Double Doubly Robust Thompson Sampling for Generalized Linear Contextual Bandits

Wonyoung Kim[1], **Kyungbok Lee**[2], **Myunghee Cho Paik**[2, 3] *

[1] Department of Industrial Engineering and Operations Research, Columbia University
[2] Department of Statistics, Seoul National University
[3] Shepherd23 Inc.
wk2389@columbia.edu, turtle107@snu.ac.kr, myungheechopaik@snu.ac.kr

Abstract

We propose a novel algorithm for generalized linear contextual bandits (GLBs) with an $\tilde{O}(\sqrt{\kappa^{-1}\phi^{-1}T})$ regret over T rounds where ϕ is the minimum eigenvalue of the covariance of contexts and κ is a lower bound of the variance of rewards. In several identified cases of $\phi^{-1} = O(d)$, where d is the dimension of contexts, our result is the first regret bound for generalized linear bandits (GLBs) achieving the order $\sqrt{d}$ without discarding the observed rewards. Previous approaches achieve the regret bound of order $\sqrt{d}$ by discarding the observed rewards, whereas our algorithm achieves the bound incorporating contexts from all arms in our double doubly robust (DDR) estimator. The DDR estimator is a subclass of doubly robust estimator but with a tighter error bound. We also provide an $O(\kappa^{-1}\phi^{-1}\log(NT)\log T)$ regret bound for N arms under a probabilistic margin condition. This is the first regret bound under the margin condition for linear models or GLMs when contexts are different for all arms but coefficients are common. We conduct empirical studies using synthetic data and real examples, demonstrating the effectiveness of our algorithm.

Introduction

In multi-armed bandits (MABs), a learner repeatedly chooses an action or arm from action sets given in an environment and observes the reward for the chosen arm. The goal is to find a rule for choosing arms to maximize the expected cumulative rewards. A linear contextual bandit is an MAB with context vectors for each arm in which the expected reward is a linear function of the corresponding context vector. Popular contextual bandit algorithms include upper confidence bound (Abbasi-Yadkori, Pál, and Szepesvári (2011), `LinUCB`) and Thompson sampling (Agrawal and Goyal (2013), `LinTS`) whose theoretical properties have been studied (Auer 2002; Chu et al. 2011; Abbasi-Yadkori, Pál, and Szepesvári 2011; Agrawal and Goyal 2014). More recently, extensions to generalized linear models (GLMs) have received significant attention. The study by Filippi et al. (2010) is one of the pioneering studies to propose contextual bandits for GLMs with regret analysis. Abeille and Lazaric (2017) extended `LinTS` to GLM rewards. In the GLM, the variance of a reward is related to its mean and the regret bound typically depends on the lower bound of the variance, κ. Faury et al. (2020) demonstrated the regret bound free of κ for a logistic case.

When we focus on the dependence of d on the regret bound, an $\tilde{O}(\kappa^{-1}\sqrt{dT})$ regret bound has been achieved by Li, Lu, and Zhou (2017), where $\tilde{O}$ denotes big-O notation up to logarithmic factors. For logistic models, Jun et al. (2021) achieved $\tilde{O}(\sqrt{dT})$. These two methods used the approach of Auer (2002) whose main idea is to carefully compose independent samples to develop an estimator and derive the bound using this independence. However, to maintain independence, the developed estimator ignores many observed rewards. Despite of this limitation, no existing algorithms have achieved a regret bound sublinear in d without using the approach of Auer (2002). We propose a novel contextual bandit algorithm for generalized linear rewards with $\tilde{O}(\sqrt{\kappa^{-1}dT})$ in several practical cases. To our knowledge, this regret bound is firstly achieved without relying on the approach of Auer (2002).

Our proposed algorithm is the first among `LinTS` variants with a regret bound of order $\sqrt{d}$ for linear or generalized linear payoff. The proposed algorithm is equipped with a novel estimator called double doubly robust (DDR) estimator which is a subclass of doubly robust (DR) estimators with a tighter error bound than conventional DR estimators. The DR estimators use contexts from unselected arms after imputing predicted responses using a class of imputation estimators. Based on these estimators, Kim, Kim, and Paik (2021) have recently proposed a `LinTS` variant for linear rewards which achieves an $\tilde{O}(d\sqrt{T})$ regret bound for several practical cases. With our proposed DDR estimator, we develop a novel `LinTS` variant for generalized linear rewards which improves the regret bound by a factor of $\sqrt{d}$.

