x\|_{H_t^{-1}}\right) \\ &\geq \mathbb{P}_t\left(U \geq c \sqrt{\dot{\mu}_{\min}^{-1}} \|x\|_{G_t^{-1}}\right)\,. \end{aligned}$$

For $c = a\sqrt{2 \log(Kn)}$, $x = x_i$, and by the union bound over all K arms, we get that event $\bar{E}_{2,t}$ in (8) occurs with probability at most $1/n$. □

4.4 Analysis of `GLM-FPL`

The regret bound of `GLM-FPL` is stated below. The analysis assumes that all feature vectors x_i have at most one non-zero entry. This assumption is discussed in Section 4.5.

Theorem 5. *The n-round regret of* `GLM-FPL` *is bounded as*

$$R(n) \leq \dot{\mu}_{\max}(c_1 + c_2)\left(1 + \frac{2}{0.15 - 2/n}\right) \times \sqrt{2dn \log(2n/d)} + (\tau + 4)\Delta_{\max}\,,$$

where

$$\begin{aligned} a &= c_1 \dot{\mu}_{\max}\,, \\ c_1 &= \sigma \dot{\mu}_{\min}^{-1} \sqrt{d \log(n/d) + 2 \log n}\,, \\ c_2 &= c_1 \dot{\mu}_{\min}^{-1} \, \dot{\mu}_{\max} \sqrt{2 \log(Kn)}\,, \end{aligned}$$

and the number of exploration rounds τ satisfies

$$\lambda_{\min}(G_\tau) \geq \max\{4\sigma^2 \dot{\mu}_{\min}^{-2}(d \log(n/d) + 2 \log n), 8a^2 \dot{\mu}_{\min}^{-2} \log n,\, 1\}\,.$$

Proof. The claim is proved in Appendix A.

The proof has three key steps. First, we bound the probability of event $\bar{E}_{1,t}$ from above (Lemma 8 in Appendix B). Second, we choose parameter a such that the probabilities of events $\bar{E}_{2,t}$ and $E_{3,t}$ are bounded for any history $\mathcal{F}_{t-1}$ (Lemma 6). Finally, we set the number of initial exploration rounds τ such that $\|\bar{\theta}_t - \theta_*\|_2 \leq 1/2$ is likely and $\|\tilde{\theta}_t - \theta_*\|_2 \leq 1$ is conditionally likely given $\mathcal{F}_{t-1}$, in any round $t \geq \tau$ (Lemma 10 in Appendix B). □

The above regret bound is also $\tilde{O}(d\sqrt{n \log K})$. The key concentration and anti-concentration lemma follows.

Lemma 6. *Let*

$$a = c_1 \dot{\mu}_{\max}\,, \quad c_2 = c_1 \dot{\mu}_{\min}^{-1} \, \dot{\mu}_{\max} \sqrt{2 \log(Kn)}\,.$$

Let $E = \left\{\|\bar{\theta}_t - \theta_*\|_2 \leq 1/2\right\}$, $E' = \left\{\|\tilde{\theta}_t - \theta_*\|_2 \leq 1\right\}$, and $\mathbb{P}_t\left(\bar{E}'\right) \leq 1/n$ on event E. Then $\mathbb{P}_t\left(\bar{E}_{2,t}\right) \leq 2/n$ on event E and $\mathbb{P}_t\left(E_{3,t}\right) \geq 0.15$.

Proof. Fix any history $\mathcal{F}_{t-1}$. By Lemma 1, where $\mathcal{D}_1 = \{(X_\ell, Y_\ell)\}_{\ell=1}^{t-1}$ and $\mathcal{D}_2 = \{(X_\ell, Y_\ell + Z_\ell)\}_{\ell=1}^{t-1}$, we get

$$\sum_{\ell=1}^{t-1} Z_\ell X_\ell = \tilde{H}_t(\tilde{\theta}_t - \bar{\theta}_t)\,,$$

