Reward-Biased Maximum Likelihood Estimation for Linear Stochastic Bandits

[1]Yu-Heng Hung, [1]Ping-Chun Hsieh, [2]Xi Liu, and [2]P. R. Kumar
[1]National Chiao Tung University
[2]Texas A&M University

Abstract

Modifying the reward-biased maximum likelihood method originally proposed in the adaptive control literature, we propose novel learning algorithms to handle the explore-exploit trade-off in linear bandit problems as well as generalized linear bandit problems. We develop novel index policies that we prove achieve order-optimality, and show that they achieve empirical performance competitive with state-of-the-art baselines in extensive experiments, while entailing low computational complexity per pull for linear bandits.

1 Introduction

The problem of decision making for an unknown dynamic system, called stochastic adaptive control [1, 2], was examined in the control theory community beginning in the 1950s. It was recognized early on by Feldbaum [3, 4] that control played a dual role, that of exciting a system to learn its dynamics, as well as satisfactorily regulating its behavior, therefore dubbed as the problem of "dual control." This leads to a central problem of identifiability: As the controller begins to converge, it ceases to learn about the behavior of the system to other control actions. This issue was quantified by Borkar and Varaiya [5] within the setting of adaptive control of Markov chains. Consider a stochastic system with a state-space X, control or action set U, modelled as a controlled Markov chain with transition probabilities Prob$(x(t+1) = j|x(t) = i, u(t) = u) = p(i, j; u, \theta_*)$ dependent on an unknown parameter θ_* lying in a known set Θ, where $x(t)$ is the state of the system at time step t, and $u(t)$ is the action taken at that time. Given a one-step reward function $r(i, u)$, let $\phi : X \times \Theta \to U$ denote the optimal stationary control law as a function of $\theta \in \Theta$ for the long-term average reward problem: $\max \frac{1}{T}\sum_{t=0}^{T-1} r(x(t), u(t))$, i.e., $u(t) = \phi(x(t), \theta)$ is the optimal action to take if the true parameter is θ. Since θ_* is unknown, consider a "certainty-equivalent" approach: At each time step t, let $\widehat{\theta}_{\mathrm{ML}}(t) \in \mathrm{argmax}_{\theta\in\Theta} \sum_{s=0}^{t-1} \log p(x(s), x(s+1), u(s), \theta)$ denote the Maximum Likelihood (ML) estimate of θ_*, with ties broken according to any fixed priority order. Then apply the action $u(t) = \phi(x(t), \widehat{\theta}_{\mathrm{ML}}(t))$ to the system. It was shown in [6] that under an irreducibility assumption, the parameter estimates $\widehat{\theta}_{\mathrm{ML}}(t)$ converge to a random limit $\check{\theta}$ satisfying

$$p(i, j, \phi(i, \check{\theta}), \check{\theta}) = p(i, j, \phi(i, \check{\theta}), \theta_*) \qquad \forall i, j \in X. \tag{1}$$

That is, the closed-loop transition probabilities under the control law $\phi(\cdot, \check{\theta})$ are correctly determined. However, the resulting feedback control law $\phi(\cdot, \check{\theta})$ need not be optimal for the true parameter θ_*.

A key observation that permitted a breakthrough on this problem was made by Kumar and Becker [6]. Denote by $J(\phi, \theta)$ the long-term average reward incurred when the stationary control law ϕ is used if the true parameter is θ, and by $J(\theta) := \mathrm{Max}_\phi J(\phi, \theta)$ the optimal long-term average reward attainable when the parameter is θ. Then,

$$J(\check{\theta}) \overset{(a)}{=} J(\phi(\cdot, \check{\theta}), \check{\theta}) \overset{(b)}{=} J(\phi(\cdot, \check{\theta}), \theta_*) \overset{(c)}{\leq} J(\theta_*). \tag{2}$$

Preprint. Under review.

where the key equality (b) that the long-term reward under $\phi(\cdot, \check{\theta})$ is the same under the parameters $\check{\theta}$ and θ_* follows from the equivalence of the closed-loop transition probabilities (1), while (a) and (c) hold trivially since $\phi(\cdot, \check{\theta})$ is optimal for $\check{\theta}$, but is not necessarily optimal for θ_*. Therefore the maximum likelihood estimator is biased in favor of parameters with *smaller* reward. To counteract this bias, [6] proposed delicately biasing the ML parameter estimation criterion in the reverse way in favor of parameters with *larger* reward by adding a term $\alpha(t)J(\theta)$ to the log-likelihood, with $\alpha(t) > 0$, $\alpha(t) \to +\infty$, and $\frac{\alpha(t)}{t} \to 0$. This results in the *Reward-Biased ML Estimate* (RBMLE):

$$\widehat{\theta}_{\text{RBMLE}}(t) \in \underset{\theta \in \Theta}{\operatorname{argmax}} \Big\{ \alpha(t)J(\theta) + \sum_{s=0}^{t-1} \log p(x(s), x(s+1), u(s), \theta) \Big\}. \tag{3}$$

This modification is delicate since $\alpha(t) = o(t)$, and therefore retains the ability of the ML estimate to estimate the closed-loop transition probabilities, i.e., (1) continues to hold, for any "frequent" limit point $\check{\theta}$ (i.e., that which occurs as a limit along a sequence with sufficient density in the integers). Hence the bias $J(\check{\theta}) \leq J(\theta)$ of (2) continues to hold. However, since $\alpha(t) \to +\infty$ the bias in favor of parameters with larger rewards ensures that

$$J(\check{\theta}) \leq J(\theta_*). \tag{4}$$

From (2,4) it follows that $J(\phi(, \cdot, \check{\theta}), \theta_*) = J(\theta_*)$, whence $\phi(\cdot, \check{\theta})$ is optimal for the unknown θ_*.

The RBMLE method holds potential as a general-purpose method for learning of dynamic systems. However, its analysis was confined to long-term average optimality, which only assures that the regret is $o(t)$. Pre-dating the UCB method of Lai and Robbins [7], it has largely remained unexplored vis-a-vis its finite-time performance as well as empirical performance on contemporary problems. Motivated by this, there has been recent interest in revisiting the RBMLE. Recently, its regret performance has been established for classical multi-armed bandits for the exponential family of measures [8]. However, classical bandits do not allow the incorporation of "context," which is important in several applications [9–13]. In this paper, we examine the RBMLE method both for linear contextual bandits as well as a more general class generalized linear bandits. Linear bandits and its variants have been popular models for abstracting the sequential decision making in various applications, such as recommender systems [9] and medical treatment [13].

This paper employs the RBMLE principle and obtains simple index policies for linear contextual bandits as well as its generalizations that have order-optimal theoretical finite-time regret performance as well as empirical performance competitive with the best currently available. (i) We adapt the RBMLE principle to linear contextual bandits by proposing a specific type of reward-bias term. An important modification is to introduce a regularization term into the RBMLE criterion. (ii) Both for usage in situations where distribution of the noise is unknown, as well as to derive simple index policies, we introduce into RBMLE the modification of using a Gaussian pseudo-likelihood function. (iii) We show that the so modified RBMLE index attains an order-optimal regret bound for general, possibly non-parametric, sub-Gaussian rewards. (iv) Based on the results for the linear bandits, we extend the techniques to the generalized linear models and show that the same regret bound can still be attained in the general case. (v) We show that the resulting RBMLE-based generalized linear bandits have a cumulative regret of the order of $\sqrt{T}\log T$. This shaves a factor of $\sqrt{\log T}$ from [14], and achieves the same regret bound as that of UCB-GLM in [15]. (vi) We also conduct extensive experiments to demonstrate the effectiveness, efficiency, and scalability of the proposed algorithms.

2 Problem Setup

We consider the stochastic contextual bandit problem with $K < +\infty$ arms, possibly large. At the beginning of each decision time $t \in \mathbb{N}$, a d-dimensional context vector $x_{t,a} \in \mathbb{R}^d$, with $\|x_{t,a}\| \leq 1$, is revealed to the learner, for each arm $a \in [K]$. The contexts $\{x_{t,a}\}$ are generated by an adaptive adversary, which determines them in an arbitrary way based on the history of all the contexts and rewards. Given the contexts, the learner selects an arm $a_t \in [K]$ and obtains the corresponding reward r_t, which is conditionally independent of all the other rewards in the past given the context $\{x_{t,a_t}\}$. We define (i) $x_t := x_{t,a_t}$, (ii) X_t as the $(t-1) \times d$ matrix in which the s-th row is $x_s^\intercal$, for all $s \in [t-1]$, (iii) $R_t := (r_1, \cdots, r_{t-1})^\intercal$ row vector of the observed rewards up to time $t-1$, and (iv) $\mathcal{F}_t = (x_1, a_1, r_1, \cdots, x_t)$ denotes the σ-algebra of all the causal information available right before r_t is observed. We assume that the rewards are linearly realizable, i.e. there exists an unknown parameter $\theta_* \in \mathbb{R}^d$ with $\|\theta_*\|_2 \le 1$, and a known, strictly increasing *link function* $\mu : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}[r_t|\mathcal{F}_t] = \mu(\theta_*^\intercal x_t)$. We assume that μ is continuously differentiable, with its derivative μ' having a supremum L_μ, and an infimum $\kappa_\mu > 0$. We call this the *generalized linear bandit* problem.

Let $a_t^* := \arg\max_{1\le i\le K} \theta_*^\intercal x_{t,i}$ be an arm that yields the largest conditional expected reward $\mathbb{E}[r_t|\mathcal{F}_t]$ at time t (with ties broken arbitrarily), and $x_t^* := x_{t,a_t^*}$. The objective of the learner is to maximize its total over a finite time horizon T, i.e., the learner aims to minimize the *total conditional expected pseudo-regret*, which we shall refer to simply as the “cumulative regret," defined as

$$\mathcal{R}(T) := \sum_{t=1}^{T} \mu(\theta_*^\intercal x_t^*) - \mu(\theta_*^\intercal x_t). \tag{5}$$

We call the problem a *standard* linear bandits problem if (i) the reward is $r_t = \theta_*^\intercal x_t + \varepsilon_t$, (ii) ε_t is a noise with $\mathbb{E}[\varepsilon_t|x_t] = 0$, and (iii) the rewards are conditionally σ-sub-Gaussian, i.e.

$$\mathbb{E}[\exp(\rho\varepsilon_t)|\mathcal{F}_t] \le \exp\big(\frac{\rho^2\sigma^2}{2}\big). \tag{6}$$

Wlog, we assume $\sigma = 1$. For standard linear bandits the link function μ is an identity and $\kappa_\mu = 1$.

3 RBMLE for Standard Linear Bandits

We begin with the derivation of the RBMLE index and its regret analysis for linear contextual bandits.

3.1 Index Derivation for Standard Linear Bandits

Let $\ell(\mathcal{F}_t;\theta)$ denote the log-likelihood of the historical observations when the true parameter is θ. Let λ be a positive constant. At each time t, the learner takes the following two steps.

1. Let $\bar\theta_t$ be any maximizer of $\Big\{\ell(\mathcal{F}_t;\theta) + \alpha(t)\cdot\max_{1\le a\le K}\theta^\intercal x_{t,a} - \frac{\lambda}{2}\|\theta\|_2^2\Big\}$.
2. Choose any arm a_t that maximizes $\bar\theta_t^\intercal x_{t,a}$.

The term $\alpha(t)\max_{1\le a\le K}\theta^\intercal x_{t,a}$ is the reward-bias. A modification to the RBMLE is the additional quadratic regularization term $\frac{\lambda}{2}\|\theta\|_2^2$, à la ridge regression. Wlog, we assume that $\lambda \ge 1$.

The above strategy can be simplified to an *index strategy*. Define the index of an arm a at time t by

$$\mathcal{I}_{t,a} := \max_{\theta}\Big\{\ell(\mathcal{F}_t;\theta) + \alpha(t)\cdot\theta^\intercal x_{t,a} - \frac{\lambda}{2}\|\theta\|_2^2,\Big\}, \tag{7}$$

and simply choose an arm a_t that has maximum index. The indexability proof is in Appendix A.

To derive indices, it is necessary to know what the log-likelihood $\ell(\mathcal{F}_t;\theta)$ is. However, in practice, the true distribution of the noise ε_t is unknown to the learner or it may not even follow any parametric distribution. We employ the Gaussian density function as a surrogate:

$$\ell(\mathcal{F}_t;\theta) = -\frac{1}{2}\sum_{s=1}^{t-1}(\theta^\intercal x_s - r_s)^2 - \frac{t-1}{2}\log(2\pi). \tag{8}$$

Hence $\bar\theta_t$ is any maximizer of $\Big\{-\sum_{s=1}^{t-1}(\theta^\intercal x_s - r_s)^2 + 2\alpha(t)\cdot\max_{1\le a\le K}\theta^\intercal x_{t,a} - \lambda\|\theta\|_2^2\Big\}$.