We also demonstrate a logarithmic cumulative regret bound under a margin condition. Obtaining bounds under this margin condition is challenging because it requires a tight prediction error bound of order $1/\sqrt{t}$ for each round t. Our DDR estimator, which has a tighter error bound than DR estimators, enables us to derive a logarithmic cumulative regret bound under the margin condition.

In our experiments, we demonstrate that our proposed algorithm performs better with less a priori knowledge

*Corresponding author

Algorithm	Regret Upper Bound	Model	Assumptions
GLM-UCB (Filippi et al. 2010)	$O(\kappa^{-1}d\sqrt{T}\log^{3/2}T)$	GLM	1-3
SupCB-GLM (Li, Lu, and Zhou 2017)	$O(\kappa^{-1}\sqrt{dT\log NT}\log T)$	GLM	1-5
TS (GLM) (Abeille and Lazaric 2017)	$O(\kappa^{-1}d^{3/2}\sqrt{T}\log T)$	GLM	1-3
Logistic UCB-2 (Faury et al. 2020)	$O(d\sqrt{T}\log T)$	Logistic	1-3
SupLogistic (Jun et al. 2021)	$O(\sqrt{dT\log NT}\log T)$	Logistic	1-5
DDRTS-GLM (Proposed)	$O(\sqrt{\kappa^{-1}dT\log NT})$	GLM	1-5, $\phi^{-1}=O(d)$

Table 1: The main orders of regret bounds for GLMs/logistic bandit algorithms (See Appendix B in Kim, Lee, and Paik (2022) for the comparison with more algorithms). For details about the assumptions, see Section .

than several generalized linear contextual bandit algorithms. Many contextual bandit algorithms require the knowledge of time horizon T (Li, Lu, and Zhou 2017; Jun et al. 2021) or a set which includes the true parameter (Filippi et al. 2010; Jun et al. 2017; Faury et al. 2020). This hinders the application of the algorithms to the real examples. Our proposed algorithm does not require the knowledge of T or the parameter set and widely applicable with superior performances.

The main contributions of this paper are as follows:

- We propose a novel generalized linear contextual bandit (GLB) algorithm that has $\tilde{O}(\sqrt{\kappa^{-1}dT})$ in several identified cases (Theorem 1). This is the first regret bound for GLBs with the order $\sqrt{d}$ without relying on the arguments of Auer (2002).
- We provide a novel estimator called DDR estimator which is a subclass of the DR estimators but has a tighter error bound than conventional DR estimators. The proposed DDR estimator uses an explicit form of the imputation estimator, which is also doubly robust and guarantees our novel theoretical results.
- We provide novel theoretical analyses that extend DR Thompson sampling (Kim, Kim, and Paik (2021), DR-LinTS) to GLBs and improve its regret bound by a factor of $\sqrt{d}$. Our analyses are different from those of DR-LinTS in deriving a new regret bound capitalizing on independence (Lemma 6) and new maximal elliptical potential lemma (Lemma 7).
- We provide an $O(\kappa^{-1}d\log NT\log T)$ regret bound under a probabilistic margin condition (Theorem 8). This is the first logarithmic regret bound for linear and generalized linear payoff with arm-specific contexts.
- Simulations using synthetic datasets and analyses with two real datasets show that the proposed method outperforms existing GLMs/logistic bandit methods.

Generalized Linear Contextual Bandits

The GLB problem considers the case in which the reward follows GLMs. Filippi et al. (2010) presented a GLB problem with a finite number of arms. In this setting, the learner faces N arms and each arm is associated with contexts. For each $i \in \{1,\dots,N\} := [N]$, we denote a d-dimensional context for the i-th arm at round t by $X_{i,t} \in \mathbb{R}^d$. At round t, the learner observes the contexts $\{X_{1,t},\dots,X_{N,t}\} := \mathcal{X}_t$, pulls one arm $a_t \in [N]$, and observes Y_t. Let $\mathcal{H}_t$ be the history at round t that contains the contexts $\{\mathcal{X}_\tau\}_{\tau=1}^t$, chosen arms $\{a_\tau\}_{\tau=1}^{t-1}$ and the corresponding rewards $\{Y_\tau\}_{\tau=1}^{t-1}$. We assume that the distribution of the reward is