where $Z_\ell \in \mathcal{N}(0, a^2)$ are i.i.d. normal random variables,

$$\tilde{H}_t = \sum_{\ell=1}^{t-1} \dot{\mu}(X_\ell^\top \theta'_t) X_\ell X_\ell^\top\,,$$

and $\theta'_t = \alpha \bar{\theta}_t + (1 - \alpha)\tilde{\theta}_t$ for some $\alpha \in [0, 1]$. Fix any $x \in \mathbb{R}^d$ and let $U = x^\top G_t^{-1} \sum_{\ell=1}^{t-1} Z_\ell X_\ell$. Then

$$x^\top G_t^{-1} \tilde{H}_t(\tilde{\theta}_t - \bar{\theta}_t) = U \sim \mathcal{N}(0, a^2 \|x\|_{G_t^{-1}}^2)\,.$$

Since U is a normal random variable, we have that

$$\mathbb{P}_t\left(U \geq a \|x\|_{G_t^{-1}}\right) \geq 0.15\,,$$
$$\mathbb{P}_t\left(U \geq c \|x\|_{G_t^{-1}}\right) \leq \exp\left[-\frac{c^2}{2a^2}\right]\,,$$

for any $c > 0$.

Since all feature vectors have at most one non-zero entry, G_t^{-1} and $\tilde{H}_t$ are diagonal, as is $G_t^{-1}\tilde{H}_t$. By the definitions of G_t and $\tilde{H}_t$, diagonal entries of $G_t^{-1}\tilde{H}_t$ are non-negative and at most $\dot{\mu}_{\max}$. Let x have at most one non-zero entry. Then $x^\top(\tilde{\theta}_t - \bar{\theta}_t)$ and $x^\top G_t^{-1}\tilde{H}_t(\tilde{\theta}_t - \bar{\theta}_t)$ have the same sign, which we use to derive

$$\begin{aligned}0.15 &\leq \mathbb{P}_t\left(U \geq a\|x\|_{G_t^{-1}}\right)\\ &\leq \mathbb{P}_t\left(\dot{\mu}_{\max}\, x^\top(\tilde{\theta}_t - \bar{\theta}_t) \geq a\|x\|_{G_t^{-1}}\right)\\ &= \mathbb{P}_t\left(x^\top(\tilde{\theta}_t - \bar{\theta}_t) \geq a\dot{\mu}_{\max}^{-1}\|x\|_{G_t^{-1}}\right) .\end{aligned}$$

For $a = c_1\dot{\mu}_{\max}$ and $x = x_1$, we get that event $E_{3,t}$ in (9) occurs with probability at least 0.15.

The diagonal entries of $G_t^{-1}\tilde{H}_t$ are non-negative, and also at least $\dot{\mu}_{\min}$ on events E and E'. So, on event E,

$$\begin{aligned}\exp\left[-\frac{c^2}{2a^2}\right] &\geq \mathbb{P}_t\left(U \geq c\|x\|_{G_t^{-1}}\right)\\ &\geq \mathbb{P}_t\left(U \geq c\|x\|_{G_t^{-1}},\, E' \text{ occurs}\right)\\ &\geq \mathbb{P}_t\left(\dot{\mu}_{\min}\, x^\top(\tilde{\theta}_t - \bar{\theta}_t) \geq c\|x\|_{G_t^{-1}}\right) - \frac{1}{n}\\ &= \mathbb{P}_t\left(x^\top(\tilde{\theta}_t - \bar{\theta}_t) \geq c\dot{\mu}_{\min}^{-1}\|x\|_{G_t^{-1}}\right) - \frac{1}{n} .\end{aligned}$$

For $c = a\sqrt{2\log(Kn)}$, $x = x_i$, and by the union bound over all K arms, we get that event $\bar{E}_{2,t}$ in (8) occurs with probability at most $2/n$. □

4.5 Discussion

The regret of GLM-TSL is $\tilde{O}(d\sqrt{n\log K})$ (Theorem 3). Up to the factor of $\sqrt{\log K}$, this matches the gap-free bounds of GLM-UCB [Filippi et al., 2010] and UCB-GLM [Li et al., 2017]. As in Agrawal and Goyal [2013b], the key idea in our analysis is to achieve optimism by inflating the covariance matrix in GLM-TSL by $a = O(\sqrt{d\log n})$. This setting is too conservative in practice. Thus, in Section 5, we also experiment with $a = O(1)$, which is known to work well in practice [Chapelle and Li, 2012, Russo et al., 2018].

