Reward-Biased Maximum Likelihood Estimation for Linear Stochastic Bandits

Yu-Heng Hung[1,2], **Ping-Chun Hsieh**[1,2], **Xi Liu**[3], **P. R. Kumar**[3]
[1] Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan
[2] Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
[3] Department of Electrical and Computer Engineering, Texas A&M University, College Station, Texas, USA
j5464654.cs08g@nctu.edu.tw, pinghsieh@nctu.edu.tw, xiliu.tamu@gmail.com, prk@tamu.edu

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 bandits problems as well as generalized linear bandits problems. We develop novel index policies that we prove achieve order-optimality, and show that they achieve empirical performance competitive with the state-of-the-art benchmark methods in extensive experiments. The new policies achieve this with low computation time per pull for linear bandits, and thereby resulting in both favorable regret as well as computational efficiency.

1 Introduction

The problem of decision making for an unknown dynamic system, called stochastic adaptive control (Kumar 1985; Kumar and Varaiya 2015), was examined in the control theory community beginning in the 1950s. It was recognized early on by Feldbaum (Feldbaum 1960a,b) 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 (Borkar and Varaiya 1979) 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 $\text{Prob}(x(t+1)=j|x(t)=i,u(t)=u)=p(i,j;u,\theta_*)$ 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 $\widehat{\theta}_{\text{ML}}(t) \in \text{argmax}_{\theta\in\Theta}\sum_{s=0}^{t-1}\log p(x(s),x(s+1),u(s),\theta)$ denote the Maximum Likelihood (ML) estimate of θ_*, with ties broken according to any fixed priority order. Then apply the action $u(t)=\phi(x(t),\widehat{\theta}_{\text{ML}}(t))$ to the system. It was shown in (Kumar and Becker 1982) that under an irreducibility assumption, the parameter estimates $\widehat{\theta}_{\text{ML}}(t)$ converge to a random limit $\check{\theta}$ satisfying

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

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 (Kumar and Becker 1982). 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 $J(\theta) := \text{Max}_\phi J(\phi,\theta)$ the optimal long-term average reward attainable when the parameter is θ. Then,

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

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, (Kumar and Becker 1982) 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):

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

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 positive 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

$$J(\check{\theta}) \geq J(\theta_*), \tag{4}$$

as shown in (Kumar and Becker 1982, Lemma 4). From (2) and (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 the 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 Upper Confidence Bound (UCB) method of Lai and Robbins (Lai and Robbins 1985), RBMLE has largely remained unexplored vis-à-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 (Liu et al. 2020). However, classical bandits do not allow the incorporation of "context," which is important in various applications (Li et al. 2010; Lu, Pál, and Pál 2010; Chapelle and Li 2011; Li, Karatzoglou, and Gentile 2016; Tewari and Murphy 2017). Therefore, the design and the proofs in (Liu et al. 2020) cannot directly apply to the more structured contextual bandit model. In this paper, we examine the RBMLE method both for linear contextual bandits as well as a more general class of generalized linear bandits. Linear bandits and their variants have been popular models for abstracting the sequential decision making in various applications, such as recommender systems (Li et al. 2010) and medical treatment (Tewari and Murphy 2017).

This paper extends the RBMLE principle and obtains simple index policies for linear contextual bandits as well as their generalizations that have provable order-optimal finite-time regret performance as well as empirical performance competitive with the best currently available. The main contributions of this paper are as follows:

- We extend the RBMLE principle to linear contextual bandits by proposing a specific type of reward-bias term. We introduce into RBMLE the modification of using a Gaussian pseudo-likelihood function, both for usage in situations where the distribution of the rewards is unknown, as well as to derive simple index policies. Different from the popular UCB-based policies, whose indices usually consist of two components: a maximum likelihood estimator and a confidence interval, RBMLE directly incorporates a reward-bias term into the log-likelihood function to guide the exploration instead of using concentration inequalities. The derived RBMLE index is thereby different from the existing indices for linear bandits.
- We show that the so modified RBMLE index attains a regret bound of $\mathcal{O}(\sqrt{T}\log T)$, which is order-optimal (within a logarithmic factor) for general, possibly non-parametric, sub-Gaussian rewards. To the best of our knowledge, this is the first provable finite-time regret guarantee of the classic RBMLE principle for contextual bandits. This bound shaves a factor of $\mathcal{O}(\sqrt{T^{\epsilon}})$ from Thompson Sampling (LinTS) (Agrawal and Goyal 2013), a factor of $\mathcal{O}(\sqrt{\log T})$ from (Chu et al. 2011), and a factor of $\mathcal{O}(\sqrt{\log^3 T})$ from Gaussian Process Upper Confidence Bound (GPUCB) with linear kernels (Srinivas et al. 2010), and achieves the same regret bound as the Information Directed Sampling (IDS) (Kirschner and Krause 2018) and Improved Gaussian Process Upper Confidence Bound (IGP-UCB) (Chowdhury and Gopalan 2017).
- We extend the techniques to the generalized linear models and show that the same regret bound of $\mathcal{O}(\sqrt{T}\log T)$ can still be attained in the general case. This shaves a factor of $\sqrt{\log T}$ from (Filippi et al. 2010), and achieves the same regret bound as UCB-GLM in (Li, Lu, and Zhou 2017).
- We conduct extensive experiments to demonstrate that the proposed RBMLE achieves an empirical regret competitive with the state-of-the-art benchmark methods while being efficient in terms of computation time. Notably, the regret performance of RBMLE is the most robust across different sample paths. The results validate that the proposed algorithm enjoys favorable regret and computation time.

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 $\|\theta_*\|_2 \leq 1$, 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 $\kappa_\mu > 0$.[1] We call this the *generalized linear bandit* problem.

Let $a_t^* := \arg\max_{1\leq i\leq K} \theta_*^\intercal x_{t,i}$ 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

[1] A further discussion about this assumption is in Appendix H.

"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.,

$$\mathbb{E}[\exp(\rho\varepsilon_t)|\mathcal{F}_t] \leq \exp\left(\rho^2\sigma^2/2\right). \tag{6}$$

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 t, the learner takes the following two steps.

1. Let $\bar{\theta}_t = \underset{\theta}{\operatorname{argmax}} \left\{ \ell(\mathcal{F}_t; \theta) + \alpha(t) \max_{a \in [K]} \theta^\intercal x_{t,a} - \frac{\lambda}{2}\|\theta\|_2^2 \right\}$.
2. Choose any arm a_t that maximizes $\bar{\theta}_t^\intercal x_{t,a}$.

The term $\alpha(t) \max_{1 \leq a \leq K} \theta^\intercal x_{t,a}$ 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 $\lambda \geq 1$.

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

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

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 $\left\{ -\sum_{s=1}^{t-1} (\theta^\intercal x_s - r_s)^2 + 2\alpha(t) \cdot \max_{1 \leq a \leq K} \theta^\intercal x_{t,a} - \lambda\|\theta\|_2^2 \right\}$.

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 LinRBMLE 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 LinRBMLE algorithm is

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

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

We summarize the LinRBMLE algorithm in Algorithm 1.

Algorithm 1 LinRBMLE Algorithm

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

Remark 1 Similar to the well-known LinUCB index $\widehat{\theta}_t^\intercal x_{t,i} + \gamma\|x_{t,i}\|_{V_t^{-1}}$ (Li et al. 2010), the LinRBMLE index is also defined as the sum of the least squares estimate and an additional exploration term. Despite this high-level resemblance, LinRBMLE has two salient features: (i) As mentioned in Section 1, the LinRBMLE index is different from the UCB-based indices as it directly incorporates a reward-bias term into the log-likelihood function to guide the exploration instead of using concentration inequalities; (ii) Under LinRBMLE, the ratio between the exploration terms of any two arms i, j is $\|x_{t,i}\|_{V_t^{-1}}^2 / \|x_{t,j}\|_{V_t^{-1}}^2$, which is more contrastive than $\|x_{t,i}\|_{V_t^{-1}} / \|x_{t,j}\|_{V_t^{-1}}$ of LinUCB. With a proper bias term, this design of LinRBMLE implicitly encourages more exploration (since $\|x_{t,i}\|_{V_t^{-1}}$ is a confidence interval). As will be seen in Section 3.2, with a proper bias term (e.g., $\alpha(t) = \sqrt{t}$), this additional exploration does not sacrifice the regret bound. Moreover, as suggested by the regret statistics in Section 5, this design makes LinRBMLE empirically more robust across different sample paths, which is of intrinsic interest.

3.2 Regret Bound for the LinRBMLE Index

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

Lemma 1 Under the standard linear bandit model,

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

The proof of Lemma 1 is in Appendix C.