It is shown in Section 3.2 that despite the likelihood misspecification, the index derived from the Gaussian density achieves the same regret bound for general non-parametric sub-Gaussian rewards.

The RBMLE index has the following explicit form, as proved in Appendix B:

Corollary 1 For the Gaussian likelihood (8), there is a unique maximizer of (7) for every arm a,

$$\bar\theta_{t,a} = V_t^{-1}(X_t^\intercal R_t + \alpha(t)x_{t,a}), \tag{9}$$

where $V_t := X_t^\intercal X_t + \lambda I$. The arm a_t chosen by the RBMLE algorithm is

$$a_t = \operatorname*{argmax}_{1 \le i \le K} \Big\{ \widehat{\theta}_t^\intercal x_{t,i} + \frac{1}{2}\alpha(t)\|x_{t,i}\|^2_{V_t^{-1}} \Big\}, \tag{10}$$

where $\widehat{\theta}_t := {V_t}^{-1} X_t^\intercal R_t$ is the least squares estimate of θ_*.

We summarize the resulting LinRBMLE algorithm in Algorithm 1:

Algorithm 1 LinRBMLE Algorithm

```
1: Input: $\alpha(t)$, $\lambda$
2: Initialization: $V_1 \leftarrow \lambda I$
3: for $t = 1, 2, \cdots$ do
4:     Observe the contexts $\{x_{t,a}\}$ for all the arms
5:     Select the action $a_t = \operatorname{argmax}_a \{\widehat{\theta}_t^\intercal x_{t,a} + \frac{1}{2}\alpha(t)\|x_{t,a}\|^2_{V_t^{-1}}\}$ and obtain $r_t$
6:     Update $V_{t+1} \leftarrow V_t + x_{t,a_t} x_{t,a_t}^\intercal$
7: end for
```

Remark 1 Similar to the well-known LinUCB index $\widehat{\theta}_t^\intercal x_{t,i} + \gamma\|x_{t,i}\|_{{V_t}^{-1}}$ [9, 16], the LinBMLE index is also defined as the sum of the least square estimate and an additional exploration term. Despite this high-level resemblance, LinRBMLE has two salient features: (i) Instead of being constructed from the confidence interval, the LinRBMLE index is derived from the machinery of biased maximum likelihood; (ii) Compared to LinUCB, the exploration term of LinRBMLE has an additional factor of $\alpha(t)\|x_{t,i}\|_{V_t^{-1}}$, which encourages more exploration for those arms with $\alpha(t)\|x_{t,i}\|_{V_t^{-1}} > 1$. This additional exploration helps the learner identify the optimal arm faster, especially for challenging problems. As will be seen in Section 3.2, with a proper bias term (e.g., $\alpha(t) = \sqrt{t}$), the additional exploration does not sacrifice the regret bound. Moreover, the regret statistics in Section 5 suggest that this design makes LinRBMLE empirically more robust across different sample paths.

3.2 Regret Bound for the LinRBMLE index

We begin regret analysis with a bound on "immediate" regret $R_t := \theta_*^\intercal(x_t^* - x_t)$.

Lemma 1 Under the standard linear bandit model,

$$R_t \le \|\theta_* - \widehat{\theta}_t\|_{V_t} \cdot \|x_t^*\|_{V_t^{-1}} - \frac{1}{2}\alpha(t)\|x_t^*\|^2_{V_t^{-1}} + \|\widehat{\theta}_t - \theta_*\|_{V_t} \cdot \|x_t\|_{V_t^{-1}} + \frac{1}{2}\alpha(t)\|x_t\|^2_{V_t^{-1}}. \tag{11}$$

The proof of Lemma 1 is in Appendix C.

Remark 2 Note that Lemma 1 highlights the main difference between the analysis of the Upper Confidence Bound (UCB)-type algorithms (e.g., [17, 16]) and that of the LinRBMLE algorithm. To arrive at a regret upper bound for LinRBMLE, it is required to handle both $\|\theta_*^\intercal - \widehat{\theta}_t\|_{V_t} \cdot \|x_t^*\|_{V_t^{-1}}$ and $\frac{1}{2}\alpha(t)\|x_t^*\|^2_{V_t^{-1}}$. While it could be challenging to quantify each individual term, we show in Theorem 1 that a tight regret upper bound can be obtained by jointly analyzing these two terms.

Theorem 1 below presents the regret bound for the BMLE algorithm; it is proved in Appendix D. Let

$$G_0(t,\delta) := \sigma\sqrt{d\log\left(\frac{\lambda+t}{\lambda\delta}\right)} + \lambda^{\frac{1}{2}}, \text{ and } G_1(t) := \sqrt{2d\log\left(\frac{\lambda+t}{d}\right)}. \tag{12}$$

Theorem 1 For the RBMLE index (10), with probability at least $1-\delta$, the cumulative regret satisfies

$$\mathcal{R}(T) = \sum_{t=1}^{T} R_t \le \big(G_0(T,\delta)\big)^2 \cdot \Big(\sum_{t=1}^{T}\frac{1}{2\alpha(t)}\Big) + \sqrt{T}G_0(T,\delta)G_1(T) + \frac{1}{2}\alpha(T)\big(G_1(T)\big)^2. \tag{13}$$

Consequently, by choosing the bias term $\alpha(t) = \sqrt{t}$, the regret bound is $\mathcal{R}(T) = \mathcal{O}(\sqrt{T}\log T)$.

4 BMLE for Linear Bandits With General Link Functions

4.1 Index Derivation for Generalized Linear Bandits

For the generalized linear case, as before, let $\bar{\theta}_t$ be any maximizer of $\big\{\ell(\mathcal{F}_t;\theta) + \alpha(t) \cdot \max_{1\leq a\leq K} \theta^\intercal x_{t,a} - \frac{\lambda}{2}\|\theta\|_2^2\big\}$. However, a major difference vis-à-vis the standard linear case is that $L_\mu > \kappa_\mu$. To handle this, we incorporate an additional factor $\eta(t)$ that is a positive-valued, strictly increasing function that satisfies $\lim_{t\to\infty}\eta(t) = \infty$, and choose any arm a_t that maximizes $\big\{\ell(\mathcal{F}_t;\bar{\theta}_{t,a}) + \eta(t)\alpha(t)\cdot\bar{\theta}_{t,a}^\intercal x_{t,a} - \frac{\lambda}{2}\|\bar{\theta}_{t,a}\|_2^2\big\}$. The regret analysis below suggests that it is sufficient to choose $\eta(t)$ to be slowly increasing, e.g., $\eta(t) = 1 + \log t$.

Next, we generalize the notion of a surrogate Gaussian likelihood discussed in Section 3.1 by considering the density functions of the canonical exponential families:

$$p(r_t|x_t) = \exp(r_t x_t^\intercal \theta_* - b(x_t^\intercal\theta_*) + c(r_t)), \tag{14}$$

where $b(\cdot):\mathbb{R}\to\mathbb{R}$ is a strictly convex function that satisfies $b'(z) = \mu(z)$, for all $z\in\mathbb{R}$, and $c(\cdot):\mathbb{R}\to\mathbb{R}$ is the normalization function. The exponential family consists of a variety of widely-used distributions, including binomial, Gaussian, and Poisson distributions. By the properties of the exponential family, $b'(x_t^\intercal\theta_*) = \mathbb{E}[r_t|x_t]$ and $b''(x_t^\intercal\theta_*) = \mathbb{V}[r_t|x_t] > 0$. By (21) and the strict convexity of $b(\cdot)$, $\ell(\mathcal{F}_t;\theta) + \alpha(t)\cdot\theta^\intercal x_{t,a}$ is strictly concave in θ and therefore has a unique maximizer. By the first-order sufficient condition,

$$\bar{\theta}_{t,a} \text{ is the unique solution to } \sum_{s=1}^{t-1}\big(r_s x_s - \mu(x_s^\intercal\bar{\theta}_{t,a})x_s\big) - \lambda\bar{\theta}_{t,a} + \alpha(t)x_{t,a} = 0. \tag{15}$$

Note that (14) is used only for index derivation and is not required in the regret analysis in Section 4.2. We summarize the resulting GLM-RBMLE algorithm for the generalized linear case in Algorithm 2.

Algorithm 2 GLM-RBMLE Algorithm

```
Input: $\alpha(t)$, $\lambda$, $\eta(t)$
for $t = 1, 2, \cdots$ do
    Observe the contexts $\{x_{t,a}\}$ for all the arms
    Calculate $\bar{\theta}_{t,a}$ by solving $\sum_{s=1}^{t-1}\big(r_s x_s - \mu(x_s^\intercal\bar{\theta}_{t,a})x_s\big) - \lambda\bar{\theta}_{t,a} + \alpha(t)x_{t,a} = 0$, for each $i$
    Select $a_t = \operatorname{argmax}_a \big\{\ell(\mathcal{F}_t;\bar{\theta}_{t,a}) + \eta(t)\alpha(t)\bar{\theta}_{t,a}^\intercal x_{t,a} - \frac{\lambda}{2}\|\bar{\theta}_{t,a}\|_2^2\big\}$ and obtain $r_t$
end for
```

4.2 Regret Bound for GLM-RBMLE for Generalized Linear Bandits

We begin the regret analysis of GLM-RBMLE with a preliminary result proved in Appendix E.

Lemma 2 For any arms i and j, there exists $\bar{\theta}_0 = \beta_0\bar{\theta}_{t,i} + (1-\beta_0)\bar{\theta}_{t,j}$ with $\beta_0\in(0,1)$ such that

$$(x_{t,i} + x_{t,j})^\intercal(\bar{\theta}_{t,j} - \bar{\theta}_{t,i}) + \alpha(t)\|x_{t,i}\|_{U_0{}^{-1}} - \alpha(t)\|x_{t,j}\|_{U_0{}^{-1}} = 0, \tag{16}$$

where $U_0 := \sum_{s=1}^{t-1}\mu'(x_s^\intercal\bar{\theta}_0)x_s x_s^\intercal + \lambda I$ is a $d\times d$ positive definite matrix.

Define $T_0 := \min\{t\in\mathbb{N} : \frac{L_\mu^3}{2\kappa_\mu^2\eta(t)} < \frac{1}{2}\}$. For ease of exposition, define the function

$$G_2(t,\delta) := \frac{\sigma}{\kappa_\mu}\sqrt{\frac{d}{2}\log(1+\frac{2t}{d}) + \log\frac{1}{\delta}}. \tag{17}$$

We also define $C_1 := 2L_\mu^4/k_\mu^4 + 1/k_\mu^2$, $C_2 := 2L_\mu^3/\kappa_\mu^2 + L_\mu/\kappa_\mu$, and $C_3 := L_\mu^2/2$.

Theorem 2 For the GLM-RBMLE index, with probability at least $1-\delta$, the cumulative regret satisfies

$$\mathcal{R}(T) \leq T_0 + C_1\alpha(T)\big(G_1(T)\big)^2 + C_2\sqrt{T}G_1(T)G_2(T,\delta) + C_3\big(G_2(T,\delta)\big)^2\sum_{t=1}^{T}\frac{1}{\alpha(t)}. \tag{18}$$

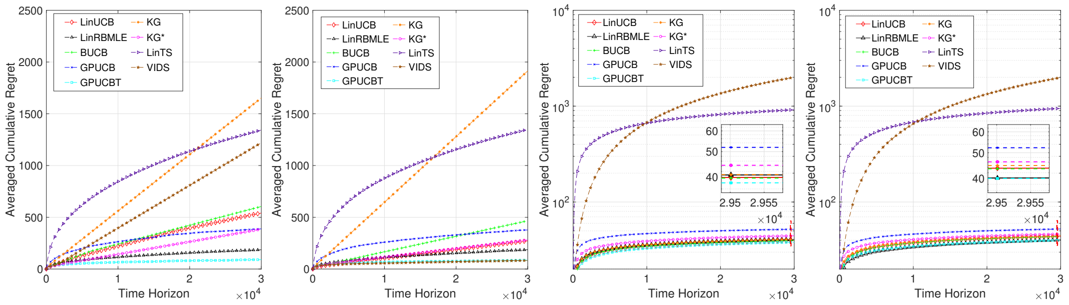
Consequently, if $\alpha(t) = \Omega(\sqrt{t})$, then $\mathcal{R}(T) = \mathcal{O}(\alpha(T)\log T)$. Otherwise, if $\alpha(t) = \mathcal{O}(\sqrt{t})$, then $\mathcal{R}(T) = \mathcal{O}\big((\sum_{t=1}^{T}\frac{1}{\alpha(t)})\log T\big)$. Hence, by choosing $\alpha(t) = \sqrt{t}$, $\mathcal{R}(T) = \mathcal{O}(\sqrt{T}\log T)$.