The regret of GLM-FPL is $\tilde{O}(d\sqrt{n\log K})$ (Theorem 5). Although the bound scales with K, d, and n similarly to that in Theorem 3, it is worse in constant factors. For instance, c_2 is additionally multiplied by $\sqrt{\dot{\mu}_{\min}^{-1}\,\dot{\mu}_{\max}}$. The number of initial exploration rounds is also higher, since we need to guarantee that $\tilde{\theta}_t$ and θ_* are close with a high probability given any $\mathcal{F}_{t-1}$. As in GLM-TSL, the suggested value of $a = O(\sqrt{d\log n})$ is too conservative in practice. Thus, we also experiment with $a = O(1)$ in Section 5.

The regret bound of GLM-FPL is proved under the assumption that feature vectors have at most one non-zero entry. We need this assumption for the following reason. We establish in Lemma 6 that

$$U = x^\top G_t^{-1}\tilde{H}_t(\tilde{\theta}_t - \bar{\theta}_t) \sim \mathcal{N}(0, a^2\|x\|_{G_t^{-1}}^2) .$$

Since $a\|x\|_{G_t^{-1}}$ is one standard deviation of U, event $U > a\|x\|_{G_t^{-1}}$ is likely. But we need event $U' = x^\top(\tilde{\theta}_t - \bar{\theta}_t) > a\|x\|_{G_t^{-1}}$ to be likely. If G_t^{-1} and $\tilde{H}_t$ have different eigenvectors, U and U' can have different signs, and it is hard to relate them due to potential rotations by $G_t^{-1}\tilde{H}_t$. Our assumption guarantees that the eigenvectors of G_t^{-1} and $\tilde{H}_t$ are identical. We leave the elimination of this assumption for future work.

The initial exploration in GLM-TSL and GLM-FPL can be implemented as follows. Let $\{v_i\}_{i=1}^d \subseteq \{x_i\}_{i=1}^K$ be any basis in $\mathbb{R}^d$ and $M = \sum_{i=1}^d v_i^\top v_i$. Then, to satisfy assumptions $\lambda_{\min}(G_\tau) \geq C$ in Theorems 3 and 5, each arm in the basis is pulled $C\lambda_{\min}^{-1}(M)$ times.

5 Experiments

We conduct two sets of experiments. In Section 5.1, we assess the empirical regret of GLM-TSL and GLM-FPL in logistic bandits. Because of its simplicity and generality, the perturbation mechanism in GLM-FPL can be easily applied to more complex models. We assess it on contextual bandit problems with neural networks in Section 5.2.

5.1 Logistic Bandit

The goal of this experiment is to show that our proposed algorithms perform well. We experiment with a *logistic bandit*, a GLM bandit where $\mu(v) = 1/(1 + \exp[-v])$ and $Y_{i,t} \sim \mathrm{Ber}(\mu(x_i^\top\theta_*))$. The number of arms is $K = 100$. To avoid bias in choosing problem instances, we generate them randomly: the feature vector of arm i is drawn uniformly at random from $[-1, 1]^d$ and the parameter vector is $\theta_* \sim \mathcal{N}(\mathbf{0}, 3d^{-2}I_d)$, where I_d is a $d \times d$ identity matrix. By design, $\mathrm{var}\left[x_i^\top\theta_*\right] = 1$, and so $x_i^\top\theta_* \in [-4, 4]$ with a high probability. We vary the number of features d from 5 to 20. The horizon is $n = 50\,000$ rounds and our results are averaged over 100 problem instances.