Remark 2 Lemma 1 highlights the main difference between the analysis of the UCB-based algorithms (e.g., (Abbasi-Yadkori, Pál, and Szepesvári 2011; Chu et al. 2011)) 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 LinRBMLE algorithm; it is proved in Appendix D. Let

$$G_0(t,\delta) := \sigma\sqrt{d\log\big((\lambda+t)/(\lambda\delta)\big)} + \lambda^{\frac{1}{2}}, \tag{12}$$

$$G_1(t) := \sqrt{2d\log\big((\lambda+t)/d\big)} \tag{13}$$

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

$$\begin{aligned}\mathcal{R}(T) = \sum_{t=1}^{T} R_t \le \big(G_0(T,\delta)\big)^2 \cdot \Big(\sum_{t=1}^{T}\frac{1}{2\alpha(t)}\Big) \\ + \sqrt{T}G_0(T,\delta)G_1(T) + \frac{1}{2}\alpha(T)\big(G_1(T)\big)^2.\end{aligned} \tag{14}$$

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

Remark 3 As mentioned in Section 1, in aspect of T, LinRBMLE achieves a better regret bound than several popular benchmark methods, including LinTS (Agrawal and Goyal 2013), SupLinUCB (Chu et al. 2011), and GPUCB with a linear kernel (Srinivas et al. 2010). Moreover, LinRBMLE achieves the same regret bound as that of IDS (Kirschner and Krause 2018) and IGP-UCB (Chowdhury and Gopalan 2017) which are two of the most competitive benchmarks. In Section 5, we show via simulations that LinRBMLE achieves an empirical regret competitive with IDS while being much more computationally efficient. LinRBMLE also has the same regret bound as that of LinUCB (Abbasi-Yadkori, Pál, and Szepesvári 2011). As LinRBMLE addresses exploration in a fundamentally different manner as discussed in Remark 1, the corresponding regret proof also differs from those of the UCB-base policies, as highlighted in Remark 2. From the simulations, we further observe that LinRBMLE significantly outperforms LinUCB in terms of both empirical mean regret and regret statistics. LinRBMLE also matches the lower bound in both d and T for infinite arm set. We give a more detail discussion in Remark 7.

4 RBMLE for Generalized Linear Bandits

4.1 Index Derivation for Generalized Linear Bandits

For the generalized linear case, as before, let $\bar{\theta}_t$ be any maximizer of $\big\{\ell(\mathcal{F}_t;\theta) + \alpha(t)\cdot\max_{1\le a\le K}\theta^\intercal x_{t,a} - \frac{\lambda}{2}\|\theta\|_2^2\big\}$. 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:

$$p(r_t|x_t) = \exp(r_t x_t^\intercal\theta_* - b(x_t^\intercal\theta_*) + c(r_t)), \tag{15}$$

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, $\bar{\theta}_{t,a}$ is the unique solution to

$$\sum_{s=1}^{t-1}\big(r_s x_s - \mu(x_s^\intercal\bar{\theta}_{t,a})x_s\big) - \lambda\bar{\theta}_{t,a} + \alpha(t)x_{t,a} = 0. \tag{16}$$

Note that (15) 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.

Remark 4 The technical reason behind incorporating $\eta(t)$ into GLM-RBMLE is as follows: As will be seen in (101)-(102) in Appendix F, the immediate regret R_t is upper bounded by the value of a quadratic function of $\|x_{t,a_t^*}\|_{U^{-1}}$, and this inequality resembles (37) for the linear case. To further bound the RHS of (101), we need the leading coefficient $L_\mu^3/(2\kappa_\mu^2\eta(t)) - 1$ to be negative. To ensure this, we propose to set $\eta(t)$ to be a positive, strictly increasing function with $\lim_{t\to\infty}\eta(t) = \infty$ such that $L_\mu^3/(2\kappa_\mu^2\eta(t)) < 1$ for all sufficiently large t. For the linear case, we can simply let $\eta(t) = 1$ since $L_\mu = \kappa_\mu = 1$ and $L_\mu^3/2\kappa_\mu^2 < 1$ automatically holds.

Algorithm 2 GLM-RBMLE Algorithm

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

4.2 Regret Bound for GLM-RBMLE for Generalized Linear Bandits

We begin the regret analysis of GLM-RBMLE by introducing the following definitions.