Remark 3 This bound improves that in [14] by a $\sqrt{\log T}$ factor and is the same as UCB-GLM [15].

5 Numerical Experiments

To evaluate performance of the proposed RBMLE methods, we conduct a comprehensive empirical comparison with other state-of-the-art methods vis-a-vis three aspects: effectiveness (cumulative regret), efficiency (computation time per decision vs. cumulative regret), and scalability (in number of arms and dimension of contexts). We paid particular attention to fairness of comparison and reproducibility of results. To ensure sample-path sameness for all methods, we compared each method over a pre-prepared dataset containing the context of each arm and the outcomes of pulling each arm over all rounds. Hence, the outcome of pulling an arm is obtained by querying the pre-prepared data instead of calling the random generator and changing its state. A few benchmarks such as Thompson Sampling (LinTS) and Variance-based Information Directed Sampling (VIDS) that rely on outcomes of random sampling in each round of decision-making are separately evaluated with the same prepared data and with the same seed. To ensure reproducibility of experimental results, we set up the seeds for the random number generators at the beginning of each experiment and provide all the codes, including seed setup in the supplementary material.

The benchmark methods compared include LinUCB [16], LinTS [18], Bayes-UCB (BUCB) [19], Gaussian Process Upper Confidence Bound (GPUCB) [20] and its variant GPUCB-Tuned (GPUCBT) [21], Knowledge Gradient (KG) and its variant KG* [22–24], and VIDS [21]. A detailed review of these methods is presented in Section 6. The values of their hyper-parameters are as follows. As suggested by Theorem 1, $\alpha(t) = \sqrt{t}$ and $\lambda = 1$ in LinRBMLE. We take $\alpha = 1$ in LinUCB and $\delta = 10^{-5}$ in GPUCB. We tune the parameter c in GPUCBT for each experiment and choose $c = 0.9$ that achieves the best performance. We follow the suggestion of [19] to choose $c = 0$ for BUCB. Respecting the restrictions in [18], we take $\delta = 0.5$ and $\epsilon = 0.9$ in LinTS. In the comparison with IDS and VIDS, we sampled 10^3 points over the interval $[0, 1]$ for q (Algorithm 4 in [21]) and take $M = 10^4$ in sampling (Algorithm 6 in [21]). In the Bayesian family of benchmark methods (LinTS, BUCB, KG, KG*, GPUCB, GPUCBT, and VIDS), the prior distribution over the unknown parameters θ_* is $\mathcal{N}(0_d, I_d)$. The comparison contains 50 trials of experiments and T rounds in each trial. We consider both contexts, "static," where the context for each arm is fixed in each experiment trial, and "time-varying," where the context for each arm changes from round to round. All contexts are drawn randomly from $\mathcal{N}(0_d, 10I_d)$ and normalized by their ℓ_2 norm.



(a) (b) (c) (d)

Figure 1: Cumulative regret averaged over 50 trials with $T = 3 \times 10^4$ and $K = 10$: (a) and (b) are under static contexts; (c) and (d) are under time-varying contexts; (a) and (c) are with $\theta_* = (-0.3, 0.5, 0.8)$; (b) and (d) are with with $\theta_* = (-0.7, -0.6, 0.1)$.

Effectiveness. Figure 1 and Table 1 illustrate the effectiveness of LinRBMLE in terms of cumulative regret. We observe that for both static and time-varying contexts, LinRBMLE achieves performance only slightly worse than the best performing algorithm, which is often GPUCBT or VIDS. However, compared to these two, LinRBMLE has some salient advantages. In contrast to LinRBMLE, GPUCBT has no guaranteed regret bound and requires tuning the hyper-parameter c to establish its outstanding performance. This restricts its applicability if pre-tuning is not possible. Compared to VIDS, the computational complexity of LinRBMLE is much lower. As shown in Table 1, LinRBMLE also exhibits better robustness with an order of magnitude or two smaller std. dev. compared to VIDS and many other benchmark methods. In Figure 1(a), VIDS appears to have not converged, but a detailed check reveals that this is only because its performance in some trials is much worse than in other trials.

Table 1: Statistics of the final cumulative regret in Figure 1(a). The best and the second-best are highlighted. 'Q' and "Std.Dev" stand for quantile and standard deviation of the total cumulative regret over 50 trails, respectively. All the values displayed here are scaled by 0.01 for more compact notations.

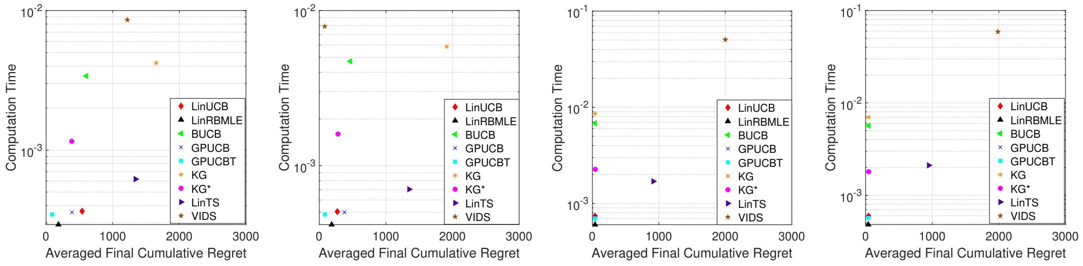
Alg.	RBMLE	LinUCB	BUCB	GPUCB	GPUCBT	KG	KG*	LinTS	VIDS
Mean	**1.86**	5.41	6.04	3.88	**0.90**	16.52	3.86	13.43	12.20
Std.Dev	**0.42**	14.87	11.78	1.19	**0.53**	26.68	10.46	2.20	74.66
Q.10	1.45	**0.04**	0.07	2.30	0.32	**0.03**	0.07	10.83	0.15
Q.25	1.62	**0.07**	0.10	3.01	0.59	**0.05**	0.10	12.44	0.29
Q.50	1.79	**0.15**	**0.14**	3.78	0.79	0.18	0.18	13.58	0.45
Q.75	1.96	1.00	1.30	4.56	1.09	23.83	**0.34**	14.25	**0.79**
Q.90	**2.31**	19.34	23.00	5.74	**1.66**	64.89	18.94	15.73	2.38
Q.95	**2.75**	30.47	36.31	5.91	**1.98**	75.96	27.18	16.78	9.40

The robustness is also reflected in variation across problem instances, e.g., the performance of VIDS is worse in the problem of Figure 1(b) than in the problem of Figure 1(a), while the performance of LinRBMLE is consistent in these two examples. The robustness of LinRBMLE across different sample paths can be largely attributed to the inclusion of the Reward Bias term $\alpha(t)$ in the index (10), which encourages more exploration even for those sample paths with small $\|x_{t,i}\|_{V_t^{-1}}$. It is worth mentioning that the advantage of VIDS compared to other methods is less obvious for time-varying contexts. Experimental results reported in [21] are restricted to the static contexts. More tables on the statistics of the the final cumulative regret in Figure 1 can be found in appendix.

Efficiency. Figure 2 presents the the averaged cumulative regret versus average computation time per decision. We observe that LinRBMLE and GPUCBT have points closest to the origin, signifying small regret along with small computation time, and outperform the other benchmark methods.

Scalability. Table 2 presents scalability of computation time per decision as K and d are varied. We observe that both LinRBMLE and GPUCBT, which are often the best among the benchmark methods have low computation time as well as better scaling when d or K are increased. Such scalability is important for big data applications such as recommender and advertising systems.

For generalized linear bandits, a similar study on effectiveness, efficiency, and scalability for GLM-RBMLE and popular benchmark methods is detailed in the supplementary material.



(a) (b) (c) (d)

Figure 2: Average computation time per decision vs. averaged cumulative regret for (a) Figure 1(a); (b) Figure 1(b); (c) Figure 1(c); (d) Figure 1(d).

6 Related Work

The RBMLE method was originally proposed in [6]. It was subsequently examined in the Markovian setting in [25–27], and in the linear quadratic Gaussian (LQG) system setting in [28–30]. A survey, circa 1985, of the broad field of stochastic adaptive control can be found in [1]. Recently it has been examined from the point of examining its regret performance in the case of non-contextual bandits with exponential family of distributions in [8]. Other than that, there appears to have been no work on examining its performance beyond long-term average optimality, which corresponds to regret of $o(t)$.

The linear stochastic bandits and their variants have been extensively studied from two main perspectives, namely the frequentist and the Bayesian approaches. From the frequentist viewpoint, one major

Table 2: Average computation time per decision for static contexts, under different values of K and d. All numbers are averaged over 50 trials with $T = 10^3$ and in 10^{-2} seconds. The best is highlighted. We set $d = 5$ and $K = 5$ as default.

Algorithm	RBMLE	LinUCB	BUCB	GPUCB	GPUCBT	KG	KG*	LinTS	VIDS
$K = 10$	**0.06**	0.07	0.66	0.07	0.07	0.82	0.23	0.34	16.02
$K = 20$	**0.38**	0.46	4.51	0.46	0.43	5.60	1.25	0.41	23.99
$K = 30$	0.64	0.77	7.86	0.77	0.74	9.85	2.09	**0.32**	26.45
$d = 10$	**0.02**	0.03	0.21	0.03	0.03	0.26	0.11	0.04	13.02
$d = 20$	**0.03**	0.04	0.30	0.04	0.04	0.37	0.15	0.24	22.29
$d = 30$	**0.04**	0.05	0.32	0.05	0.05	0.40	0.17	0.41	27.78

line of research is to leverage the least squares estimator and enforce exploration by constructing an upper confidence bound (UCB), introduced in the LinRel algorithm by Auer [31]. The idea of UCB was later extended to the LinUCB policy, which is simpler to implement and has been tested extensively via experiments [9]. While being simple and empirically appealing approaches, the primitive versions of the above two algorithms are rather difficult to analyze due to the statistical dependencies among the observed rewards. To obtain proper regret bounds, both policies were analyzed with the help of a more complicated master algorithm. To address this issue, Dani *et al.* [32] proposed to construct a confidence ellipsoid, which serves as an alternative characterization of UCB, and proved that the resulting algorithm achieved an order-optimal regret bound (up to a poly-logarithmic factor). Later, sharper characterizations of the confidence ellipsoid were presented by Rusmevichientong and Tsitsiklis [33] and Abbasi-Yadkori et al. [17] thereby improving the regret bound. Given the success of UCB-type algorithms for linear bandits, the idea of a confidence set was later extended to the generalized linear case [14, 15] to study a broader class of linear stochastic bandit models. Differing from the above UCB-type approaches, as a principled frequentist method, the proposed RBMLE algorithm guides the exploration toward potentially reward-maximizing model parameters by applying a bias to the log-likelihood. Most related is the work by Liu *et al.* [8], which adapted the RBMLE principle for stochastic multi-armed bandits and presented the regret analysis as well as extensive numerical experiments. However, [8] focused on the non-contextual bandit problems, and the presented results cannot directly apply to the more structured linear bandit model.

Instead of viewing model parameters as deterministic unknown variables, the Bayesian approaches assume a prior distribution to facilitate the estimation of model parameters. As one of the most popular Bayesian methods, Thompson sampling (TS) [34] approaches the exploration issue by sampling the posterior distribution. For linear bandit models, TS has been tested in large-scale experiments [11] and shown to enjoy order-optimal regret bounds in various bandit settings [18, 35–38]. On the other hand, Bayesian strategies can also be combined with the notion of UCB for exploration, as in the popular GP-UCB [20] and Bayes-UCB [19] algorithms. Alternative exploration strategies for linear bandits have also been considered from the perspective of explicit information-theoretic measures. In [21], Russo and Van Roy proposed a promising algorithm called information-directed sampling (IDS), which makes decisions based on the ratio between the square of expected regret and the information gain. As the evaluation of mutual information requires computing high-dimensional integrals, VIDS, a variant of IDS, was proposed to approximate the information ratio by sampling, while still achieving competitive empirical regret performance. Compared to IDS and its variants, the proposed RBMLE enjoys a closed-form index and is therefore computationally more efficient. Another promising solution is the Knowledge Gradient (KG) approach [23, 22], which enforces exploration by taking a one-step look-ahead measurement. While being empirically competitive, it remains unknown whether KG and its variants have a provable near-optimal regret bound. In contrast, the proposed RBMLE enjoys provable order-optimal regret for standard linear as well as generalized linear bandits.