Our baselines are two UCB algorithms, GLM-UCB [Filippi et al., 2010] and UCB-GLM [Li et al., 2017]. We experiment with two designs for each evaluated algorithm, *theory* (as analyzed) and *informal* (practical). For GLM-TSL, we use a from Theorem 3 and $a = 1$, for which (4) reduces to sampling from the Laplace approximation. For GLM-FPL, we use a from Theorem 5 and $a = 0.5$. We choose the latter since a in Theorem 5 is half that in Theorem 3 in logistic models, since $\dot{\mu}_{\max} = 0.25$. We also implement GLM-UCB and UCB-GLM with tighter confidence intervals, $0.5\|x\|_{G^{-1}}$, where x is the feature vector of the arm, G is the sample

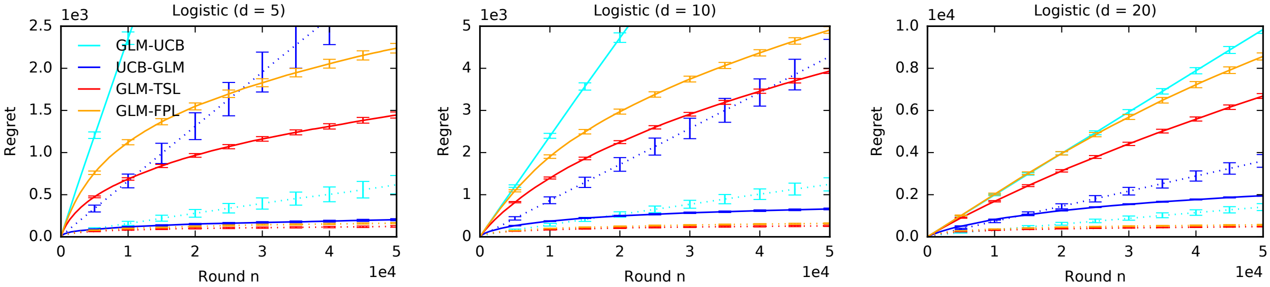

Figure 1: Evaluation of GLM-TSL and GLM-FPL in logistic bandits. The n-round regret is shown as a function of n. The solid and dotted lines represented theory-suggested and informal designs, respectively.

covariance matrix, and 0.5 is the maximum standard deviation of rewards in logistic models. All algorithms pull d linearly independent arms initially and $\dot{\mu}_{\min}$ is set to the most optimistic value of 0.25.

Our results are shown in Figure 1. We observe that theory GLM-TSL and GLM-FPL outperform theory GLM-UCB, but not theory UCB-GLM. The latter is known from prior algorithm designs. In particular, when LinTS [Agrawal and Goyal, 2013b] is implemented as analyzed, it fails to outperform LinUCB [Abbasi-Yadkori et al., 2011]; but it does outperform it when the theory-suggested posterior scaling is relaxed. This is indeed how LinTS is usually implemented. Informal GLM-UCB and UCB-GLM fail, and have linear regret in n. On the other hand, informal GLM-TSL and GLM-FPL have low regret, sublinear in n. We conclude that GLM-TSL and GLM-FPL have state-of-the-art performance in logistic bandits.

5.2 Deep Bandit

The second experiment is on contextual bandit problems, which are generated as follows. We fix a supervised learning dataset $\mathcal{D}$ and a target label c. The examples with label c have random rewards $\mathrm{Ber}(0.75)$ while the other examples have random rewards $\mathrm{Ber}(0.25)$. In round t, the agent is presented $K = 10$ random examples $x_{i,t}$ from $\mathcal{D}$, which are arms. The agent learns a single generalization model that maps feature vector $x_{i,t}$ to its expected reward. The goal of the agent is to learn a good mapping quickly. Since our generalization models are imperfect, our evaluation metric is the *average per-round reward* in n rounds, which we define as $\sum_{t=1}^{n} Y_t / n$.

