Provably Optimal Algorithms for Generalized Linear Contextual Bandits

Chun-Ting Kuo
Department of Computer Science
National Chiao Tung University
jack041286.eic09g@nctu.edu.tw

1 Introduction

1.1 Introduction

Most theoretical analyses on contextual bandits so far are on linear bandits, so they want to design new methods on generalized linear contextual bandits. In this work, they proposed an upper confidence bound based algorithm for generalized linear contextual bandits, which achieves an $O(\sqrt{dT})$ optimal regret bound 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.

In this paper, they 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. 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.

1.2 Related Work

Contextual bandit problems are originally motivated by applications in clinical trials (Woodroofe, 1979).The doctor needs to decide, in a sequential manner, which of them to use based on the patient's profiles. 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. Logistic regression model based algorithms have been shown to have substantial improvements over linear models (Li et al., 2012). Logistic regression is a kind of Generalized linear model. 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 using GLMs perform well empirically (Li et al., 2012). Another line, EXP-type algorithms, which can be applied to almost any model classes (Auer et al., 2002b). EXP4 algorithm (Beygelzimer et al., 2010; Agarwal et al., 2014) give an $O(\sqrt{dKT})$ regret that is near-optimal.

So, the problem is, can we find an efficient algorithm to achieve the optimal convergence rate for generalized linear bandits? In this paper, they propose a GLM version of the UCB algorithm called SupCB-GLM that achieves a optimal regret over T rounds of order $O(\sqrt{dT})$.

Preprint. Under review.

2 Problem Formulation

2.1 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, $\{x_{t,a}|a \in [K]\} \subset R^d$, which $|x_{t,a}| \leq 1$. Each feature vector $x_{t,a}$ is associated with an unknown stochastic reward $y_{t,a} \in [0,1]$. In a generalized linear models(GLM), in which there is an unknown $\theta^* \in R^d$, and a fixed, strictly increasing link function $\mu : R \to R$ such that $E[Y|X] = \mu(X'\theta^*)$, where X is the chosen action's feature and Y the corresponding reward. The agent's goal is to maximize the cumulative expected rewards over T . We can define optimal reward at round t, by
$a_t^* = argmax_{a\in[K]}\ \mu(x'_{t,a}\theta^*)$. Then, the agent's total regret of following strategy π can be expressed as follows

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

Note that $R_T(\pi)$ is in general a random variable due to the possible randomness in π. Denote by $X_t = x_{t,a_t}$ $Y_t = y_{t,a_t}$, X_t is a random variable because the agent chooses current action based on previous rewards $Y_t \in [0,1]$. In 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. 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,

$$E[e^{\lambda\epsilon_t}] \leq e^{\lambda^2\sigma^2/2} \tag{2}$$

2.2 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 (including Gaussian, binomial, Poisson, gamma,...), and its density, parameterized by $\theta \in \Theta$, can be written as

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

where $\eta \in R^+$ is a known scale parameter; n, g and h are three normalization functions mapping from to .That is infinitely differentiable satisfying $\dot{m}(X\theta^*) = E[Y|X] = \mu(X'\theta^*)$ and $\ddot{m}(X\theta^*) = V(Y|X)$. 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

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

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

$$\hat{\theta} \in arg\max_{\theta\in\Theta}\sum_{t=1}^{n}[Y_tX_t'\theta - m(X_t'\theta)]$$

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

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

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

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

The assumption of (4) told us the condition of $\lambda_{min}(V_n)$. I think the condition of $\lambda_{min}(V_n)$ is neccessary for estimating generalized linear models. So, the Theorem 1 telling us that characterizes the behavior of MLE on every direction. It implies that $x'(\hat{\theta_n} - \theta^*)$ has a sub-Gaussian tail bound for any $x \in R^d$.