7 Conclusion

In this paper, we employ the Reward Biased Maximum Likelihood principle originally proposed for adaptive control, to contextual bandits. RBMLE leads to a simple index policy for standard linear bandits. We show through regret analysis as well as empirical evaluation, that the RBMLE policy is competitive with the state-of-the art with respect to both performance as well as computation time per decision.

Broader Impact

Linear bandits as well as the generalized models serve as a powerful framework for sequential decision making in various critical applications, such as clinical trials [39], mobile health [13], personalized recommender [9] and online advertising systems [11], etc. The rising volume of datasets in these applications requires learning algorithms that are more effective, efficient and scalable. The study in this paper contributes a new family of frequentist approaches to this community. These approaches are proved to be order-optimal and demonstrate strong empirical performance with respect to measures of effectiveness, efficiency and scalability. As such, the proposed approaches are expected to further improve user experience in applications and benefit business stakeholders. The proposed approaches are inspired by an early adaptive control framework. This framework has been applied in many adaptive control applications [1, 25, 26, 28, 27, 29, 30]. However, analysis of its finite-time performance has been missing for decades. Our study takes a very first step towards understanding its finite-time performance in the contextual bandit setting.

Unfortunately, as in many other contextual bandit studies, our model does not take into account the fairness issue in learning the unknown parameters. For instance, it may happen that during the learning process, contextual bandit algorithms may consistently discriminate against some specific groups of users based on their social, economic, racial and sexual characteristics. Ensuring fairness may therefore require additional constraints on automated selection procedures. Such a study can contribute to general studies on the undesirable biases of machine learning algorithms [40].

References

[1] P. R. Kumar. A survey of some results in stochastic adaptive control. *SIAM Journal on Control and Optimization*, 23(3):329–380, 1985.

[2] P. R. Kumar and Pravin Varaiya. *Stochastic systems: Estimation, identification, and adaptive control*. 1986.

[3] A. A. Feldbaum. Dual control theory. i. *Avtomatika i Telemekhanika*, 21(9):1240–1249, 1960.

[4] A. A. Feldbaum. Dual control theory. ii. *Avtomatika i Telemekhanika*, 21(11):1453–1464, 1960.

[5] Vivek Borkar and Pravin Varaiya. Adaptive control of markov chains, i: Finite parameter set. *IEEE Transactions on Automatic Control*, 24(6):953–957, 1979.

[6] P. R. Kumar and Arthur Becker. A new family of optimal adaptive controllers for markov chains. *IEEE Transactions on Automatic Control*, 27(1):137–146, 1982.

[7] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[8] Xi Liu, Ping-Chun Hsieh, Yu-Heng Huang, Anirban Bhattacharya and P. R. Kumar. Exploration through bias: Revisiting biased maximum likelihood estimation in stochastic multi-armed bandits. In *International Conference on Machine Learning*, 2020.

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

[10] Tyler Lu, Dávid Pál, and Martin Pál. Contextual multi-armed bandits. In *Proceedings of the Thirteenth international conference on Artificial Intelligence and Statistics*, pages 485–492, 2010.

[11] Olivier Chapelle and Lihong Li. An empirical evaluation of Thompso sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.

[12] Shuai Li, Alexandros Karatzoglou, and Claudio Gentile. Collaborative filtering bandits. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 539–548, 2016.

[13] Ambuj Tewari and Susan A Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pages 495–517. Springer, 2017.

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

[15] 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-Volume 70*, pages 2071–2080. JMLR. org, 2017.

[16] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

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

[18] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135, 2013.

[19] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On bayesian upper confidence bounds for bandit problems. In *Artificial intelligence and statistics*, pages 592–600, 2012.

[20] Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 1015–1022, 2010.

[21] Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018.

[22] Ilya O Ryzhov, Peter I Frazier, and Warren B Powell. On the robustness of a one-period look-ahead policy in multi-armed bandit problems. *Procedia Computer Science*, 1(1):1635–1644, 2010.

[23] Ilya O Ryzhov, Warren B Powell, and Peter I Frazier. The knowledge gradient algorithm for a general class of online learning problems. *Operations Research*, 60(1):180–195, 2012.

[24] Bogumił Kamiński. Refined knowledge-gradient policy for learning probabilities. *Operations Research Letters*, 43(2):143–147, 2015.

[25] P. R. Kumar and Woei Lin. Optimal adaptive controllers for unknown markov chains. *IEEE Transactions on Automatic Control*, 27(4):765–774, 1982.

[26] P. R. Kumar. Simultaneous identification and adaptive control of unknown systems over finite parameter sets. *IEEE Transactions on Automatic Control*, 28(1):68–76, 1983.

[27] Vivek Borkar. The Kumar-Becker-Lin scheme revisited. *Journal of optimization theory and applications*, 66(2):289–309, 1990.

[28] P. R. Kumar. Optimal adaptive control of linear-quadratic-Gaussian systems. *SIAM Journal on Control and Optimization*, 21(2):163–178, 1983.

[29] Marco Campi and P. R. Kumar. Adaptive linear quadratic Gaussian control: the cost-biased approach revisited. *SIAM Journal on Control and Optimization*, 36(6):1890–1907, 1998.

[30] Maria Prandini and Marco Campi. Adaptive LQG control of input-output systems—a cost-biased approach. *SIAM Journal on Control and Optimization*, 39(5):1499–1519, 2000.

[31] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

[32] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*, 2008.

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

[34] 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.

[35] Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of Thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

[36] Marc Abeille, Alessandro Lazaric, et al. Linear Thompson sampling revisited. *Electronic Journal of Statistics*, 11(2):5165–5197, 2017.

[37] Shipra Agrawal and Navin Goyal. Near-optimal regret bounds for Thompson sampling. *Journal of the ACM (JACM)*, 64(5):1–24, 2017.

[38] Bianca Dumitrascu, Karen Feng, and Barbara Engelhardt. Pg-ts: Improved Thompson sampling for logistic contextual bandits. In *Advances in neural information processing systems*, pages 4624–4633, 2018.

[39] Yogatheesan Varatharajah, Brent Berry, Sanmi Koyejo, and Ravishankar Iyer. A contextual-bandit-based approach for informed decision-making in clinical trials. *arXiv preprint arXiv:1809.00258*, 2018.

[40] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pages 325–333, 2016.

[41] Dimitri P Bertsekas. *Nonlinear programming*. Athena Scientific, 1999.

Appendix

A Proof of Indexability of the Strategy (7)

Recall from Section 3.1 that $\bar{\theta}_t$ denotes a maximizer of the following problem:

$$\max_{\theta}\Big\{\ell(\mathcal{F}_t;\theta)+\alpha(t)\cdot\max_{1\leq a\leq K}\theta^{\intercal}x_{t,a}-\frac{\lambda}{2}\|\theta\|_2^2\Big\}. \tag{19}$$

Define

$$\bar{\mathcal{A}}_t := \operatorname*{argmax}_{a}\ \bar{\theta}_t^{\intercal}x_{t,a}, \tag{20}$$

$$\bar{\Theta}_{t,a} := \operatorname*{argmax}_{\theta}\Big\{\ell(\mathcal{F}_t;\theta)+\alpha(t)\cdot\theta^{\intercal}x_{t,a}-\frac{\lambda}{2}\|\theta\|_2^2\Big\}. \tag{21}$$

For each arm a, consider an estimator $\bar{\theta}_{t,a}\in\bar{\Theta}_{t,a}$. Subsequently, define an index set

$$\bar{\mathcal{A}}_t' := \operatorname*{argmax}_{1\leq a\leq K}\Big\{\ell(\mathcal{F}_t;\bar{\theta}_{t,a})+\alpha(t)\cdot\bar{\theta}_{t,a}^{\intercal}x_{t,a}-\frac{\lambda}{2}\|\bar{\theta}_{t,a}\|_2^2\Big\}. \tag{22}$$

Theorem 3 $\bar{\mathcal{A}}_t=\bar{\mathcal{A}}_t'$.

Proof The proof follows from the fact that any maximizer of the original double maximization problem in (19) remains a maximizer after interchanging the order of the max operators. By the definition of $\bar{\Theta}_t$ and $\bar{\mathcal{A}}_t$ in (19) and (20), given any $\bar{\theta}_t\in\bar{\Theta}_t$, any arm $a\in\bar{\mathcal{A}}_t$ is a maximizer of the optimization problem $\max_i\ \bar{\theta}_t^{\intercal}x_{t,i}$. We know

$$\operatorname*{argmax}_{1\leq i\leq K}\bar{\theta}_t^{\intercal}x_{t,i}=\operatorname*{argmax}_{1\leq i\leq K}\big\{\ell(\mathcal{F}_t;\bar{\theta}_t)+\alpha(t)\cdot\bar{\theta}_t^{\intercal}x_{t,i}\big\}. \tag{23}$$

Moreover,

$$\begin{aligned}
\max_{1\leq i\leq K}\big\{\ell(\mathcal{F}_t;\bar{\theta}_t)+\alpha(t)\cdot\bar{\theta}_t^{\intercal}x_{t,i}\big\} &= \ell(\mathcal{F}_t;\bar{\theta}_t)+\alpha(t)\cdot\max_{1\leq i\leq K}\bar{\theta}_t^{\intercal}x_{t,i} && (24)\\
&= \max_{\theta}\big\{\ell(\mathcal{F}_t;\theta)+\alpha(t)\cdot\max_{1\leq i\leq K}\theta^{\intercal}x_{t,i}\big\} && (25)\\
&= \max_{\theta}\max_{1\leq i\leq K}\big\{\ell(\mathcal{F}_t;\theta)+\theta^{\intercal}x_{t,i}\big\} && (26)\\
&= \max_{1\leq i\leq K}\max_{\theta}\big\{\ell(\mathcal{F}_t;\theta)+\theta^{\intercal}x_{t,i}\big\} && (27)\\
&= \max_{1\leq i\leq K}\big\{\ell(\mathcal{F}_t;\bar{\theta}_{t,i})+\bar{\theta}_{t,i}^{\intercal}x_{t,i}\big\}, && (28)
\end{aligned}$$

where (24) follows since $\ell(\mathcal{F}_t;\bar{\theta}_t)$ is independent of i, (25) holds by the definition of $\bar{\theta}_t$, (26)-(27) hold by the fact that the optimal value remains unchanged after interchanging the order of the two max operators, and (28) follows from the definition of $\bar{\theta}_{t,i}$. Therefore, by (23)-(28), $\bar{\mathcal{A}}_t=\bar{\mathcal{A}}_t'$. □

B Proof of Corollary 1

By substituting the Gaussian likelihood for $\ell(\mathcal{F}_t;\theta)$ in (21), the resulting objective in (21) becomes a strictly concave function and enjoys a unique maximizer. By the first-order necessary optimality condition [41], it is easy to verify that (9) is indeed the unique solution to (21). Subsequently, based on (9) and Theorem 3, we know the arm chosen by the RBMLE algorithm at each time t is

$$\begin{aligned}
a_t &= \operatorname*{argmax}_{1\leq i\leq K}\Big\{-(X_t\bar{\theta}_{t,i}-R_t)^{\intercal}(X_t\bar{\theta}_{t,i}-R_t)+\alpha(t)\bar{\theta}_{t,i}^{\intercal}x_{i,t}-\lambda\|\bar{\theta}_{t,i}\|_2^2\Big\} && (29)\\
&= \operatorname*{argmax}_{1\leq i\leq K}\Big\{-\bar{\theta}_{t,i}(X_t^{\intercal}X_t+\lambda I)\bar{\theta}_{t,i}+(2R_t^{\intercal}X_t+2\alpha(t)x_{t,i})\bar{\theta}_{t,i}-R_t^{\intercal}R_t\Big\} && (30)\\
&= \operatorname*{argmax}_{1\leq i\leq K}\Big\{(X_t^{\intercal}R_t+\alpha(t)x_{t,i})^{\intercal}(X_t^{\intercal}X_t+\lambda I)^{-1}(X_t^{\intercal}R_t+\alpha(t)x_{t,i})-R_t^{\intercal}R_t\Big\} && (31)\\
&= \arg\max_{1\leq i\leq K}\Big\{\widehat{\theta}_t^{\intercal}x_{t,i}+\frac{1}{2}\alpha(t)\|x_{t,i}\|_{V_t^{-1}}^2\Big\}. && (32)
\end{aligned}$$

where (29)-(31) hold by substituting (9) in (22), and (32) follows from the definition of $\widehat{\theta}_t$. □