We experiment with two large-scale datasets: MNIST and Fashion MNIST. *MNIST* [Lecun et al., 1998] is a dataset of 60 thousand 28×28 gray-scale images of handwritten digits, from 0 to 9. *Fashion MNIST* [Xiao et al., 2017] is a dataset of 60 thousand 28×28 gray-scale images in 10 fashion categories. We generate 500 bandit instances for each dataset, 50 for each class in that dataset. The horizon is $n = 10\,000$ rounds and we report the average reward over all instances in each dataset.

We implement GLM-FPL with the neural network generalization in Keras [Chollet et al., 2015]. The neural network has a single fully-connected hidden layer with 50 units. The output layer is a sigmoid. We experiment with both ReLU and tanh activation functions in the hidden layer. The output layer is a sigmoid. In each round, the model is updated using the adaptive optimizer Adam [Kingma and Ba, 2015], where the learning rate is 0.001 and the minibatch contains 32 most recent examples. These settings are default in Keras. Yogi [Zaheer et al., 2018] could be used instead of Adam. The rewards of the training examples are perturbed with i.i.d. $\mathcal{N}(0, a^2)$ noise where $a = 1$. We call this algorithm DeepFPL.

We consider two baselines. The first is a follow-the-leader variant of DeepFPL where $a = 0$. We call it DeepFL. The second is a variant of *Neural Linear*, the best method in a recent large empirical study [Riquelme et al., 2018]. This approach learns a representation separately of the bandit problem and applies an existing bandit algorithm to it. We learn the representation in m percent of initial rounds by exploring randomly. The representation is the same neural network as in DeepFPL. After learning, we chop its head off and use the rest to embed feature vectors. The bandit algorithm is GLM-FPL and we call this combined approach repGLM-FPL. We experiment with m from 1% to 20%.

Our results are reported in Figure 2. We observe three major trends. First, DeepFPL achieves high average rewards of at least 0.5, which is close to the theoretical optimum $0.25\,(1/K)^K + 0.75\,(1 - (1/K)^K) \approx 0.576$ in both our problems. Second, DeepFPL outperforms DeepFL. This shows that exploration is beneficial, since the only difference between DeepFPL and DeepFL is that DeepFPL perturbs rewards to explore. Third, DeepFPL outperforms all variants of repGLM-FPL. This shows that interleaving of representation learning and exploration is beneficial. Also note that the best setting of m in repGLM-FPL depends on the problem. For instance, at $n = 10\,000$ rounds, 1% and 5% exploration is comparable in the first two plots, while 5% exploration is superior in the last plot. DeepFPL does not need any such tunable parameter.

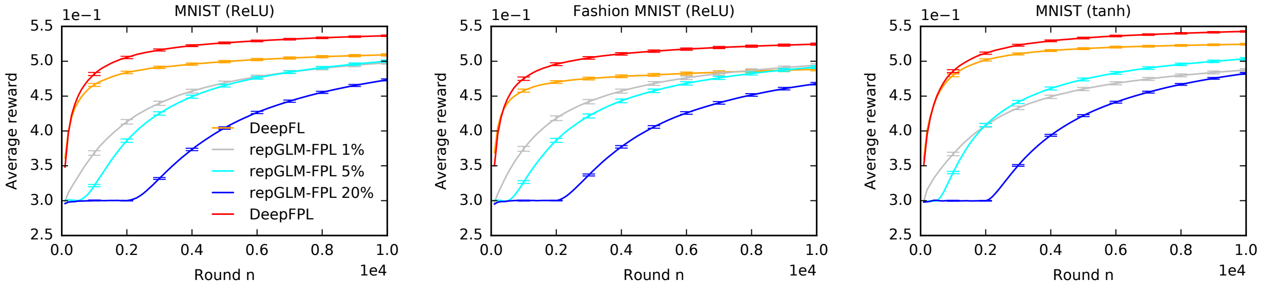

Figure 2: Evaluation of `DeepFPL` on contextual bandit problems in Section 5.2.

This experiment shows that `GLM-FPL` generalizes easily to complex models and works well. While it does not have regret guarantees in these models, it should be of interest to practitioners.