$$\mathbb{P}\left(Y_t = y \middle| \mathcal{H}_t, a_t\right) \propto \exp\left\{y\theta_{a_t,t} - b\left(\theta_{a_t,t}\right)\right\},$$

where $\theta_{a_t,t} = \xi_{a_t,t} := X_{a_t,t}^T\beta^*$, with the function b known and assumed to be twice differentiable. Then we have $\mathbb{E}\left[Y_t \middle| \mathcal{H}_t, a_t\right] = b'\left(\theta_{a_t,t}\right) = \mu(\xi_{a_t,t})$ and $\mathbb{V}\left[Y_t \middle| \mathcal{H}_t, a_t\right] = b''\left(\theta_{a_t,t}\right)$, for some unknown $\beta^* \in \mathbb{R}^d$. Let $a_t^* := \arg\max_{i=1,\dots,N}\left\{\mu(X_{i,t}^T\beta^*)\right\}$ be the optimal arm that maximizes the expected reward at round t. We define the regret at round t by $\texttt{regret}(t) := \mu(X_{a_t^*,t}^T\beta^*) - \mu(X_{a_t,t}^T\beta^*)$. Our goal is to minimize the sum of regrets over T rounds, $R(T) := \sum_{t=1}^T \texttt{regret}(t)$. The total number of rounds T is finite but possibly unknown.

Related Works

Table 1 summarizes the comparison of main regret orders for GLMs/logistic bandit algorithms. Filippi et al. (2010) extended LinUCB and proposed an algorithm for GLB (GLM-UCB) with a regret bound of $O(\kappa^{-1}d\sqrt{T}\log^{3/2}T)$, where κ is a lower bound of μ'. Abeille and Lazaric (2017) extended LinTS to GLBs and demonstrated a regret bound of $O(\kappa^{-1}d^{3/2}\sqrt{T}\log T)$. In contrast to linear models, the Gram matrix in GLMs depends on the mean and thus the regret bound has the factor of κ^{-1} which can be large. Faury et al. (2020) solved this problem by proposing an UCB-based algorithm for logistic models with a regret bound of $O(d\sqrt{T}\log T)$ free of κ.

Other related works have focused on achieving $\tilde{O}(\sqrt{dT})$ regret bounds using the argument of Auer (2002). The SupCB-GLM algorithm (Li, Lu, and Zhou 2017) is extended from GLM-UCB (Filippi et al. 2010) yielding an $O(\kappa^{-1}\sqrt{dT\log TN}\log T)$ regret bound, whereas SupLogistic (Jun et al. 2021) is extended from Logistic UCB-2 (Faury et al. 2020) with an $O(\sqrt{dT\log TN}\log T)$ regret bound. SupCB-GLM and SupLogistic have the best-known regret bounds for contextual bandits with GLM and logistic models, respectively. However, they do not incorporate many observed rewards to achieve independence, resulting in additional rounds for their estimators to achieve certain precision level.

The DR method has been employed in the bandit literature for linear payoffs (Kim and Paik 2019; Dimakopoulou et al. 2019; Kim, Kim, and Paik 2021). Except in the work by Dimakopoulou et al. (2019), the merit of this technique is

Algorithm 1: Double Doubly Robust Thompson Sampling for Generalized Linear Contextual Bandits (DDRTS-GLM)

INPUT: exploration parameter v, regularization parameter λ, the number of maximum possible resampling $M_t > 0$, threshold value for resampling $\gamma \in [(N+1)^{-1}, N^{-1})$, $S > 0$.
Initialize $V_1 := \lambda I_d$, $\widehat{\beta}_t = 0_d$ and sample a_1 from $\{1, \ldots, N\}$, randomly
for $t \geq 2$ **do**
  Initialize $n = 1$ and observe contexts $\mathcal{X}_t$
  Sample $\tilde{\beta}_{1,t}, \ldots \tilde{\beta}_{N,t}$ from $\mathcal{N}(\widehat{\beta}_{t-1}, v^2 V_{t-1}^{-1})$, independently and observe $m_t = \arg\max_i \mu(X_{i,t}\tilde{\beta}_{i,t})$
  if $\tilde{\pi}_{m_t,t} \leq \gamma$ and $n \leq M_t$ **then**
    Set $n \leftarrow n+1$ and go to line 5
  else
    Set $a_t = m_t$, play arm a_t and observe reward Y_{m_t}
  end if
  Compute $V_t = \sum_{\tau=1}^{t} \sum_{i=1}^{N} \mu'(X_{i,\tau}^T \widehat{\beta}_{t-1}) X_{i,\tau} X_{i,\tau}^T + \lambda I_d$, $\breve{\beta}_t$ and $Y_{i,\tau}^{\breve{\beta}_t}$, then solve (3) to update $\widehat{\beta}_t$
end for