C Proof of Lemma 1

Proof By the definition of regret for the linear bandit model,

$$
\begin{aligned}
R_t &= \theta_*^\intercal x_t^* - \theta_*^\intercal x_t \qquad &(33)\\
&= (\theta_* - \widehat{\theta}_t)x_t^* + \widehat{\theta}_t^\intercal x_t^* - \theta_*^\intercal x_t &(34)\\
&\leq (\theta_* - \widehat{\theta}_t)x_t^* + \widehat{\theta}_t x_t + \frac{1}{2}\alpha(t)\|x_t\|^2_{V_t^{-1}} - \frac{1}{2}\alpha(t)\|x_t^*\|^2_{V_t^{-1}} - \theta_*^\intercal x_t &(35)\\
&= (\theta_* - \widehat{\theta}_t)x_t^* + (\widehat{\theta}_t - \theta_*)x_t + \frac{1}{2}\alpha(t)\|x_t\|^2_{V_t^{-1}} - \frac{1}{2}\alpha(t)\|x_t^*\|^2_{V_t^{-1}}, &(36)
\end{aligned}
$$

where (35) follows from the RBMLE index (10). Let $V_t^{1/2}$ and $V_t^{-1/2}$ denote square-roots, satisfying $V_t = V_t^{1/2}V_t^{1/2}$ and $V_t^{-1} = V_t^{-1/2}V_t^{-1/2}$, unique since V_t is positive definite. The result (11) follows by replacing the vector multiplication of $(\theta_* - \widehat{\theta}_t)^\intercal x_t$ and $(\widehat{\theta}_t - \theta_*)^\intercal x_t$ in (36) by $(\theta_* - \widehat{\theta}_t)^\intercal V_t^{1/2}V_t^{-1/2}x_t$ and $(\widehat{\theta}_t - \theta_*)^\intercal V_t^{1/2}V_t^{-1/2}x_t$, and applying the Cauchy-Schwarz inequality. □

D Proof of Theorem 1

Before proving Theorem 1, we first introduce the following useful lemmas. Recall that $V_t = \sum_{s=1}^t x_s x_s^\intercal + \lambda I$. Moreover, recall that

$$G_0(t,\delta) := \sigma\sqrt{d\log\big(\frac{\lambda+t}{\lambda\delta}\big)} + \lambda^{\frac{1}{2}} \tag{37}$$

$$G_1(t) := \sqrt{2d\log\big(\frac{\lambda+t}{d}\big)} \tag{38}$$

Lemma 3 For any time $t \geq 1$, with probability at least $1-\delta$,

$$\|\theta_* - \widehat{\theta}_t\|_{V_t} \cdot \|x_t^*\|_{V_t^{-1}} - \frac{1}{2}\alpha(t)\|x_t^*\|^2_{V_t^{-1}} \leq \frac{1}{2\alpha(t)}\big(G_0(t,\delta)\big)^2. \tag{39}$$

Proof (Lemma 3) First, we obtain an upper bound by completing the square of the left-hand side of (39) as

$$
\begin{aligned}
&\|\theta_* - \widehat{\theta}_t\|_{V_t} \cdot \|x_t^*\|_{V_t^{-1}} - \frac{1}{2}\alpha(t)\|x_t^*\|^2_{V_t^{-1}} &(40)\\
&= -\frac{1}{2}\alpha(t)\Big(\|x_t^*\|_{V_t^{-1}} - \frac{\|\theta_* - \widehat{\theta}_t\|_{V_t}}{\alpha(t)}\Big)^2 + \frac{1}{2}\frac{\|\theta_* - \widehat{\theta}_t\|^2_{V_t}}{\alpha(t)} &(41)\\
&\leq \frac{1}{2}\frac{\|\theta_* - \widehat{\theta}_t\|^2_{V_t}}{\alpha(t)}. &(42)\\
& &(43)
\end{aligned}
$$

Moreover, by Theorem 2 in [17], we know that with probability at least $1-\delta$,

$$\|\theta_* - \widehat{\theta}_t\|_{V_t} \leq \sigma\sqrt{d\log\big(\frac{\lambda+t}{\lambda\delta}\big)} + \lambda^{\frac{1}{2}} = G_0(t,\delta). \tag{44}$$

Therefore, we can conclude that (39) indeed holds. □

Lemma 4 With probability at least $1-\delta$,

$$\sum_{t=1}^{T}\big(\|\widehat{\theta}_t - \theta_*\|_{V_t} \cdot \|x_t\|_{V_t^{-1}}\big) \leq \sqrt{T}\cdot G_0(T,\delta)G_1(T) = \mathcal{O}(\sqrt{T}\log T). \tag{45}$$

Proof (Lemma 4) By Lemma 11 of [17], the fact that $\|x_{t,a}\|_2 \leq 1$ and $\lambda \geq 1$, and the Cauchy-Schwarz inequality, we have

$$\sum_{t=1}^{T} \|x_t\|_{V_t{}^{-1}} \leq \sqrt{T} \cdot G_1(T). \tag{46}$$

By moving the term $\|\widehat{\theta}_t - \theta_*\|_{V_t}$ outside the summation in (45) and then applying (44), we obtain

$$\sum_{t=1}^{T} \Big(\|\widehat{\theta}_t - \theta_*\|_{V_t} \cdot \|x_t\|_{V_t^{-1}} \Big) = \sqrt{T} \cdot G_0(T, \delta) G_1(T). \tag{47}$$

This implies that (45) indeed holds. □

Lemma 5

$$\sum_{t=1}^{T} \alpha(t) \|x_t\|_{V_t{}^{-1}}^2 \leq \alpha(T) \big(G_1(T)\big)^2 = \mathcal{O}(\alpha(T) \log T). \tag{48}$$

Proof (Lemma 5) by Lemma 11 of [17] and the fact that $\|x_{t,a}\|_2 \leq 1$ and $\lambda \geq 1$, we know

$$\sum_{t=1}^{T} \|x_t\|_{V_t{}^{-1}}^2 \leq \big(G_1(T)\big)^2 = \mathcal{O}(\log T). \tag{49}$$

By moving the bias term outside the summation (48), we have

$$\sum_{t=1}^{T} \alpha(t) \|x_t\|_{V_t^{-1}}^2 \leq \alpha(T) \sum_{t=1}^{T} \|x_t\|_{V_t^{-1}}^2 \leq \alpha(T) \big(G_1(T)\big)^2 = \mathcal{O}(\alpha(T) \log T). \tag{50}$$

□

Remark 4 Note that the first inequality in (50) might seem fairly conservative. However, it cannot be improved as can be seen from the following example: Define a function $f : \mathbb{N} \to \mathbb{R}$ as: $f(t) = k + \frac{1}{t}$ if $t = 2^k$, and $f(t) = \frac{1}{t}$, otherwise. It is easy to check that $\log T \leq \sum_{t=1}^{T} f(t) \leq 2 \log T$, and $\sum_{t=1}^{T} \alpha(t) f(t) = \theta(\alpha(T) \log T)$.

Now we are ready to prove Theorem 1.

Proof (Theorem 1) By combining (11) and Lemmas 3-5, we know

$$\mathcal{R}(T) = \sum_{t=1}^{T} R_t \leq \big(G_0(T, \delta)\big)^2 \cdot \sum_{t=1}^{T} \frac{1}{2\alpha(t)} + \sqrt{T} G_0(T, \delta) G_1(T) + \frac{1}{2} \alpha(T) \big(G_1(T)\big)^2. \tag{51}$$

By choosing $\alpha(t) = \sqrt{t}$, the regret bound is $\mathcal{R}(T) = \mathcal{O}(\sqrt{T} \log T)$. □

E Proof of Lemma 2

By (15),

$$\sum_{s=1}^{t-1} \big(r_s x_s - \mu(x_s^\intercal \bar{\theta}_{t,i}) x_s\big) - \lambda \bar{\theta}_{t,i} + \alpha(t) x_{t,i} = 0, \tag{52}$$

$$\sum_{s=1}^{t-1} \big(r_s x_s - \mu(x_s^\intercal \bar{\theta}_{t,j}) x_s\big) - \lambda \bar{\theta}_{t,j} + \alpha(t) x_{t,j} = 0. \tag{53}$$

Moreover, by the mean value theorem, there exists $\beta_0 \in (0, 1)$ and $\underline{\theta} = \beta_0 \bar{\theta}_{t,i} + (1 - \beta_0) \bar{\theta}_{t,j}$ such that

$$\sum_{s=1}^{t-1} \mu(x_s^\intercal \bar{\theta}_{t,i}) x_s + \lambda \bar{\theta}_{t,i} - \sum_{s=1}^{t-1} \mu(x_s^\intercal \bar{\theta}_{t,j}) x_s - \lambda \bar{\theta}_{t,i} \tag{54}$$

$$= \Big[\sum_{s=1}^{t} \mu'(x_s^\intercal \underline{\theta}) x_s x_s^\intercal + \lambda I \Big] (\bar{\theta}_{t,i} - \bar{\theta}_{t,j}) = U_0 (\bar{\theta}_{t,i} - \bar{\theta}_{t,j}). \tag{55}$$

Multiplying both sides of (52)-(53) by the row vector $(x_{t,i}+x_{t,j})^\intercal {U_0}^{-1}$ yields

$$(x_{t,i}+x_{t,j})^\intercal {U_0}^{-1}\Big(\sum_{s=1}^{t-1}\big(r_s x_s-\mu(x_s^\intercal\bar{\theta}_{t,i})x_s\big)-\lambda\bar{\theta}_{t,i}\Big)+\alpha(t)(x_{t,i}+x_{t,j})^\intercal {U_0}^{-1}x_{t,i}=0, \tag{56}$$

$$(x_{t,i}+x_{t,j})^\intercal {U_0}^{-1}\Big(\sum_{s=1}^{t-1}\big(r_s x_s-\mu(x_s^\intercal\bar{\theta}_{t,j})x_s\big)-\lambda\bar{\theta}_{t,j}\Big)+\alpha(t)(x_{t,i}+x_{t,j})^\intercal {U_0}^{-1}x_{t,j}=0. \tag{57}$$

By combining (56)-(56) and eliminating the common terms, we conclude that

$$(x_{t,i}+x_{t,j})^\intercal(\bar{\theta}_{t,j}-\bar{\theta}_{t,i})+\alpha(t)\|x_{t,i}\|_{{U_0}^{-1}}-\alpha(t)\|x_{t,j}\|_{{U_0}^{-1}}=0. \tag{58}$$

□

F Proof of Theorem 2

For each time t, we denote the estimate of θ without applying the bias term as $\widehat{\theta}_t$, which satisfies the first-order necessary condition $\nabla_\theta(\ell(\mathcal{F}_t;\theta)-\frac{\lambda}{2}\|\theta\|_2^2)|_{\theta=\widehat{\theta}_t}=0$. Equivalently,

$$\sum_{s=1}^{t-1}\big(r_s x_s-\mu(x_s^\intercal\widehat{\theta}_t)x_s\big)-\lambda\widehat{\theta}_t=0. \tag{59}$$

Recall that $V_t=\sum_{s=1}^{t-1}x_s x_s^\intercal+\lambda I$, where I denotes the $d\times d$ identity matrix. Without loss of generality, we may assume that $L_\mu\geq 1$ and $\kappa_\mu\leq 1$ (as these can be easily achieved by adding a constant scaling factor to the link function). Before proving Theorem 2, we first establish several preliminary results.