6 Related Work

In the infinite arm setting, Abeille and Lazaric [2017] proved that the regret of `GLM-TSL` is $\tilde{O}(d^{\frac{3}{2}}\sqrt{n})$. We prove that it is $\tilde{O}(d\sqrt{n \log K})$ when the number of arms is K. This is an improvement of $\sqrt{d / \log K}$ in our setting. We also match the result of Abeille and Lazaric [2017] in the infinite arm setting. Specifically, if the space of arms was discretized on an ε-grid, and this discretization would not change the order of the regret, the number of arms would be $K = \varepsilon^{-d}$ and $\sqrt{\log K} = \sqrt{d \log(1/\varepsilon)}$. Our analysis is different from Abeille and Lazaric [2017] and is more like that of Agrawal and Goyal [2013b]. We also match, up to the factor of $\sqrt{\log K}$, the bounds of most non-randomized GLM bandit algorithms [Filippi et al., 2010, Zhang et al., 2016, Li et al., 2017, Jun et al., 2017], which are $\tilde{O}(d\sqrt{n})$.

Dong et al. [2019] proved that the n-round Bayes regret of `GLM-TSL` is $\tilde{O}(d\sqrt{n})$. This bound is for a weaker performance metric than in this work, the Bayes regret; applies only to logistic bandits; and makes strong assumptions on the features of arms and θ_*. However, it does not depend on $\dot{\mu}_{\min}$, which is a significant advance.

Similarly to `GLM-TSL`, we prove that the regret of `GLM-FPL` is $\tilde{O}(d\sqrt{n \log K})$. This regret bound is under the assumption that feature vectors have at most one non-zero entry. Although limited, this result is non-trivial since the number of potentially optimal arms is $2d$, two per dimension. This is the first frequentist regret bound for exploration by Gaussian noise perturbations in a non-linear model. The good empirical performance of `GLM-FPL` (Section 5) suggests that the regret bound should hold in general, and we leave the more general analysis as future work.

`GLM-TSL` is a variant of Thompson sampling. Thompson sampling [Thompson, 1933, Agrawal and Goyal, 2013a, Russo et al., 2018] is relatively well understood in linear bandits [Agrawal and Goyal, 2013b, Valko et al., 2014]. However, it is difficult to extend it to non-linear problems because their posterior distributions are complex and have to be approximated. In general, posterior approximations in bandits are computationally costly and lack regret guarantees [Gopalan et al., 2014, Kawale et al., 2015, Lu and Van Roy, 2017, Riquelme et al., 2018, Lipton et al., 2018, Liu et al., 2018]. We provide guarantees in this work.

`GLM-FPL` is a follow-the-perturbed-leader algorithm [Hannan, 1957, Kalai and Vempala, 2005]. We can also view it as randomized least-squares value iteration [Osband et al., 2016] applied to bandits. Our instance, additive Gaussian noise in a GLM, is novel. `GLM-FPL` is also closely related to perturbed-history exploration [Kveton et al., 2019c,a,b]. Kveton et al. [2019b] proposed a logistic bandit algorithm that explores by perturbing its history with Bernoulli noise. This algorithm was not analyzed and is less general than `GLM-FPL`, as it is only for logistic bandits.

7 Conclusions

We study two randomized algorithms for GLM bandits, `GLM-TSL` and `GLM-FPL`. The key idea in both algorithms is to explore by perturbing the maximum likelihood estimate in round t. We analyze `GLM-TSL` and `GLM-FPL`, and prove that their n-round regret is $\tilde{O}(d\sqrt{n \log K})$. Both `GLM-TSL` and `GLM-FPL` perform well empirically in logistic bandits. `GLM-FPL` can be easily generalized to more complex problems. Our experiments with neural networks are very encouraging, and indicate that `GLM-FPL` can be analyzed beyond GLM bandits. We plan to conduct such analyses in future work.

Our analysis is under the assumption that the feature vectors of arms are fixed and do not change over time. This assumption can be lifted. The only part of the proof that changes is that the number of initial exploration rounds τ after which $\lambda_{\min}(G_\tau)$ (Theorems 3 and 5) is large enough becomes a random variable. Li et al. [2017] analyzed this random variable and we can directly reuse their result.