Proof of Theorem 1 :
The maximum-likelihood estimation can be written as the solution to the following equation

$$\sum_{i=1}^{n}(Y_i - \mu(X_i'\theta))X_i = 0$$

Define $G(\theta) := \sum_{i=1}^{n}(\mu(X_i'\theta) - \mu(X_i'\theta^*))X_i$, we have

$$G(\theta^*) = 0 \quad and \quad G(\hat{\theta}) = \sum_{i=1}^{n}\epsilon_i X_i$$

$$G(\theta_1) - G(\theta_2) = \left[\sum_{i=1}^{n}\mu(X_i'\theta)X_i X_i'\right](\theta_1 - \theta_2) := F(\theta)(\theta_1 - \theta_2)$$

since $\dot{\mu} > 0$ and $\lambda_{min}(V) > 0$, for any $\theta_1 \neq \theta_2$, we have

$$||G(\theta)||^2_{V^{-1}} = ||G(\theta) - G(\theta^*)||^2_{V^{-1}} = (\theta - \theta^*)'F(\theta)V^{-1}F(\theta)(\theta - \theta^*)$$

due to the fact $F(\theta) \geq \kappa_\eta V$

$$||G(\theta)||^2_{V^{-1}} \geq \kappa_\eta^2 \lambda_{min}(V)||\theta - \theta^*||^2$$

On the other hand, Lemma A of Chen et al. (1999) implies that

$$||G(\theta)||_{V^{-1}} \leq \kappa_\eta \eta \sqrt{\lambda_{min}(V)}$$

So that we can obtain and finish the proof

$$|x'(\hat{\theta_n} - \theta^*)| \leq \frac{3\sigma}{\kappa}\sqrt{log(1/\delta)}||x||_{V_n^{-1}}$$

We can know the condition of $\lambda_{min}(V_n)$ in Theorem 1, and it can satisfied under mild condition such as Proposition 1. Assume the mild condition $\lambda_{min}(V_n) \geq B$ in Proposition 1, 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, B^d. Furthermore, let $\sum := E[X_t X_t']$ 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*

$$n \geq (\frac{C_1\sqrt{d} + C_2\sqrt{log(1/\delta)}}{\lambda_{min}(\sum)})^2 + \frac{2B}{\lambda_{min}(\sum)}$$

The Proposition 1 tells us when it satisfied under the mild condition $\lambda_{min}(V_n) \geq B$. For simplicity, we will drop the subscript n when there is no ambiguity. Therefore, n is not important, V_n is denoted V and so on.

$$V_n = \sum_{t=1}^{n} X_t X_t' := V$$

3 Theoretical Analysis

In this section, we present two algorithms. While the first algorithm UCB-GLM is computationally more efficient, the second algorithm SupCB-GLM has a provable optimal regret bound .

3.1 Algorithm UCB-GLM

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_t 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.

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
$$\sum_{i=1}^{t-1} \left(Y_i - \mu\left(X'_i\theta\right)\right) X_i = 0$$
2. Choose $a_t = \text{argmax}_{a\in[K]} \left(X'_{t,a}\hat{\theta}_t + \alpha \left\|X_{t,a}\right\|_{V_t^{-1}}\right)$
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 τ and α. 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 is only needed to ensure is invertible. The feature vectors X_t depend on the previous rewards. Consequently, the rewards $\{Y_i, i \in [t]\}$ may not be independent given $\{X_i, i \in [t]\}$. Another parameter α is used to control the amount of exploration.The larger the α is, the more exploration will be used.

Theorem 2 *Fix any $\delta > 0$. There exists a universal constant $C > 0$, such that if we run UCB-GLM with then, with probability at least $1 - 2\delta$, the regret of the algorithm is upper bounded by*

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

The theorem 2 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. By choosing $\delta = \frac{1}{T}$ and using the fact that $R_T \leq T$, this high probability result implies a bound on the expected regret: $E[R_T] = (d\sqrt{T})$.

Proof of Theorem 2 :
To facilitate our proof of Theorem 2, we first present two technical lemmas

Lemma 1 *Let $\{X_t\}_{t=1}^{\infty}$ be the sequence in R^d satisfying $||X_t|| \leq 1$. Define $X_0 = 0$ 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 $n > 0$,*

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

Lemma 2 *Suppose $\lambda_{min}(V_{m+1}) \geq 1$. For $\delta \in [\frac{1}{T}, 1)$, define event*

$$\varepsilon_\triangle := \left\{ || \triangle_t ||_{V_t} \leq \frac{\sigma}{\kappa}\sqrt{\frac{d}{2} log(1 + 2t/d) + log(1/\delta)} \right\}$$

Then, event $\varepsilon_\triangle$ holds for all $t \geq \tau$ with probability at least $1 - \delta$.

Then, we can know the upper bound of $||X_t||_{V_t^{-1}}$ and $|| \triangle_t ||_{V_t}$ from Lemma 1 and Lemma 2.