Lemma 6 For any arm i,

$$\|\widehat{\theta}_t-\bar{\theta}_{t,i}\|_{V_t}\leq\frac{1}{\kappa_\mu}\alpha(t)\|x_{t,i}\|_{{V_t}^{-1}}. \tag{60}$$

Proof (Lemma 6) For each time t, define a “helper function" $Z_t(\cdot):\mathbb{R}^d\rightarrow\mathbb{R}^d$ by

$$Z_t(\theta):=\sum_{s=1}^{t-1}\big(\mu(x_s^\intercal\theta)-\mu(x_s^\intercal\theta_*)\big)x_s+\lambda(\theta-\theta_*). \tag{61}$$

It is easy to verify that $Z_t(\theta_*)=0$. By (15),

$$Z_t(\widehat{\theta}_t)-Z_t(\bar{\theta}_{t,i})=\sum_{s=1}^{t-1}\Big(\big(\mu(x_s^\intercal\widehat{\theta}_t)-\mu(x_s^\intercal\bar{\theta}_{t,i})\big)x_s\Big)+\lambda(\widehat{\theta}_t-\bar{\theta}_{t,i})=-\alpha(t)x_{t,i}. \tag{62}$$

Next, we consider upper and lower bounds on the inner product of $\widehat{\theta}_t-\bar{\theta}_{t,i}$ and $Z_t(\widehat{\theta}_t)-Z_t(\bar{\theta}_{t,i})$. For the upper bound,

$$(\widehat{\theta}_t-\bar{\theta}_{t,i})^\intercal(Z_t(\widehat{\theta}_t)-Z_t(\bar{\theta}_{t,i}))=-\alpha(t)(\widehat{\theta}_t-\bar{\theta}_{t,i})^\intercal x_{t,i} \tag{63}$$

$$\leq\alpha(t)\|\widehat{\theta}_t-\bar{\theta}_{t,i}\|_{V_t}\cdot\|x_{t,i}\|_{{V_t}^{-1}}, \tag{64}$$

where (63) follows from (62), and (64) holds by the Cauchy-Schwarz inequality. Similarly, we obtain a lower bound as

$$(\widehat{\theta}_t-\bar{\theta}_{t,i})^\intercal(Z_t(\widehat{\theta}_t)-Z_t(\bar{\theta}_{t,i}))\geq(\widehat{\theta}_t-\bar{\theta}_{t,i})^\intercal\kappa_\mu V_t(\widehat{\theta}_t-\bar{\theta}_{t,i}) \tag{65}$$

$$=\kappa_\mu\|\widehat{\theta}_t-\bar{\theta}_{t,i}\|_{V_t}^2. \tag{66}$$

By combining (64) and (66), we conclude that (60) indeed holds. □

Based on Lemma 2, given $\bar{\theta}_t$ and $\bar{\theta}_{t,a_t^*}$, there must exist a constant $\beta_0\in(0,1)$, satisfying $\underline{\theta}=\beta_0\bar{\theta}_t+(1-\beta_0)\bar{\theta}_{t,a_t^*}$, such that

$$(x_t+x_{t,a_t^*})^\intercal(\bar{\theta}_{t,a_t^*}-\bar{\theta}_t)+\alpha(t)\|x_t\|_{U^{-1}}-\alpha(t)\|x_{t,a_t^*}\|_{U^{-1}}=0, \tag{67}$$

where the matrix U is defined as

$$U = \sum_{s=1}^{t-1} \mu'(x_s^\intercal \underline{\theta}) x_s x_s^\intercal + \lambda I. \tag{68}$$

For ease of notation, we define the L_2-regularized log-likelihood as

$$\ell_\lambda(\mathcal{F}_t; \theta) := \ell(\mathcal{F}_t; \theta) - \frac{\lambda}{2} \|\theta\|_2^2 \tag{69}$$

Lemma 7 For any arm i, the L_2-regularized log-likelihood satisfies

$$\ell_\lambda(\mathcal{F}_t; \bar{\theta}_t) - \ell_\lambda(\mathcal{F}_t; \bar{\theta}_{t,i}) \le \frac{L_\mu}{2\kappa_\mu^2} \cdot \alpha(t)^2 \|x_{t,i}\|^2_{V_t{}^{-1}}. \tag{70}$$

Proof (Lemma 7) We quantify the difference in log-likelihood under $\bar{\theta}_t$ and $\bar{\theta}_{t,i}$ with the help of $\widehat{\theta}_t$. Denoting the Hessian of $\ell_\lambda(\mathcal{F}_t; \theta)$ with respect to θ by $H_\ell(\theta)$, we have

$$H_\ell(\theta) = \sum_{s=1}^{t-1} -\mu'(x_s^\intercal \theta) x_s x_s^\intercal - \lambda I, \tag{71}$$

and hence $H_\ell(\theta)$ is negative-definite. By the boundedness of μ', we also know that

$$H_\ell(\theta) \succeq -L_\mu(V_t - \lambda I) - \lambda I \succeq -L_\mu V_t. \tag{72}$$

Consequently,

$$\begin{aligned}
\ell_\lambda(\mathcal{F}_t; \bar{\theta}_t) - \ell_\lambda(\mathcal{F}_t; \bar{\theta}_{t,i}) &= \big(\ell_\lambda(\mathcal{F}_t; \bar{\theta}_t) - \ell_\lambda(\mathcal{F}_t; \widehat{\theta}_t)\big) + \big(\ell_\lambda(\mathcal{F}_t; \widehat{\theta}_t) - \ell_\lambda(\mathcal{F}_t; \bar{\theta}_{t,i})\big) && (73)\\
&= \underbrace{\frac{1}{2}(\bar{\theta}_t - \widehat{\theta}_t)^\intercal H_\ell(\theta')(\bar{\theta}_t - \widehat{\theta}_t)}_{\le 0} - \frac{1}{2}(\bar{\theta}_{t,i} - \widehat{\theta}_t)^\intercal H_\ell(\theta'')(\bar{\theta}_{t,i} - \widehat{\theta}_t) && (74)\\
&\le \frac{1}{2} L_\mu \cdot \|\bar{\theta}_{t,i} - \widehat{\theta}_t\|^2_{V_t} && (75)\\
&\le \frac{L_\mu}{2\kappa_\mu^2} \cdot \alpha(t)^2 \|x_{t,i}\|^2_{V_t{}^{-1}}, && (76)
\end{aligned}$$

where (74) follows from (59) and the Taylor expansion of $\ell_\lambda(\mathcal{F}_t; \theta)$ at $\theta = \widehat{\theta}_t$ up to the quadratic term (with $\theta' = \xi' \bar{\theta}_t + (1 - \xi')\widehat{\theta}_t$ and $\theta'' = \xi'' \bar{\theta}_{t,i} + (1 - \xi'')\widehat{\theta}_t$ for some $\xi', \xi'' \in [0, 1]$), (75) holds by (72), and (76) is a direct result of Lemma 6. □

As will be seen presently, the regret bound involves several quantities concerning the norms of the differences in the estimators of θ and the norms of the context vectors. Recalling that $\widehat{\theta}_t$ denotes the estimator of θ without applying the bias term, we first establish several useful inequalities in the following Lemma 8. For ease of exposition, we discuss the Loewner order of the two key matrices V_t and U. For any two symmetric matrices A, B, we write $A \preceq B$ if $B - A$ is a positive semi-definite matrix. Similarly, we write $A \succeq B$ if $A - B$ is positive semi-definite. By (68), the boundedness of the first-order derivative of μ and that $L_\mu \ge 1$, we know

$$U \preceq L_\mu(V_t - \lambda I) + \lambda I = L_\mu V_t + (\lambda - L_\mu \lambda) I \preceq L_\mu V_t. \tag{77}$$

Similarly, by the fact that $\kappa_\mu \le 1$, we have

$$U \succeq \kappa_\mu(V_t - \lambda I) + \lambda I = \kappa_\mu V_t + (\lambda - \kappa_\mu \lambda) I \succeq \kappa_\mu V_t. \tag{78}$$

Lemma 8 The following inequalities hold with probability one:

$$\|\widehat{\theta}_t - \bar{\theta}_t\|_U \cdot \|x_{t,a_t^*}\|_{U^{-1}} \le \frac{L_\mu^2}{\kappa_\mu} \alpha(t) \|x_t\|_{U^{-1}} \cdot \|x_{t,a_t^*}\|_{U^{-1}}, \tag{79}$$

$$\|\theta_* - \widehat{\theta}_t\|_U \cdot \|x_{t,a_t^*}\|_{U^{-1}} \le L_\mu \|\theta_* - \widehat{\theta}_t\|_{V_t} \cdot \|x_{t,a_t^*}\|_{U^{-1}}, \tag{80}$$

$$\|\theta_* - \widehat{\theta}_t\|_U \cdot \|x_t\|_{U^{-1}} \le \frac{L_\mu}{\kappa_\mu} \|\theta_* - \widehat{\theta}_t\|_{V_t} \cdot \|x_t\|_{V_t{}^{-1}}, \tag{81}$$

$$\|\widehat{\theta}_t - \bar{\theta}_{t,a_t^*}\|_U \cdot \|x_t\|_{U^{-1}} \le \frac{L_\mu^2}{\kappa_\mu} \alpha(t) \|x_t\|_{U^{-1}} \cdot \|x_{t,a_t^*}\|_{U^{-1}}. \tag{82}$$

Proof (Lemma 8) For (79), it can be shown that

$$\|\widehat{\theta}_t - \bar{\theta}_t\|_U \cdot \|x_{t,a_t^*}\|_{U^{-1}} \leq L_\mu \|\widehat{\theta}_t - \bar{\theta}_t\|_{V_t} \cdot \|x_{t,a_t^*}\|_{U^{-1}} \tag{83}$$

$$\leq L_\mu \Big(\frac{1}{\kappa_\mu}\alpha(t)\|x_{t,a_t^*}\|_{V_t{}^{-1}}\Big)\|x_{t,a_t^*}\|_{U^{-1}} \tag{84}$$

$$\leq \frac{L_\mu^2}{\kappa_\mu}\alpha(t)\|x_t\|_{U^{-1}} \cdot \|x_{t,a_t^*}\|_{U^{-1}}, \tag{85}$$

where (83) and (85) hold by the definition of U in (68) and the boundedness of the first-order derivative of μ, and (84) is a direct result of Lemma 6. Similarly, (82) can be shown by following the same procedure as (83)-(85). For (80) and (81), by the definition of U and the boundedness of the first-order derivative of μ, it is easy to verify that (80) and (81) indeed hold. □

Now we are ready to prove Theorem 2.

Proof (Theorem 2) To begin with, recall from Section 4.1 that at each time t, GLM-RBMLE selects an arm from the index set $\bar{\mathcal{A}}_t''$ defined as

$$\bar{\mathcal{A}}_t'' := \operatorname*{argmax}_{1\leq a\leq K} \Big\{\ell(\mathcal{F}_t; \bar{\theta}_{t,a}) + \eta(t)\alpha(t)\cdot \bar{\theta}_{t,a}^\intercal x_{t,a} - \frac{\lambda}{2}\|\bar{\theta}_{t,a}\|_2^2\Big\}, \tag{86}$$

Recall that the *immediate regret* is defined as $R_t = \mu(\theta_*^\intercal x_{t,a_t^*}) - \mu(\theta_*^\intercal x_t)$. By (67), under the GLM-RBMLE index in (86),

$$0 \geq \Big(\bar{\theta}_{t,a_t^*}^\intercal x_{t,a_t^*} + \frac{\ell_\lambda(\mathcal{F}_t; \bar{\theta}_{t,a_t^*})}{\eta(t)\alpha(t)}\Big) - \Big(\bar{\theta}_t^\intercal x_t + \frac{\ell_\lambda(\mathcal{F}_t; \bar{\theta}_t)}{\eta(t)\alpha(t)}\Big) \tag{87}$$

$$= \bar{\theta}_t^\intercal x_{t,a_t^*} - \bar{\theta}_{t,a_t^*}^\intercal x_t \tag{88}$$

$$- \alpha(t)\|x_t\|_{U^{-1}} + \alpha(t)\|x_{t,a_t^*}\|_{U^{-1}} - \frac{\ell_\lambda(\mathcal{F}_t; \bar{\theta}_t)}{\eta(t)\alpha(t)} + \frac{\ell_\lambda(\mathcal{F}_t; \bar{\theta}_{t,a_t^*})}{\eta(t)\alpha(t)}. \tag{89}$$