Define $T_0 := \min\{t\in\mathbb{N} : \frac{L_\mu^3}{2\kappa_\mu^2\eta(t)} < \frac{1}{2}\}$. Recall that $G_1(t)$ is defined in (13). For ease of exposition, we also 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 $C_1 := 2L_\mu^4/k_\mu^4 + 1/\sqrt{\kappa_\mu}$, $C_2 := 2L_\mu^3/\kappa_\mu^2 + L_\mu/\kappa_\mu$, and $C_3 := L_\mu^2/2$.

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

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

Therefore, if $\alpha(t) = \Omega(\sqrt{t})$, then $\mathcal{R}(T) = \mathcal{O}(\alpha(T)\log T)$; 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 $\alpha(t) = \sqrt{t}$, $\mathcal{R}(T) = \mathcal{O}(d\sqrt{T}\log T)$.

Remark 5 This bound improves that in (Filippi et al. 2010) by a $\sqrt{\log T}$ factor and is the same as that of UCB-GLM (Li, Lu, and Zhou 2017).

5 Numerical Experiments

To evaluate the 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 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 the 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.

To present a comprehensive numerical study similar to (Russo and Van Roy 2018), the benchmark methods compared include LinUCB (Chu et al. 2011), LinTS (Agrawal and Goyal 2013), Bayes-UCB (BUCB) (Kaufmann, Cappé, and Garivier 2012), GPUCB (Srinivas et al. 2010) and its variant GPUCB-Tuned (GPUCBT) (Russo and Van Roy 2018), Knowledge Gradient (KG) and its variant KG* (Ryzhov, Frazier, and Powell 2010; Ryzhov, Powell, and Frazier 2012; Kamiński 2015), and VIDS (Russo and Van Roy 2018). A detailed review of these methods is presented in Section 6. The values of their hyper-parameters are as follows. For LinRBMLE, as suggested by Theorem 1, we choose $\alpha(t) = \sqrt{t}$ without any hyper-parameter tuning, and $\lambda = 1$ which is a common choice in ridge regression and is not sensitive to the empirical regret. 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 (Kaufmann, Cappé, and Garivier 2012) to choose $c = 0$ for BUCB. Respecting the restrictions in (Agrawal and Goyal 2013), 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 and take $M = 10^4$ in sampling (Algorithm 4 and 6 in (Russo and Van Roy 2018)). 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.

The procedure for generating the synthetic dataset is as follows: (i) All contexts are drawn randomly from $\mathcal{N}(0_d, 10I_d)$ and normalized by their ℓ_2 norm; (ii) At time t, the reward of each arm i is sampled independently from $\mathcal{N}(\mu(\theta_*^\intercal x_{t,i}), 1)$. In each test case, we consider a fixed θ_* and randomly generate the contexts, which lead to different mean rewards across the arms. This scheme for generating the synthetic dataset has been widely adopted in the bandit literature, such as (Abbasi-Yadkori, Pál, and Szepesvári 2011; Dumitrascu, Feng, and Engelhardt 2018; Kirschner and Krause 2018); (iii) As IDS-based approaches are known to be time-consuming, we choose $d = 3$ as suggested by (Kirschner and Krause 2018) for the experiments involving regret comparison in order to finish enough simulation steps within a reasonable amount of time. For the scalability experiments, we reduce the number of rounds T to allow the choice of larger d's.

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 computation time of LinRBMLE is two orders of magnitude smaller, as will be shown in Figure 2. 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. 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 (Russo and Van Roy 2018) are restricted to the static contexts. More statistics of final cumulative regret in Figure 1 are provided in the appendix.

Efficiency. Figure 2 presents the averaged cumulative re-

gret versus average computation time per decision. We observe that LinRBMLE and GPUCBT have points closest to the origin, signifying small regret simultaneously with small computation time, and outperform the other 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. LinRBMLE is slightly better than LinUCB in terms of computation time under various K and d since the calculation of LinUCB index requires an additional square-root operation. 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 Appendix G.