Fix t and let $X_t^* = x_{t,a_t^*}$ and $\triangle t = \hat{\theta}_t - \theta^*$, where $a_t^* = argmax_{a \in [K]} \mu(x'_{t,a}\theta^*)$ is an optimal action at round t. The regret of algorithm UCB-GLM can be upper bounded as

$$R_T = \sum_{t=1}^{\tau} (\mu(\langle X_t^*, \theta^* \rangle) - \mu(\langle X_t, \theta^* \rangle)) + \sum_{t=\tau+1}^{T} (\mu(\langle X_t^*, \theta^* \rangle) - \mu(\langle X_t, \theta^* \rangle))$$

In exploration part, $(1 \sim \tau)$, we can get an unknown stochastic reward $y_{t,a} \in [0, 1]$ in each round, so the regret upper bound in τ round is τ

$$\sum_{t=1}^{\tau} (\mu(\langle X_t^*, \theta^* \rangle) - \mu(\langle X_t, \theta^* \rangle)) \leq \tau$$

In exploitation part, $(\tau + 1 \sim T)$, that μ is an increasing Lipschitz function with Lipschitz constant L_μ and the μ function is bounded between 0 and 1

$$\sum_{t=\tau+1}^{T} (\mu(\langle X_t^*, \theta^* \rangle) - \mu(\langle X_t, \theta^* \rangle)) \leq L_\mu \sum_{t=\tau+1}^{T} (\langle X_t^*, \theta^* \rangle - \langle X_t, \theta^* \rangle)$$

So, the regret over T can be written as

$$R_T \leq \tau + L_\mu \sum_{t=\tau+1}^{T} (\langle X_t^*, \theta^* \rangle - \langle X_t, \theta^* \rangle)$$

Because the selection of a_t in UCB-GLM implies

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

So, we can rewrite as

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

By Cauchy-Schwartz inequality,

$$\leq \alpha \left(||X_t||_{V_t^{-1}} - ||X_t^*||_{V_t - 1} \right) + ||X_t^* - X_t||_{V_t^{-1}} || \triangle_t ||_{V_t}$$

Now, We have the following two lemmas to bound $||X_t||_{V_t^{-1}}$, and $|| \triangle_t ||_{V_t}$, respectively. Their proofs are deferred to the appendix

By Lemma 2, We now choose $\alpha = \frac{\sigma}{\kappa} \sqrt{\frac{d}{2} log(1 + 2t/d) + log(1/\delta)}$. For all $t \geq \tau$, we know $|| \triangle_t ||_{V_t} \leq \alpha$.

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

Combining the above with Lemma 1 yield,

$$\sum_{t=\tau+1}^{T} (\langle X_t^*, \theta^* \rangle - \langle X_t, \theta^* \rangle) \leq 2\alpha \sqrt{2Tdlog(\frac{T}{d})} \leq \frac{2d\sigma}{\kappa} log(\frac{T}{d\delta}) \sqrt{T}$$

Finally, we can obtain

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

So, we can show an $\tilde{O}(d\sqrt{T})$ regret bound of UCB-GLM algorithm.

3.2 Algorithm SupCB-GLM

Although the algorithm UCB-GLM performs sufficiently well in practice, it is unclear whether it can achieve the optimal rates of $O(\sqrt{dTlogK})$, when K is fixed and small. The key technical difficulty in analyzing UCB-GLM is the dependence between samples. In order to create independent samples for linear contextual bandits, we propose another algorithm SupCB-GLM which uses algorithm CB-GLM as a sub-routine.

Algorithm 2: CB-GLM

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

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

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

2. $V_t = \sum_{i \in \Psi(t)} X_i X_i'$
3. **For** $a \in A,$ **do**

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

End For

Algorithm 3: SupCB-GLM

Input: tuning parameter $\alpha, \tau,$ the number of trials T.
Initialization: randomly choose $a_t \in [K]$ for $t \in [\tau]$
Set $S = [log_2 T],\ F = \{a1, ..., a\tau\}$ and $\Psi_0 = \Psi_1 = ... = \Psi_s = \varnothing$

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

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

2. While $a_t =$ Null,

a. Run CB-GLM with α and $\Psi_s \cup F$ to calculate $m_{t,a}^{(s)}$ and $w_{t,a}^{(s)}$ for all $\alpha \in A_s$.