References

Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, pages 2312–2320, 2011.

Marc Abeille and Alessandro Lazaric. Linear Thompson sampling revisited. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for Thompson sampling. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, pages 99–107, 2013a.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, pages 127–135, 2013b.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

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

Kani Chen, Inchi Hu, and Zhiliang Ying. 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.

Francois Chollet et al. Keras. https://keras.io, 2015.

Shi Dong, Tengyu Ma, and Benjamin Van Roy. On the performance of thompson sampling on logistic bandits. In *Proceedings of the 32nd Annual Conference on Learning Theory*, 2019.

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

Aditya Gopalan, Shie Mannor, and Yishay Mansour. Thompson sampling for complex online problems. In *Proceedings of the 31st International Conference on Machine Learning*, pages 100–108, 2014.

James Hannan. Approximation to Bayes risk in repeated play. In *Contributions to the Theory of Games*, volume 3, pages 97–140. Princeton University Press, Princeton, NJ, 1957.

Kwang-Sung Jun, Aniruddha Bhargava, Robert Nowak, and Rebecca Willett. Scalable generalized linear bandits: Online computation and hashing. In *Advances in Neural Information Processing Systems 30*, pages 98–108, 2017.

Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

Jaya Kawale, Hung Bui, Branislav Kveton, Long Tran-Thanh, and Sanjay Chawla. Efficient Thompson sampling for online matrix-factorization recommendation. In *Advances in Neural Information Processing Systems 28*, pages 1297–1305, 2015.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

Branislav Kveton, Csaba Szepesvari, Mohammad Ghavamzadeh, and Craig Boutilier. Perturbed-history exploration in stochastic multi-armed bandits. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019a.

Branislav Kveton, Csaba Szepesvari, Mohammad Ghavamzadeh, and Craig Boutilier. Perturbed-history exploration in stochastic linear bandits. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, 2019b.

Branislav Kveton, Csaba Szepesvari, Sharan Vaswani, Zheng Wen, Mohammad Ghavamzadeh, and Tor Lattimore. Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3601–3610, 2019c.

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

Tor Lattimore and Csaba Szepesvari. *Bandit Algorithms*. Cambridge University Press, 2019.

Yann Lecun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2071–2080, 2017.

Zachary Lipton, Xiujun Li, Jianfeng Gao, Lihong Li, Faisal Ahmed, and Li Deng. BBQ-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 5237–5244, 2018.

Bing Liu, Tong Yu, Ian Lane, and Ole Mengshoel. Customized nonlinear bandits for online response selection in neural conversation models. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 5245–5252, 2018.

Xiuyuan Lu and Benjamin Van Roy. Ensemble sampling. In *Advances in Neural Information Processing Systems 30*, pages 3258–3266, 2017.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall, 1989.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 2377–2386, 2016.

Carlos Riquelme, George Tucker, and Jasper Snoek. Deep Bayesian bandits showdown: An empirical comparison of Bayesian deep networks for Thompson sampling. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on Thompson sampling. *Foundations and Trends in Machine Learning*, 11 (1):1–96, 2018.

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

Michal Valko, Remi Munos, Branislav Kveton, and Tomas Kocak. Spectral bandits for smooth graph functions. In *Proceedings of the 31st International Conference on Machine Learning*, pages 46–54, 2014.

R. Wolke and H. Schwetlick. Iteratively reweighted least squares: Algorithms, convergence analysis, and numerical comparisons. *SIAM Journal on Scientific and Statistical Computing*, 9(5):907–921, 1988.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017. URL http://arxiv.org/abs/1708.07747.

Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In *Advances in Neural Information Processing Systems 31*, pages 9793–9803, 2018.

Lijun Zhang, Tianbao Yang, Rong Jin, Yichi Xiao, and Zhi-Hua Zhou. Online stochastic linear optimization under one-bit feedback. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 392–401, 2016.