6 Related Work

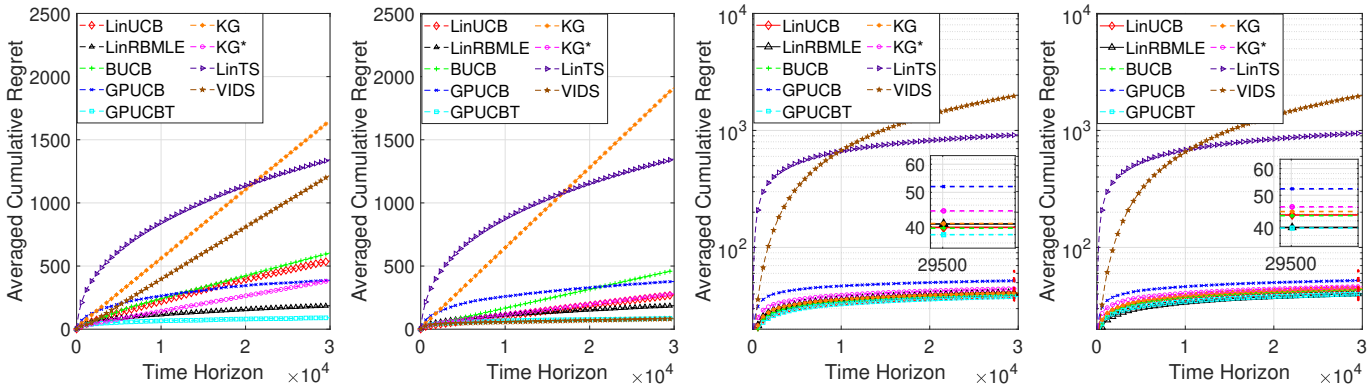
The RBMLE method was originally proposed in (Kumar and Becker 1982). It was subsequently examined in the Markovian setting in (Kumar and Lin 1982; Kumar 1983b; Borkar 1990), and in the linear quadratic Gaussian (LQG) system setting in (Kumar 1983a; Campi and Kumar 1998; Prandini and Campi 2000). A survey, circa 1985, of the broad field of stochastic adaptive control can be found in (Kumar 1985). 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 (Liu et al. 2020). 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 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 2002). The idea of UCB was later extended to the LinUCB policy, which is simpler to implement and has been tested extensively via experiments (Li et al. 2010). 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, Hayes, and Kakade 2008) 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 2010) and (Abbasi-Yadkori, Pál, and Szepesvári 2011) 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 (Filippi et al. 2010; Li, Lu, and Zhou 2017) to study a broader class of linear stochastic bandit models. Differing from the above UCB-type approaches, as a principled frequentist method, the 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. 2020), which adapted the RBMLE principle for stochastic multi-armed bandits and presented the regret analysis as well as extensive numerical experiments. However, (Liu et al. 2020) 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) (Thompson 1933) approaches the exploration issue by sampling the posterior distribution. For linear bandit models, TS has been tested in large-scale experiments (Chapelle and Li 2011) and shown to enjoy order-optimal regret bounds in various bandit settings (Agrawal and Goyal 2013; Russo and Van Roy 2016; Abeille, Lazaric et al. 2017; Agrawal and Goyal 2017; Dumitrascu, Feng, and Engelhardt 2018). On the other hand, Bayesian strategies can also be combined with the notion of UCB for exploration, as in the popular GPUCB (Srinivas et al. 2010) and Bayes-UCB (Kaufmann, Cappé, and Garivier 2012) algorithms. However, to the best of our knowledge, there is no regret guarantee for Bayes-UCB in the linear bandit setting (Urteaga and Wiggins 2017). Alternative exploration strategies for linear bandits have also been considered from the perspective of explicit information-theoretic measures. (Russo and Van Roy 2018) 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 (Ryzhov, Powell, and Frazier 2012; Ryzhov, Frazier, and Powell 2010), 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 extend the Reward Biased Maximum Likelihood principle originally proposed for adaptive control, to contextual bandits. LinRBMLE leads to a simple index policy for standard linear bandits. Through both theoretical regret analysis and simulations, we prove that the regret performance of LinRBMLE is competitive with the state-of-the-art methods while being computationally efficient. Given the favorable trade-off of regret and computation time, RBMLE is a promising approach for contextual bandits.