b. If $w_{t,a}^{(s)} > 2^{-s}$ for some $a \in A_s$,

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

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

set $a_t = arg \max_{a \in A_s} m_{t,a}^{(s)},$ update $\Psi_s = \Psi_s \cup \{t\}$

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

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

$s \leftarrow s + 1$

End For

At stage s, we set the confidence level at stage s to be 2^{-s}. Let $w_{t,a}^{(s)} = \alpha ||x_{t,a}||$, which α is used to control the amount of exploration. At each round t, it will save trajectory into Ψ_s. If $w_{t,a}^{(s)} > 2^{-s}$, we need to do more exploration on $x_{t,a}$, and select the action we chose last step. Otherwise, $w_{t,a}^{(s)} \leq 2^{-s}$, we do exploitation under a set Ψ_s. But if $w_{t,a}^{(s)} \leq 1/\sqrt{T}$, we also do more exploration but need to choose the new action. That means the action we chose last step was wrong.

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

$$T_0 = \Omega\left(\frac{\sigma^2}{\kappa^4} max\left\{d^3, \frac{log(TK/\delta)}{d}\right\}\right) \tag{6}$$

the regret of the algorithm is bounded as

$$R_T \leq 45(\sigma L_\mu/\kappa)\sqrt{logT\ log(TK/\delta)}\sqrt{dT}$$

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

$$E[R_T] = O\left((logT)^{1.5}\sqrt{dTlogK}\right)$$

$Proof\ of\ Theorem\ 3$:
To facilitate our proof of Theorem 3, we first present three technical lemmas,

Lemma 3 *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.*

Lemma 4 *Fix $\delta > 0$. Choose in SupCB-GLM $\tau = \sqrt{dT}$ and $\alpha = \frac{3\sigma}{\kappa}\sqrt{2log(TK/\delta)}$. Suppose T satisfies condition (6). Define the following event:*

$$\varepsilon_X := \left\{|m_{t,a}^{(s)} - x'_{t,a}\theta^*| \leq w_{t,a}^{(s)}, \forall \in [\tau+1, T], s \in [S], a \in [K]\right\} \tag{7}$$

Then, event ε_X holds with probability as least $1-\delta$

Lemma 5 *Suppose that event ε_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*

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

Define $V_{s,t} = \sum_{t\in\Psi_s(T)} X_t X_t'$, then by Lemma 1,

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

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

$$|\Psi_s(T)| \leq 2^s\alpha(\delta)\sqrt{2d\ log(T/d)|\Psi_s(T)|}. \tag{8}$$

Let Ψ_0 be the collection of trial 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, $\{\tau, \tau+1, ..., T\} = \Psi_0 \cup (\cup_{s=1}^S \Psi_s(T))$.
The regret over T can be written as,

$$R_T = \sum_{t=1}^{\tau}\left(\mu(x_{t,a_t^*}^*, \theta^*) - \mu(x_{t,a_t}^*, \theta^*)\right) + \sum_{t=\tau+1}^{T}\left(\mu(x_{t,a_t^*}^*, \theta^*) - \mu(x_{t,a_t}^*, \theta^*)\right)$$

In exploration part, $(1 \sim \tau)$, we can get an unknown stochastic reward $y_{t,a} \in [0,1]$ in each round. If we set $\tau = \sqrt{dT}$, we have

$$\sum_{t=1}^{\tau}\left(\mu(x_{t,a_t^*}^*, \theta^*) - \mu(x_{t,a_t}^*, \theta^*)\right) \leq \tau = \sqrt{dT}$$

In exploitation part, $(\tau+1 \sim T)$, assumption that $0 \leq \mu \leq 1$, and it can be separated two form, $t \in \Psi_0$ and $t \in \{\Psi_1, \Psi_2, ...\Psi_S\}$

$$\sum_{t=\tau+1}^{T} \Big(\mu(x^*_{t,a^*_t}, \theta^*) - \mu(x^*_{t,a_t}, \theta^*)\Big) \leq \sum_{t\in\Psi_0} \Big(\mu(x^*_{t,a^*_t}, \theta^*) - \mu(x^*_{t,a_t}, \theta^*)\Big) + \sum_{s=1}^{S} \sum_{t\in\Psi_s(T)} \Big(\mu(x^*_{t,a^*_t}, \theta^*) - \mu(x^*_{t,a_t}, \theta^*)\Big)$$