the use of full contexts along with pseudo-rewards. Our approach shares the merit of using full contexts but also uses a new estimator with a tighter error bound and more elaborate pseudo-rewards than conventional DR methods. This calls for new regret analyses which are summarized in Lemmas 6 and 7.

Under a *probabilistic* margin condition, Bastani and Bayati (2020) and Bastani, Bayati, and Khosravi (2021) presented $O(\log T)$ regret bounds for (generalized) linear contextual bandits when contexts are common to all arms but coefficients are arm-specific. In a setting in which contexts are arm-specific but with common coefficients, previous works have shown regret bounds under *deterministic* margin conditions (Dani, Hayes, and Kakade 2008; Abbasi-Yadkori, Pál, and Szepesvári 2011). In the latter setting, our bound is the first regret bound under a *probabilistic* margin condition for linear models or GLMs.

Proposed Methods

Subclass of Doubly Robust Estimator: Double Doubly Robust (DDR) Estimator

For linear contextual bandits, Kim, Kim, and Paik (2021) employed a DR estimator whose merit is to use all contexts, selected or unselected. To use unselected contexts in the DR estimator, the authors replaced the unobserved reward for the unselected context with a pseudo-reward defined as follows:

$$Y_{i,t}^{\breve{\beta}_t} := \left\{1 - \frac{\mathbb{I}(a_t = i)}{\pi_{i,t}}\right\} \mu(X_{i,t}^T \breve{\beta}_t) + \frac{\mathbb{I}(a_t = i)}{\pi_{i,t}} Y_t, \quad (1)$$

where $\pi_{i,t} = \mathbb{P}(a_t = i | \mathcal{H}_t) > 0$ [1]is the selection probability and $\breve{\beta}_t$ is an imputation estimator for β^* at round t. Kim, Kim, and Paik (2021) studied the case of $\mu(x) = x$ and proposed to set $\breve{\beta}_t$ as any $\mathcal{H}_t$-measurable estimator that renders $Y_{i,t}^{\breve{\beta}_t}$ unbiased for the conditional mean because of $\mathbb{E}(\mathbb{I}(a_t = i)|\mathcal{H}_t) = \pi_{i,t}$. This choice of imputation estimators covers a wide class of DR estimators. In this paper, we propose a subclass of the DR estimators called DDR estimator for GLB problem. The proposed subclass estimator has an explicit form of $\breve{\beta}_t$ which is crucial to our novel theoretical results.

To introduce our imputation estimator $\breve{\beta}_t$, we begin by developing a new bounded estimator,

$$\widehat{\beta}_t^S := \begin{cases} \widehat{\beta}_t^A & \text{if } \left\|\widehat{\beta}_t^A\right\|_2 \leq S \\ S\frac{\widehat{\beta}_t^A}{\left\|\widehat{\beta}_t^A\right\|_2} & \text{otherwise} \end{cases}$$

for some $S > 0$, where $\widehat{\beta}_t^A$ is the solution to the score equation, $\sum_{\tau=1}^{t} \left\{Y_\tau - \mu\left(X_{a_\tau,\tau}^T \beta\right)\right\} X_{a_\tau,\tau} = 0$. Now the imputation estimator $\breve{\beta}_t$ at round t is the solution to

$$\sum_{\tau=1}^{t} \sum_{i=1}^{N} \left\{Y_{i,\tau}^{\widehat{\beta}_t^S} - \mu\left(X_{i,\tau}^T \beta\right)\right\} X_{i,\tau} - \lambda\beta = 0,$$

where $\lambda > 0$ is a regularization parameter. Regardless of the value of S, the pseudo-reward $Y_{i,\tau}^{\widehat{\beta}_t^S}$ is unbiased. Different from DR estimators, the imputation estimate $\breve{\beta}_t$ is also robust and satisfies