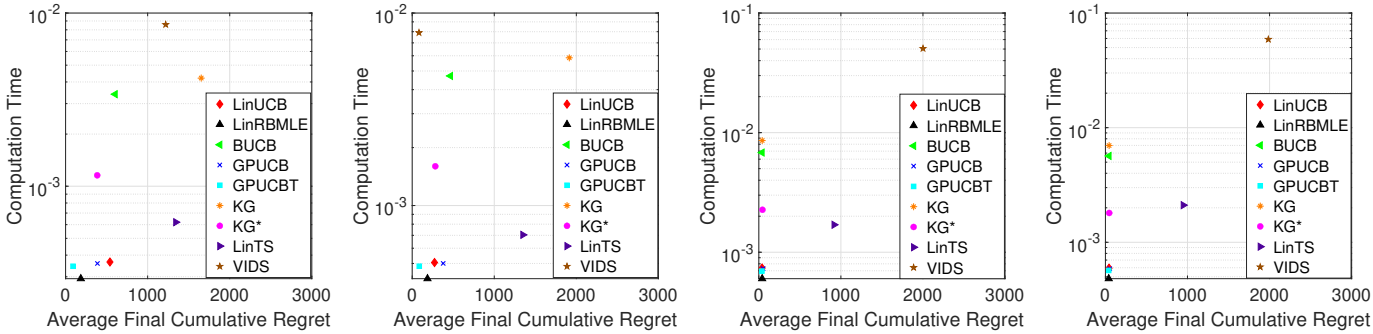


(a) (b) (c) (d)

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 $\theta_* = (-0.7, -0.6, 0.1)$.

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

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.



(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).

Algorithm	RBMLE	LinUCB	BUCB	GPUCB	GPUCBT	KG	KG*	LinTS	VIDS
$d = 100, K = 100$	**0.127**	0.149	1.157	0.147	0.145	1.107	0.401	0.192	5.054
$d = 200, K = 100$	**0.213**	0.24	1.237	0.234	0.233	1.168	0.488	0.561	9.239
$d = 300, K = 100$	**0.303**	0.339	1.467	0.334	0.332	1.386	0.599	1.374	19.876
$d = 100, K = 200$	0.233	0.273	2.25	0.268	0.266	2.155	1.021	**0.205**	6.218
$d = 200, K = 200$	**0.373**	0.421	2.455	0.41	0.409	2.31	1.168	0.586	13.838
$d = 300, K = 200$	**0.452**	0.503	2.636	0.496	0.495	2.455	1.258	1.418	28.652

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 $T = 10^2$ and in 10^{-2} seconds. The best is highlighted.

Acknowledgments

This material is based upon work partially supported by the Ministry of Science and Technology of Taiwan under Contract No. MOST 108-2636-E-009-014 and Contract No. MOST 109-2636-E-009-012. This material is also based upon work partially supported by NSF under Science & Technology Center Grant CCF-0939370, Contract CCF-1934904, and Contract OMA-2037890, the U.S. Army Research Office under Contract No. W911NF-18-10331, and U.S. ONR under Contract No. N00014-18-1-2048. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office, Office of Naval Research, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Ethical 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 (Varatharajah et al. 2018), mobile health (Tewari and Murphy 2017), personalized recommender (Li et al. 2010) and online advertising systems (Chapelle and Li 2011), 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 (Kumar 1985; Kumar and Lin 1982; Kumar 1983b,a; Borkar 1990; Campi and Kumar 1998; Prandini and Campi 2000). 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 (Joseph et al. 2016).

References

Abbasi-Yadkori, Y.; Pál, D.; and Szepesvári, C. 2011. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, 2312–2320.

Abeille, M.; Lazaric, A.; et al. 2017. Linear Thompson sampling revisited. *Electronic Journal of Statistics* 11(2): 5165–5197.

Agrawal, S.; and Goyal, N. 2013. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, 127–135.

Agrawal, S.; and Goyal, N. 2017. Near-optimal regret bounds for Thompson sampling. *Journal of the ACM (JACM)* 64(5): 1–24.

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

Bertsekas, D. P. 1999. *Nonlinear programming*. Athena Scientific.

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

Borkar, V.; and Varaiya, P. 1979. Adaptive control of Markov chains, I: Finite parameter set. *IEEE Transactions on Automatic Control* 24(6): 953–957.

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

Chapelle, O.; and Li, L. 2011. An empirical evaluation of Thompson sampling. In *Advances in neural information processing systems*, 2249–2257.

Chowdhury, S. R.; and Gopalan, A. 2017. On kernelized multi-armed bandits. *arXiv preprint arXiv:1704.00445* .

Chu, W.; Li, L.; Reyzin, L.; and Schapire, R. 2011. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 208–214.

Dani, V.; Hayes, T. P.; and Kakade, S. M. 2008. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*.

Dumitrascu, B.; Feng, K.; and Engelhardt, B. 2018. PG-TS: Improved Thompson sampling for logistic contextual bandits. In *Advances in neural information processing systems*, 4624–4633.