So, the regret over T can be written as,

$$R_T \leq \sqrt{dT} + \sum_{t\in\Psi_0} \Big(\mu(x^*_{t,a^*_t}, \theta^*) - \mu(x^*_{t,a_t}, \theta^*)\Big) + \sum_{s=1}^{S} \sum_{t\in\Psi_s(T)} \Big(\mu(x^*_{t,a^*_t}, \theta^*) - \mu(x^*_{t,a_t}, \theta^*)\Big)$$

By Lemma 5, if a_t selected by 2c, $t \in \Psi_0$

$$\sum_{t\in\Psi_0} \Big(\mu(x^*_{t,a^*_t}, \theta^*) - \mu(x^*_{t,a_t}, \theta^*)\Big) \leq T \cdot (\frac{2L_\mu}{\sqrt{T}})$$

By Lemma 5, if a_t selected by 2b, $t \in \Psi_s$

$$\sum_{s=1}^{S} \sum_{t\in\Psi_s(T)} \Big(\mu(x^*_{t,a^*_t}, \theta^*) - \mu(x^*_{t,a_t}, \theta^*)\Big) \leq \sum_{s=1}^{S} |\Psi_s(T)| \cdot (\frac{8L_\mu}{2^{s_t}})$$

Rewrite the regret,

$$R_T \leq \sqrt{dT} + T \cdot \frac{2L_\mu}{\sqrt{T}} + \sum_{s=1}^{S} L_\mu \cdot 2^{3-s} \cdot |\Psi_s(T)|$$

By the inequality (8),

$$R_T \leq \sqrt{dT} + 2L_\mu\sqrt{T} + 8L_\mu\alpha(\delta) \sum_{s=1}^{S} \sqrt{2d\ log(\frac{T}{d})|\Psi_s(T)|}$$

By Cauchy-Schwartz inequality,

$$R_T \leq \sqrt{dT} + 2L_\mu\sqrt{T} + 8L_\mu\alpha(\delta)\sqrt{2d\ log(\frac{T}{d})}\sqrt{ST}$$

$S = log_2 T$ and $\alpha = \frac{3\sigma}{\kappa}\sqrt{2log(TK/\delta)}$,

$$\begin{aligned} R_T &= \sqrt{dT} + 2L_\mu\sqrt{T} + 8L_\mu\frac{3\sigma}{\kappa}\sqrt{2log(TK/\delta)}\sqrt{2d\ log(\frac{T}{d})}\sqrt{Tlog_2 T} \\ &= \sqrt{dT} + 2L_\mu\sqrt{T} + \frac{48}{\sqrt{log2}}(\sigma L_\mu/\kappa)\sqrt{logT\ log(TK/\delta)log(\frac{T}{d})}\sqrt{dT} \\ &\leq 45(\sigma L_\mu/\kappa)\sqrt{logT\ log(TK/\delta)log(\frac{T}{d})}\sqrt{dT} \end{aligned}$$

with probability at least $1-\delta$. This completes the proof of the high-probability result.

This paper show that the regret bound is $45(\sigma L_\mu/\kappa)\sqrt{logT\ log(TK/\delta)log(\frac{T}{d})}\sqrt{dT}$, but I only can calculate the regret bound is $87.49(\sigma L_\mu/\kappa)\sqrt{logT\ log(TK/\delta)log(\frac{T}{d})}\sqrt{dT}$.
But it does not matter for this discussion of $\tilde{O}(\sqrt{dTlogK})$ regret bound

Theorem 3 demonstrates an $\tilde{O}(\sqrt{dTlogK})$ regret bound for the algorithm SupCB-GLM. It has been proved in (Chu et al. 2011) 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.

4 Conclusion

In this paper, they propose two algorithms. While the first algorithm UCB-GLM is computationally more efficient, the second algorithm SupCB-GLM has a provable optimal regret bound. I carefully go through their pseudo code and the proof of each theorem and lemma detailed. I think their SupCB-GLM method is really creativity and innovativity. The key technical difficulty in analyzing UCB-GLM is the dependence between samples but SupCB-GLM is independent samples. Therefore, they also explain why the feature vector can be independent of each other. Let $w_{t,a}^{(s)} = \alpha||x_{t,a}||$. If $w_{t,a}^{(s)} > 2^{-s}$, we need to do more exploration on $x_{t,a}$, otherwise, we choose action in a closed set Ψ_s. This idea is really creativity, and also achieve $\tilde{O}(\sqrt{dTlogK})$ optimal regret bound with high probability $1 - \delta$.