$$\left\|\breve{\beta}_t - \beta^*\right\|_2 \leq \sqrt{\frac{\kappa}{Nd}}, \quad (2)$$

when $t \geq \mathcal{T}_* = \Omega(\kappa^{-3}\phi^{-2}Nd^2 \log T)$. The proof of (2) and detailed expression of $\mathcal{T}_*$ is in Appendix E.1 of Kim, Lee, and Paik (2022). With this newly defined imputation estimator $\breve{\beta}_t$ and the corresponding pseudo-reward (1), the proposed DDR estimator $\widehat{\beta}_t$ is defined as the solution to,

$$U_t(\beta) := \sum_{\tau=1}^{t} \sum_{i=1}^{N} \left\{Y_{i,\tau}^{\breve{\beta}_t} - \mu(X_{i,\tau}^T \beta)\right\} X_{i,\tau} - \lambda\beta = 0. \quad (3)$$

This DDR estimator uses not only a robust imputation estimator, but also a more elaborate pseudo-reward than that in DR estimators. Specifically, for all $\tau \in [t]$, DR estimators use $Y_{i,\tau}^{\breve{\beta}_\tau}$, whereas our DDR estimator computes $Y_{i,\tau}^{\breve{\beta}_t}$, updating pseudo-rewards on the basis of the most up-to-date imputation estimator. Both the imputation estimator with a tighter estimation error bound and elaborate pseudo-rewards result in a subsequent reduction of the prediction error bound for the DDR estimator, which plays a crucial role in reducing $\sqrt{d}$ in the regret bound compared to that of Kim, Kim, and Paik (2021).

Double Doubly Robust Thompson Sampling

Our proposed algorithm, double doubly robust Thompson sampling algorithm for generalized linear bandits (DDRTS-GLM) is presented in Algorithm 1. At each round $t \geq 2$, the algorithm samples $\tilde{\beta}_{i,t}$ from the distribution $\mathcal{N}(\widehat{\beta}_{t-1}, v^2 V_{t-1}^{-1})$ for each $i \in [N]$ independently. Define $\tilde{Y}_{i,t} := \mu(X_{i,t}^T \tilde{\beta}_{i,t})$ and let $m_t := \arg\max_i \tilde{Y}_{i,t}$ be a

[1]The selection probability is nonzero for all arms in Thompson sampling with Gaussian prior.

candidate action. After observing m_t, compute $\tilde{\pi}_{m_t,t} := \mathbb{P}(\tilde{Y}_{m_t,t} = \max_i \tilde{Y}_{i,t}|\mathcal{H}_t)$. If $\tilde{\pi}_{m_t,t} > \gamma$, the arm m_t is selected, i.e., $a_t = m_t$. Otherwise, the algorithm resamples $\tilde{\beta}_{i,t}$ until it finds another arm satisfying $\tilde{\pi}_{i,t} > \gamma$ up to a predetermined fixed value M_t.

DDRTS-GLM requires additional computations because of the resampling and the DDR estimator. The computation for resampling is invoked when $\tilde{\pi}_{m_t,t} \leq \gamma$ but this does not occur often in practice. In computing $\tilde{\pi}_{m_t,t}$, we refer to Section H in Kim, Kim, and Paik (2021) which proposed a Monte-Carlo estimate and showed that the estimate is efficiently computable. Furthermore, the estimate is consistent and does not affect the theoretical results. Most additional computations in DDRTS-GLM occurs in estimating $\widehat{\beta}_t$ which requires the imputation estimator $\breve{\beta}_t$ and contexts of all arms. This additional computation of DDRTS-GLM is a minor cost of achieving a regret bound sublinear to d and superior performance compared to existing GLB algorithms.

Regret Analysis

In this section, we present two regret bounds for DDRTS-GLM: an $O(\sqrt{\kappa^{-1}dT\log NT})$ regret bound (Theorem 1) and an $O(\kappa^{-1}d\log NT\log T)$ regret bound under a margin condition (Theorem 8).

A Regret Bound of DDRTS-GLM

We provide the following assumptions.

Assumption 1 (Boundedness)**.** There exists $S^* > 0$ such that $\|X_{i,t}\|_2 \leq 1$ and $\|\beta^*\|_2 \leq S^*$ for all $i \in [N]$ and $t \in [T]$. The value of S^* is possibly unknown.