Faury, L.; Abeille, M.; Calauzènes, C.; and Fercoq, O. 2020. Improved Optimistic Algorithms for Logistic Bandits. *arXiv preprint arXiv:2002.07530* .

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

Feldbaum, A. A. 1960b. Dual control theory. II. *Avtomatika i Telemekhanika* 21(11): 1453–1464.

Filippi, S.; Cappe, O.; Garivier, A.; and Szepesvári, C. 2010. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, 586–594.

Joseph, M.; Kearns, M.; Morgenstern, J. H.; and Roth, A. 2016. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, 325–333.

Jun, K.-S.; Bhargava, A.; Nowak, R.; and Willett, R. 2017. Scalable generalized linear bandits: Online computation and hashing. In *Advances in Neural Information Processing Systems*, 99–109.

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

Kaufmann, E.; Cappé, O.; and Garivier, A. 2012. On Bayesian upper confidence bounds for bandit problems. In *Artificial intelligence and statistics*, 592–600.

Kirschner, J.; and Krause, A. 2018. Information Directed Sampling and Bandits with Heteroscedastic Noise. In *Conference On Learning Theory*, 358–384.

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

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

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

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

Kumar, P. R.; and Lin, W. 1982. Optimal adaptive controllers for unknown Markov chains. *IEEE Transactions on Automatic Control* 27(4): 765–774.

Kumar, P. R.; and Varaiya, P. 2015. *Stochastic systems: Estimation, identification, and adaptive control*. SIAM.

Lai, T. L.; and Robbins, H. 1985. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* 6(1): 4–22.

Li, L.; Chu, W.; Langford, J.; and Schapire, R. E. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, 661–670.

Li, L.; Lu, Y.; and Zhou, D. 2017. Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2071–2080. JMLR. org.

Li, S.; Karatzoglou, A.; and Gentile, C. 2016. Collaborative filtering bandits. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 539–548.

Liu, X.; Hsieh, P.-C.; Hung, Y.-H.; Bhattacharya, A.; and Kumar, P. R. 2020. Exploration Through Reward Biasing: Reward-Biased Maximum Likelihood Estimation for Stochastic Multi-Armed Bandits. In *International Conference on Machine Learning*.

Lu, T.; Pál, D.; and Pál, M. 2010. Contextual multi-armed bandits. In *Proceedings of the Thirteenth international conference on Artificial Intelligence and Statistics*, 485–492.

Oh, M.-h.; and Iyengar, G. 2019. Multinomial Logit Contextual Bandits. [Online]. Available from: Reinforcement Learning for Real Life (RL4RealLife) Workshop in the 36th International Conference on Machine Learning https://openreview.net/pdf/0d5b5b3cba7fbcaef861fd82231e5512ac8c1b5e.pdf.

Prandini, M.; and Campi, M. 2000. Adaptive LQG Control of Input-Output Systems—A Cost-biased Approach. *SIAM Journal on Control and Optimization* 39(5): 1499–1519.

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

Russo, D.; and Van Roy, B. 2016. An information-theoretic analysis of Thompson sampling. *The Journal of Machine Learning Research* 17(1): 2442–2471.

Russo, D.; and Van Roy, B. 2018. Learning to optimize via information-directed sampling. *Operations Research* 66(1): 230–252.

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

Ryzhov, I. O.; Powell, W. B.; and Frazier, P. I. 2012. The knowledge gradient algorithm for a general class of online learning problems. *Operations Research* 60(1): 180–195.

Srinivas, N.; Krause, A.; Kakade, S.; and Seeger, M. 2010. 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*, 1015–1022.

Tewari, A.; and Murphy, S. A. 2017. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, 495–517. Springer.

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

Urteaga, I.; and Wiggins, C. H. 2017. Bayesian bandits: balancing the exploration-exploitation tradeoff via double sampling. *arXiv:1709.03162* .

Varatharajah, Y.; Berry, B.; Koyejo, S.; and Iyer, R. 2018. A Contextual-bandit-based Approach for Informed Decision-making in Clinical Trials. *arXiv preprint arXiv:1809.00258* .

Zhang, L.; Yang, T.; Jin, R.; Xiao, Y.; and Zhou, Z.-H. 2016. Online stochastic linear optimization under one-bit feedback. In *International Conference on Machine Learning*, 392–401.