5 Extensions and Outlook

While SupCB-GLM algorithm has already achieved a provable optimal regret bound. In the future, we can add some assumption to improve the regret bound of UCB-GLM.

5.1 A better regret bound for UCB-GLM algorithm

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

$$\sum_{t=\tau+1}^{T} \lambda_{min}^{-1/2}(V_t) \leq c\sqrt{T} \tag{9}$$

holds with probability at least $1 - \delta$. Then, the regret of algorithm is bounded by

$$R_T \leq \frac{CL_\mu \sigma}{\kappa} \sqrt{dTlog(T/\delta)} \tag{10}$$

with probability at least $1 - 2\delta$, where C is a positive universal constant.
The difficulty lies in the condition in (9) is hard to check and may be violated in some cases. For example, when t is large enough, our estimator $\hat{\theta}_t$ is very close to θ^*. If we assume a positive gap between $\langle X_{t,a_t^*}, \theta^* \rangle$ and $\langle X_{t,a_t}, \theta^* \rangle$ for all $a \neq a_t^*$. It is unlikely the feature vector x_{t,a_t^*}lies in a low-dimensional subspace of R^d. 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. So, I think we can think a complete and strict proof in the future, to improve UCB-GLM and achieve both efficiency and optimal regret bound.

5.2 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.
I think it is still a theoretically interesting question.

Appendix

Proof of Lemma 1 *Let $\{X_t\}_{t=1}^{\infty}$ be the sequence in R^d satisfying $||X_t|| \leq 1$. Define $X_0 = 0$ 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 $n > 0$,*

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

$Proof\ of\ Lemma\ 1$:
By Abbasi-Yadkori et al. (2011, Lemma 11), we have

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

Note that $tr(V_{m+1}) = \sum_{t=1}^{m} ||X_t^2|| \leq m$ and that $detV_{m+1} = \Pi_{i=1}^{d}\lambda_i \geq \lambda_{min}^d(V_{m+1}) \geq 1$, where$\{\lambda_i\}$ are the eigenvalues of V_{m+1}. Applying Cauchy-Schwartz inequality yields

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

Proof of Lemma 2 *Suppose $\lambda_{min}(V_{m+1}) \geq 1$. For $\delta \in [\frac{1}{T}, 1)$, define event*

$$\varepsilon_{\triangle} := \left\{ || \triangle_t ||_{V_t} \leq \frac{\sigma}{\kappa}\sqrt{\frac{d}{2}log(1+2t/d) + log(1/\delta)} \right\}$$

Then, event $\varepsilon_{\triangle}$ holds for all $t \geq \tau$ with probability at least $1-\delta$.

$Proof\ of\ Lemma\ 2$:
Define $G_t(\theta) = \sum_{i=1}^{t-1}(\mu(X_i'\theta) - \mu(X_i'\theta^*))$, and $Z_t = \sum_{i=1}^{t-1}\epsilon_i X_i$. Following the same argument as in the proof of Theorem 1, we have $G_t(\hat{\theta}_t) = Z_t$ and

$$||G_t(\theta)||_{V_t^{-1}}^2 \geq \kappa^2 ||\theta - \theta^*||_{V_t}^2$$

for any $\theta \in \{\theta : ||\theta - \theta^*|| \leq 1\}$. Combining the following two formulas and completes the proof.

Proof of Lemma 3 *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.*

$Proof\ of\ Lemma\ 3$:
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.

Proof of Lemma 4 *Fix $\delta > 0$. Choose in SupCB-GLM $\tau = \sqrt{dT}$ and $\alpha = \frac{3\sigma}{\kappa}\sqrt{2log(TK/\delta)}$. Suppose T satisfies condition (6). Define the following event:*

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

Then, event ε_X holds with probability as least $1-\delta$

$Proof\ of\ Lemma\ 4$:

By Lemma 3, we have independent samples now. Then to apply Theorem 1, the key is to lower bound the minimum eigenvalue of Vt. 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, . . . , X_τ 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 $T_0 = \Omega\left(\frac{\sigma^2}{\kappa^4} max\{d^3, \frac{log(TK/\delta)}{d}\}\right)$.

References