Assumption 2 (Bounded rewards)**.** There exists $B > 0$ such that $|Y_t| \leq B$ for all $t \in [T]$ almost surely.

Assumption 3 (Mean function)**.** Define a set of vicinity of all possible β^* by $\mathcal{B}_r^* := \{\beta : \|\beta\|_2 \leq r + S^*\}$. Then there exists $r > 0$ such that the mean function μ is twice continuously differentiable on $\{x^T\beta : \|x\|_2 \leq 1, \beta \in \mathcal{B}_r^*\}$, and $\kappa := \inf_{\|x\|_2\leq 1,\beta\in\mathcal{B}_r^*} \mu'\left(x^T\beta\right) > 0$. This implies that $|\mu\prime| \leq L_1$ and $|\mu''| \leq L_2$ on the bounded set $\{x^T\beta : \|x\|_2 \leq 1, \beta \in \mathcal{B}_r^*\}$ for some $L_1, L_2 \in (0,\infty)$.

Assumption 4 (Independently identically distributed contexts)**.** Define the set of all contexts at round $t \in [T]$ by $\mathcal{X}_t := \{X_{1,t},\ldots,X_{N,t}\}$. Then the stochastic contexts, $\mathcal{X}_1,\ldots,\mathcal{X}_T$ are independently generated from a fixed distribution $\mathcal{P}_X$. At each round t, the contexts $X_{1,t},\ldots,X_{N,t}$ can be correlated with each other.

Assumption 5 (Positive definiteness of the covariance of the contexts)**.** There exists a positive constant $\phi > 0$ such that $\lambda_{\min}\left(\mathbb{E}[N^{-1}\sum_{i=1}^N X_{i,t}X_{i,t}^T]\right) \geq \phi$ for all t.

Assumptions 1-3 are standard in the GLB literature (see e.g. Filippi et al. (2010); Jun et al. (2017); Russac et al. (2021)) except that Assumption 1 does not require the knowledge of S^*. Assumptions 4 and 5 were used by Li, Lu, and Zhou (2017) and Jun et al. (2021) which achieved regret bounds that have only $\sqrt{d}$. Under these assumptions we present the regret bound of DDRTS-GLM in the following theorem.

Theorem 1. *(A regret bound of* DDRTS-GLM*) Suppose Assumptions 1-5 hold. For any $\gamma \in [1/(N+1), 1/N)$ and $\delta \in (0,1)$ set $v = (\kappa/L_1)\{2\log(N/(1-\gamma N))\}^{-1/2}$ and $M_t = \log(t^2/\delta)/\log(1/(1-\gamma))$ in Algorithm 1. Then with probability at least $1-8\delta$, the regret bound of* DDRTS-GLM *is bounded by*

$$R(T) \leq L_1 S^* \mathcal{T}_* + O\left(\phi^{-3}\kappa^{-3}\log^2 T\right) + (16L_1 + 32L_1^2)\sqrt{\frac{6T}{\kappa\phi}\log\frac{2NT}{\delta}}. \quad (4)$$

The term $\mathcal{T}_*$ represents the number of rounds required for the imputation estimator to satisfy (2). Since the order of $\mathcal{T}_*$ is $O(\log T)$ with respect to T, the first term in (4) is not the main order term. For ϕ in the second and third terms, Lemma 2 identifies the cases of $\phi^{-1} = O(d)$.

Lemma 2. *For $i \in [N]$, let p_i be the density for the marginal distribution of $X_i \in \mathbb{R}^d$. Suppose that $0 < p_{\min} < p_i(x)$ for all $i \in [N]$ and x such that $\|x\|_2 \leq 1$. Then we have*

$$\lambda_{min}\left(\mathbb{E}\left[N^{-1}\sum_{i=1}^N X_i X_i^T\right]\right) \geq \frac{p_{\min} vol\left(\mathcal{B}_d\right)}{(d+2)},$$

where $\mathcal{B}_d$ represents the l_2-unit ball in $\mathbb{R}^d$.

Remark 3. When p_i is the uniform density then $\phi^{-1} = d+2$. For the truncated multivariate normal distribution with mean 0_d and covariance Σ, $\phi^{-1} = (d+2)\exp\left(\frac{\lambda_{\min}(\Sigma)^{-1} - \lambda_{\max}(\Sigma)^{-1}}{2}\right)$.