Hence,

$$R_t \leq L_\mu \cdot (\theta_*^\intercal x_{t,a_t^*} - \theta_*^\intercal x_t) \tag{90}$$

$$= L_\mu \cdot \big[(\theta_* - \bar{\theta}_t)^\intercal x_{t,a_t^*} - \theta_*^\intercal x_t + \bar{\theta}_t^\intercal x_{t,a_t^*}\big] \tag{91}$$

$$\leq L_\mu \cdot \Bigg[(\theta_* - \bar{\theta}_t)^\intercal x_{t,a_t^*} - \theta_*^\intercal x_t \tag{92}$$

$$+ \Big(x_t^\intercal \bar{\theta}_{t,a_t^*} + \alpha(t)\|x_t\|_{U^{-1}} - \alpha(t)\|x_{t,a_t^*}\|_{U^{-1}} + \frac{\ell_\lambda(\mathcal{F}_t; \bar{\theta}_t)}{\eta(t)\alpha(t)} - \frac{\ell_\lambda(\mathcal{F}_t; \bar{\theta}_{t,a_t^*})}{\eta(t)\alpha(t)}\Big)\Bigg] \tag{93}$$

$$= L_\mu \cdot \Bigg[(\theta_* - \bar{\theta}_t)^\intercal x_{t,a_t^*} + (\bar{\theta}_{t,a_t^*} - \theta_*)^\intercal x_t \tag{94}$$

$$+ \alpha(t)\|x_t\|_{U^{-1}} - \alpha(t)\|x_{t,a_t^*}\|_{U^{-1}} + \frac{\ell_\lambda(\mathcal{F}_t; \bar{\theta}_t)}{\eta(t)\alpha(t)} - \frac{\ell_\lambda(\mathcal{F}_t; \bar{\theta}_{t,a_t^*})}{\eta(t)\alpha(t)}\Bigg] \tag{95}$$

$$\leq L_\mu \cdot \Bigg[\|\theta_* - \bar{\theta}_t\|_U \cdot \|x_{t,a_t^*}\|_{U^{-1}} + \|\bar{\theta}_{t,a_t^*} - \theta_*\|_U \cdot \|x_t\|_{U^{-1}} \tag{96}$$

$$+ \alpha(t)\|x_t\|_{U^{-1}} - \alpha(t)\|x_{t,a_t^*}\|_{U^{-1}} + \frac{\ell_\lambda(\mathcal{F}_t; \bar{\theta}_t)}{\eta(t)\alpha(t)} - \frac{\ell_\lambda(\mathcal{F}_t; \bar{\theta}_{t,a_t^*})}{\eta(t)\alpha(t)}\Bigg], \tag{97}$$

where (90) follows from the the boundedness of the derivative of μ, (92)-(95) hold by (87)-(89), and (96)-(96) is a direct result of the Cauchy-Schwarz inequality with respect to the norm induced by the matrix U. Next, we provide an upper bound for each term in (96)-(97):

- $\|\theta_* - \bar{\theta}_t\|_U \cdot \|x_{t,a_t^*}\|_{U^{-1}}$: We can obtain an upper bound by applying the Cauchy-Schwarz inequality and (79)-(80) in Lemma 8, as

$$\|\theta_* - \bar{\theta}_t\|_U \cdot \|x_{t,a_t^*}\|_{U^{-1}} \leq \big(\|\theta_* - \widehat{\theta}_t\|_U + \|\widehat{\theta}_t - \bar{\theta}_t\|_U\big)\|x_{t,a_t^*}\|_{U^{-1}} \tag{98}$$

$$\leq L_\mu\|\theta_* - \widehat{\theta}_t\|_{V_t} \cdot \|x_{t,a_t^*}\|_{U^{-1}} + \frac{L_\mu^2}{\kappa_\mu}\alpha(t)\|x_t\|_{U^{-1}} \cdot \|x_{t,a_t^*}\|_{U^{-1}}. \tag{99}$$

- $\|\bar{\theta}_{t,a_t^*} - \theta_*\|_U \cdot \|x_t\|_{U^{-1}}$: By the Cauchy-Schwarz inequality and (82) in Lemma 8,

$$\|\bar{\theta}_{t,a_t^*} - \theta_*\|_U \cdot \|x_t\|_{U^{-1}} \le \Big(\|\bar{\theta}_{t,a_t^*} - \widehat{\theta}_t\|_U + \|\widehat{\theta}_t - \theta_*\|_U\Big)\|x_t\|_{U^{-1}} \tag{100}$$

$$\le \frac{L_\mu^2}{\kappa_\mu}\alpha(t)\|x_t\|_{U^{-1}} \cdot \|x_{t,a_t^*}\|_{U^{-1}} + \frac{L_\mu}{\kappa_\mu}\|\widehat{\theta}_t - \theta_*\|_U \cdot \|x_t\|_{V_t^{-1}}. \tag{101}$$

- $\alpha(t)\|x_t\|_{U^{-1}}$: It is easy to verify that

$$\alpha(t)\|x_t\|_{U^{-1}} \le \frac{1}{\kappa_\mu^2}\alpha(t)\|x_t\|_{V_t^{-1}}. \tag{102}$$

- $\frac{\ell_\lambda(\mathcal{F}_t;\bar{\theta}_t)}{\eta(t)\alpha(t)} - \frac{\ell_\lambda(\mathcal{F}_t;\bar{\theta}_{t,a_t^*})}{\eta(t)\alpha(t)}$: By Lemma 7, we know

$$\frac{\ell_\lambda(\mathcal{F}_t;\bar{\theta}_t)}{\eta(t)\alpha(t)} - \frac{\ell_\lambda(\mathcal{F}_t;\bar{\theta}_{t,a_t^*})}{\eta(t)\alpha(t)} \le \frac{L_\mu}{2\eta(t)\kappa_\mu^2} \cdot \alpha(t)\|x_{t,a_t^*}\|^2_{V_t^{-1}} \le \frac{L_\mu^3}{2\eta(t)\kappa_\mu^2} \cdot \alpha(t)\|x_{t,a_t^*}\|^2_{U^{-1}}. \tag{103}$$

By combining (96)-(97) and the above upper bounds, we have

$$R_t \le L_\mu\bigg[\Big(\big(\frac{L_\mu^3}{2\kappa_\mu^2\eta(t)} - 1\big)\alpha(t)\Big) \cdot \|x_{t,a_t^*}\|^2_{U^{-1}} + \Big(\frac{2L_\mu^2}{\kappa_\mu}\alpha(t)\|x_t\|_{U^{-1}} + L_\mu\|\theta_* - \widehat{\theta}_t\|_{V_t}\Big)\|x_{t,a_t^*}\|_{U^{-1}} \tag{104}$$

$$+ \Big(\frac{L_\mu}{\kappa_\mu}\|\widehat{\theta}_t - \theta_*\|_{V_t} \cdot \|x_t\|_{V_t^{-1}} + \frac{1}{\kappa_\mu^2}\alpha(t)\|x_t\|_{V_t^{-1}}\Big)\bigg]. \tag{105}$$

Note that (104)-(105) can be interpreted as a quadratic function of $\|x_{t,a_t^*}\|_{U^{-1}}$. Recall that $T_0 := \min\{t \in \mathbb{N} : \frac{L_\mu^3}{2\kappa_\mu^2\eta(t)} < 1\}$. Therefore, for any $t \ge T_0$, by completing the square,

$$R_t \le L_\mu\bigg[\frac{\alpha(t)}{4(1 - \frac{L_\mu^3}{2\kappa_\mu^2\eta(t)})}\Big(\frac{2L_\mu^2}{\kappa_\mu}\|x_t\|_{U^{-1}} + L_\mu\frac{\|\theta_* - \widehat{\theta}_t\|_{V_t}}{\alpha(t)}\Big)^2 \tag{106}$$

$$+ \frac{L_\mu}{\kappa_\mu}\|\widehat{\theta}_t - \theta_*\|_{V_t} \cdot \|x_t\|_{V_t^{-1}} + \frac{1}{\kappa_\mu^2}\alpha(t)\|x_t\|^2_{V_t^{-1}}\bigg]. \tag{107}$$

Based on (106)-(107), to bound the cumulative regret, we need the following properties. Recall that $G_1(t)$ and $G_2(t,\delta)$ are defined as

$$G_1(t) := \sqrt{2d\log\big(\frac{\lambda + t}{d}\big)} \tag{108}$$

$$G_2(t,\delta) := \frac{\sigma}{\kappa_\mu}\sqrt{\frac{d}{2}\log(1 + \frac{2t}{d}) + \log\big(\frac{1}{\delta}\big)}. \tag{109}$$

- Note that by Lemma 11 of [17] and the fact that $\|x_{t,a}\|_2 \le 1$ and $\lambda \ge 1$,

$$\sum_{t=1}^{T}\|x_t\|^2_{V_t^{-1}} \le \big(G_1(T)\big)^2. \tag{110}$$

  Moreover, (110) also implies that

$$\sum_{t=1}^{T}\alpha(t)\|x_t\|^2_{V_t^{-1}} \le \alpha(T)\big(G_1(T)\big)^2. \tag{111}$$

- By combining (110) and the Cauchy-Schwarz inequality, we have

$$\sum_{t=1}^{T}\|x_t\|_{V_t^{-1}} \le \sqrt{T} \cdot G_1(T). \tag{112}$$

- By Lemma 3 in [15] and since the minimum eigenvalue $\lambda_{\min}(V_t) \geq \lambda \geq 1$, for any $\delta \in [1/T, 1)$, we know with probability at least $1 - \delta$, the following result holds:

$$\|\widehat{\theta}_t - \theta_*\|_{V_t} \leq G_2(t, \delta), \quad \forall t \in \mathbb{N}. \tag{113}$$

- By combining (112) and (113), we thereby know that with probability at least $1 - \delta$,

$$\sum_{t=1}^{T} \|x_t\|_{V_t{}^{-1}} \cdot \|\widehat{\theta}_t - \theta_*\|_{V_t} \leq \sqrt{T} \cdot G_1(T) G_2(T, \delta). \tag{114}$$

- Based on (113), we further know that with probability at least $1 - \delta$,

$$\sum_{t=1}^{T} \frac{\|\widehat{\theta}_t - \theta_*\|_{V_t}^2}{\alpha(t)} \leq \big(G_2(T, \delta)\big)^2 \cdot \sum_{t=1}^{T} \frac{1}{\alpha(t)}. \tag{115}$$

Summing up, by (106)-(115), the cumulative regret can be upper bounded as follows: With probability at least $1 - \delta$,

$$\sum_{t=1}^{T} R_t \leq T_0 + \sum_{t=T_0+1}^{T} C_1 \alpha(t) \|x_t\|_{V_t{}^{-1}}^2 + C_2 \|x_t\|_{V_t{}^{-1}} \cdot \|\widehat{\theta}_t - \theta_*\|_{V_t} + C_3 \frac{\|\widehat{\theta}_t - \theta_*\|_{V_t}^2}{\alpha(t)} \tag{116}$$

$$\leq T_0 + C_1 \alpha(T) \big(G_1(T)\big)^2 + C_2 \sqrt{T} G_1(T) G_2(T, \delta) + C_3 \big(G_2(T, \delta)\big)^2 \sum_{t=1}^{T} \frac{1}{\alpha(t)}, \tag{117}$$

where $C_1 := \frac{2L_\mu^4}{k_\mu^4} + \frac{1}{k_\mu^2}$, $C_2 := \frac{2L_\mu^3}{\kappa_\mu^2} + \frac{L_\mu}{\kappa_\mu}$, and $C_3 := \frac{L_\mu^2}{2}$. Therefore, if $\alpha(t) = \Omega(\sqrt{t})$, then $\mathcal{R}(T) = \mathcal{O}(\alpha(T) \log T)$; Otherwise, if $\alpha(t) = \mathcal{O}(\sqrt{t})$, then $\mathcal{R}(T) = \mathcal{O}\big((\sum_{t=1}^{T} \frac{1}{\alpha(t)}) \log T\big)$. Hence, by choosing $\alpha(t) = \sqrt{t}$, we obtain a cumulative regret bound of $\mathcal{R}(T) = \mathcal{O}(\sqrt{T} \log T)$. □

G Additional Experimental Results

In this section, we present the additional experimental results for both linear bandits and the generalized case. Throughout the experiments, we set the random seed to be 46.

G.1 Linear Bandits

To begin with, Tables 3, 4, and 5 present the mean, standard deviation, and quantiles of the experiments described in Figures 1(b), 1(c), and 1(d), respectively. Similar to what we observed from Table 1, LinRBMLE still exhibits better robustness than VIDS and most of the other benchmark methods under both static contexts and time-varying contexts. On the other hand, similar to Table 2, Table 6 presents the average computation time per decision under time-varying contexts and different K and d. We observe that the scalability of LinRBMLE in terms of number of actions and context dimension is still preserved in the scenario of time-varying contexts.

G.2 Generalized Linear Bandits

For the generalized linear bandits, we perform a similar study on the effectiveness, efficiency, and scalability of GLM-RBMLE and the popular benchmark methods. The benchmark methods that are compared with GLM-RBMLE include UCB-GLM [15] and Laplace-TS [11] (Algorithm 3 in [11]). The configurations of the three methods are as follows. We use $\alpha(t) = \sqrt{t}$, $\eta(t) = 1 + \log t$, and $\lambda = 1$ for GLM-RBMLE, as suggested in Section 4. Under UCB-GLM, after τ rounds of initial random selection, the arm with the largest $x_{t,a}^\intercal \widehat{\theta}_t + \chi \|x_{t,a}\|_{V_t{}^{-1}}$ is selected at each time t. As suggested by [15], we take $\chi = \frac{\sigma}{\kappa_\mu} \sqrt{\frac{d}{2} \log(1 + 2T/d) + \log(1/\delta)}$ with $\delta = 0.1$, and let $\tau = K$. For Laplace-TS, we set the regularization parameter to be 1. Throughout the experiments of the generalized linear model, we consider the *logistic* link function, i.e. $\mu(z) = 1/(1 + \exp(-z))$, for

Table 3: Statistics of the final cumulative regret in Figure 1(b). The best and the second-best are highlighted. ‘Q’ and “Std.Dev” stand for quantile and standard deviation of the total cumulative regret over 50 trails, respectively. All the values displayed here are scaled by 0.01 for more compact notations.

Alg.	RBMLE	LinUCB	BUCB	GPUCB	GPUCBT	KG	KG*	LinTS	VIDS
Mean	1.86	2.72	4.66	3.77	**0.86**	19.14	2.81	13.49	**0.83**
Std.Dev	**0.45**	10.64	14.63	1.42	**0.65**	35.38	8.37	2.10	1.30
Q.10	1.48	**0.05**	0.09	2.08	0.38	**0.04**	0.09	10.51	0.21
Q.25	1.63	**0.06**	0.10	2.72	0.49	**0.05**	0.12	12.23	0.30
Q.50	1.77	**0.12**	0.13	3.73	0.66	**0.10**	0.16	13.70	0.43
Q.75	1.99	0.36	**0.27**	4.35	0.91	18.06	**0.26**	14.92	0.55
Q.90	2.39	2.83	5.64	6.06	**1.64**	87.14	6.58	16.16	**1.22**
Q.95	**2.55**	8.86	39.66	6.64	**2.06**	100.66	19.38	16.64	4.57

Table 4: Statistics of the final cumulative regret in Figure 1(c). The best and the second-best are highlighted. ‘Q’ and “Std.Dev” stand for quantile and standard deviation of the total cumulative regret over 50 trails, respectively. All the values displayed here are scaled by 0.01 for more compact notations.

Alg.	RBMLE	LinUCB	BUCB	GPUCB	GPUCBT	KG	KG*	LinTS	VIDS
Mean	0.41	0.40	**0.40**	0.52	**0.38**	0.41	0.44	9.17	20.01
Std.Dev	0.17	0.19	0.15	**0.14**	0.15	**0.14**	0.16	0.25	0.65
Q.10	**0.25**	0.25	**0.24**	0.38	0.25	0.26	0.28	8.87	19.20
Q.25	0.30	0.28	**0.27**	0.43	**0.27**	0.32	0.35	9.01	19.61
Q.50	0.37	**0.35**	0.36	0.50	**0.34**	0.38	0.40	9.15	20.05
Q.75	0.47	0.47	0.49	0.56	**0.44**	**0.46**	0.53	9.25	20.35
Q.90	**0.57**	0.60	0.64	0.64	**0.51**	0.61	0.63	9.49	20.75
Q.95	**0.63**	**0.62**	0.69	0.72	0.65	0.71	0.70	9.73	21.09

Table 5: Statistics of the final cumulative regret in Figure 1(d). The best and the second-best are highlighted. ‘Q’ and “Std.Dev” stand for quantile and standard deviation of the total cumulative regret over 50 trails, respectively. All the values displayed here are scaled by 0.01 for more compact notations.

Alg.	RBMLE	LinUCB	BUCB	GPUCB	GPUCBT	KG	KG*	LinTS	VIDS
Mean	**0.40**	0.44	0.44	0.52	**0.40**	0.45	0.46	9.48	19.85
Std.Dev	0.19	0.18	0.19	**0.11**	**0.12**	0.18	0.16	0.32	0.66
Q.10	**0.21**	0.25	**0.23**	0.40	0.26	0.25	0.29	8.95	19.24
Q.25	**0.30**	0.32	0.30	0.43	0.31	**0.29**	0.33	9.34	19.37
Q.50	**0.39**	0.43	0.41	0.53	**0.39**	0.45	0.43	9.53	19.70
Q.75	**0.46**	0.53	0.57	0.62	**0.48**	0.54	0.56	9.70	20.12
Q.90	**0.52**	0.60	0.62	0.67	**0.55**	0.65	0.64	9.89	20.83
Q.95	**0.70**	0.77	**0.70**	**0.70**	**0.59**	0.74	0.75	9.92	21.02

Table 6: Average computation time per decision in time-varying contexts under different values of K and d. All numbers are averaged over 50 trials with $T = 10^3$ and in 10^{-4} seconds. The best one is highlighted.

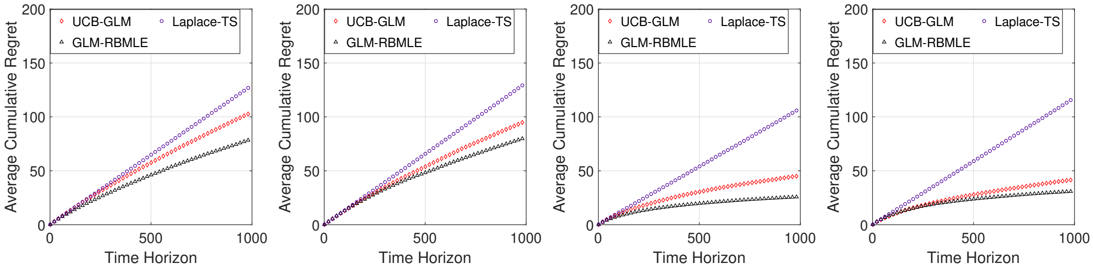
Alg.	RBMLE	UCB	BUCB	GPUCB	GPUCBT	KG	KG*	LinTS	VIDS
K=10	**6.02**	7.32	67.42	7.35	6.87	83.52	22.84	27.81	1500.03
K=20	38.78	47.65	486.22	49.28	46.26	604.77	133.90	**38.46**	2390.61
K=30	67.94	81.90	827.39	83.96	78.32	1021.50	214.76	**27.14**	2601.90
d = 10	**2.18**	2.69	20.00	2.66	2.47	24.86	10.04	4.46	1300.33
d = 20	**3.55**	4.56	30.65	4.23	4.02	37.85	15.67	25.25	2157.48
d = 30	**4.87**	5.86	37.59	5.86	5.52	46.30	19.83	54.48	2818.37

all $z \in \mathbb{R}$. Similar to the experiments for LinRBMLE, for each comparison we consider both static contexts as well as time-varying contexts. The comparison contains 50 trials of experiments and T rounds in each experiment. As the algorithms are computationally more intense for general linear bandits than for those for linear bandits, the time horizon is reduced to $T = 10^3$ in the experiments for the generalized linear bandits.

Effectiveness. Figure 3 and Tables 7-10 show the effectiveness of GLM-RBMLE in terms of cumulative regret. Under both static and time-varying contexts, GLM-RBMLE achieve the best mean regret performance in all the four configurations. Similar to LinRBMLE, based on the results of standard deviation and regret quantiles, GLM-RBMLE also exhibits better robustness across sample paths than the two popular benchmark methods. Specifically when contexts are static, GLM-RBMLE has lower standard deviation and 0.95 quantile compare to UCB-GLM and Laplace-TS. We can characterize the statistical stability by standard deviation and quantiles so we give the result that GLM-RBMLE has better stability than others. On the other hand, in Figure 3, Laplace-TS appears to have not converged, but the corresponding regret quantiles provided by Tables 7-10 reveal that this is only because its performance in some trials is much worse than that in other trials.

Efficiency. Figures 4 shows the averaged cumulative regret versus computation time per decision. We observe that GLM-RBMLE achieves the smallest average regret at the cost of a higher computation time compared to UCB-GLM.

Scalability. Table 11 presents computation time per decision as K and d are varied. We observe that under $K = 5$ and $d = 10, 20, 30$, the computation time per decision of GLM-RBMLE and UCB-GLM are comparable and much smaller than that of Laplace-TS. On the other hand, under $d = 5$ and $K = 10, 20, 30$, we also observe that the computation time of GLM-RBMLE is proportional to the number of arms, as indicated by Line 4 of Algorithm 2. It remains an interesting open question how to improve the scalability of GLM-RBMLE in terms of number of arms.



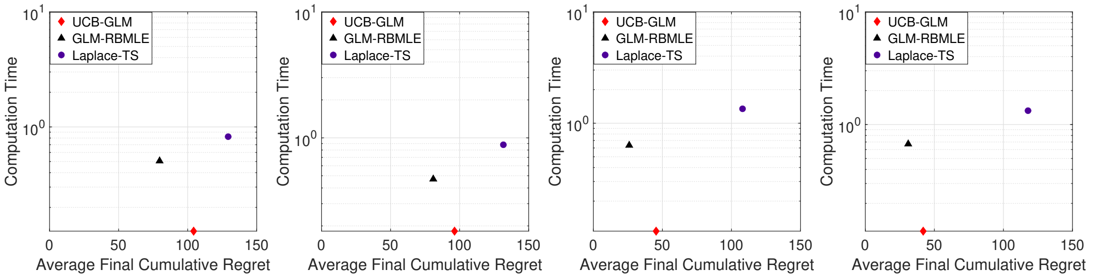
(a)

(b)

(c)

(d)

Figure 3: Cumulative regret averaged over 50 trials with $T = 10^3$ and $K = 10$ on generalized linear bandits: (a) and (b) are under static contexts; (c) and (d) are under time-varying contexts; (a) and (c) are with $\theta_* = (0.3, -0.5, 0.2, -0.7, -0.1)$; (b) and (d) are with $\theta_* = (0.2, -0.8, -0.5, 0.1, 0.1)$.



(a)

(b)

(c)

(d)

Figure 4: Average computation time per decision vs. average final cumulative regret for (a) Figure 3(a); (b) Figure 3(b); (c) Figure 3(c); (d) Figure 3(d).

Table 7: Statistics of the final cumulative regret in Figure 3(a). The best one is highlighted.

Algorithm	**GLM-RBMLE**	**UCB-GLM**	**Laplace-TS**
Mean Final Regret	**79.66**	104.31	129.31
Standard Deviation	**20.86**	31.52	87.92
Quantile .10	55.53	69.60	**11.65**
Quantile .25	65.02	83.95	**58.37**
Quantile .50	**78.56**	106.78	124.07
Quantile .75	**91.83**	125.10	197.74
Quantile .90	**106.03**	140.75	259.94
Quantile .95	**108.87**	153.24	264.79

Table 8: Statistics of the final cumulative regret in Figure 3(b). The best one is highlighted.

Algorithm	**GLM-RBMLE**	**UCB-GLM**	**Laplace-TS**
Mean Final Regret	**80.94**	96.34	131.69
Standard Deviation	**25.38**	30.94	90.99
Quantile .10	58.86	60.90	**11.50**
Quantile .25	63.85	72.74	**53.86**
Quantile .50	**78.12**	95.25	125.30
Quantile .75	**92.96**	119.07	188.75
Quantile .90	**114.39**	131.07	248.53
Quantile .95	**131.95**	143.54	292.39

Table 9: Statistics of the final cumulative regret in Figure 3(c). The best one is highlighted.

Algorithm	**GLM-RBMLE**	**UCB-GLM**	**Laplace-TS**
Mean Final Regret	**25.95**	45.41	107.99
Standard Deviation	9.30	**8.25**	57.90
Quantile .10	**15.92**	35.73	34.02
Quantile .25	**19.68**	38.57	65.03
Quantile .50	**23.11**	44.98	101.27
Quantile .75	**29.84**	51.50	145.02
Quantile .90	**35.71**	55.93	173.38
Quantile .95	**42.36**	60.32	213.75

Table 10: Statistics of the final cumulative regret in Figure 3(d). The best one is highlighted.

Algorithm	**GLM-RBMLE**	**UCB-GLM**	**Laplace-TS**
Mean Final Regret	**31.08**	41.93	117.81
Standard Deviation	13.40	**6.50**	62.84
Quantile .10	**18.81**	34.87	32.58
Quantile .25	**21.64**	37.09	75.70
Quantile .50	**29.48**	41.97	119.50
Quantile .75	**36.09**	45.66	163.05
Quantile .90	**48.10**	51.15	203.09
Quantile .95	55.04	**54.06**	219.47

Table 11: Average computation time per decision for static contexts in generalized linear bandit model, under different values of K and d. All numbers are averaged over 50 trials with $T = 10^2$ and in seconds.

Algorithm	**GLM-RBMLE**	**UCB-GLM**	**Laplace-TS**
$K = 5, d = 10$	0.0275	0.0089	0.0675
$K = 5, d = 20$	0.0407	0.0216	0.2110
$K = 5, d = 30$	0.0519	0.0461	0.3691
$K = 10, d = 5$	0.0406	0.0041	0.0305
$K = 20, d = 5$	0.0823	0.0039	0.0331
$K = 30, d = 5$	0.1225	0.0037	0.0333