When there is a lower bound for the marginal density of X_i, the main order of the regret bound is $O(\sqrt{\kappa^{-1}dT\log NT})$. The best known regret bound for GLBs is the $O(\kappa^{-1}\sqrt{dT\log N}\log T)$ regret bound of SupCB-GLM, and our bound is improved by $\kappa^{-1/2}\log T$. To our knowledge, our bound is the best among previously proven bounds for GLB algorithms. Furthermore, this is the first regret bound sublinear in d among LinTS variants. For logistic bandits, our regret bound is comparable with the bound of SupLogistic in terms of d. Even though our bound has extra $\kappa^{-1/2}$, SupLogistic has extra $\log T$. If $\log T > \kappa^{-1/2}$, the proposed method has a tighter bound than SupLogistic.

In the case of linear payoffs, our regret bound is $O(\sqrt{dT\log NT})$, which is tighter than that of previously known linear contextual bandit algorithms. Although the lower bound $\Omega(\sqrt{dT\log N\log T})$ obtained by Li, Wang, and Zhou (2019) is larger than our upper bound, this is not a contradiction because the lower bound does not apply to our setting because of Assumptions 4 and 5.

Key Derivations for the Regret Bound

In this subsection, we show how the improvement of the regret bound is possible. For each i and t, let $\Delta_i(t) := \mu(X_{a_t^*,t}^T\beta^*) - \mu(X_{i,t}^T\beta^*)$, and $\mathcal{D}_i(t) := |\mu(X_{i,t}^T\widehat{\beta}_{t-1}) - \mu(X_{i,t}^T\beta^*)|$. Denote the weighted Gram matrix by $W_t := \sum_{\tau=1}^t\sum_{i=1}^N \mu'(X_{i,\tau}^T\beta^*)X_{i,\tau}X_{i,\tau}^T + \lambda I$. We define a set of

super-unsaturated arms at round t as

$$S_t := \Big\{ i \in [N] : \Delta_i(t) \le \mathcal{D}_i(t) + \mathcal{D}_{a_t^*}(t) + \sqrt{\kappa L_1} \sqrt{\|X_{i,t}\|^2_{W_{t-1}^{-1}} + \left\|X_{a_t^*,t}\right\|^2_{W_{t-1}^{-1}}} \Big\}. \tag{5}$$

The following lemma shows that a_t is in S_t with well-controlled $\pi_{a_t,t}$ with high probability.

Lemma 4. *Suppose Assumptions 1 and 3 hold. Let S_t be the super-unsaturated arms defined in* (5)*. For any $\gamma \in [1/(N+1), 1/N)$ and $\delta \in (0,1)$ set $v = (\kappa/L_1)\{2\log(N/(1-\gamma N))\}^{-1/2}$ and $M_t = \log(t^2/\delta)/\log(1/(1-\gamma))$ in Algorithm 1. Then the action a_t selected by* `DDRTS-GLM` *satisfies*

$$\mathbb{P}\left(\bigcap_{t=\mathcal{T}_*}^{T} \{a_t \in S_t\} \cap \bigcap_{t=1}^{T} \{\pi_{a_t,t} > \gamma\} \right) \ge 1 - \delta. \tag{6}$$

Remark 5. The second event in (6) helps bound $\mathbb{I}(a_t = i)/\pi_{i,t}$ in the pseudo-reward (1).

If a_t is in S_t, the instantaneous regret is bounded by

$$\texttt{regret}(t) = \Delta_{a_t}(t) \le 2 \max_{i \in [N]} \left\{ \mathcal{D}_i(t) + \sqrt{\kappa L_1} \|X_{i,t}\|_{W_{t-1}^{-1}} \right\}. \tag{7}$$

Kim, Kim, and Paik (2021) adopted a similar approach but had a different super-unsaturated set, resulting in a different bound for $\texttt{regret}(t)$. For comparison, in the case of $\mu(x) = x$, Kim, Kim, and Paik (2021) proposed DR estimator $\widehat{\beta}_t^{DR}$ to derive

$$\texttt{regret}(t) \le 2 \left\| \widehat{\beta}_{t-1}^{DR} - \beta^* \right\|_2 + \sqrt{\|X_{a_t,t}\|^2_{W_{t-1}^{-1}} + \left\|X_{a_t^*,t}\right\|^2_{W_{t-1}^{-1}}},$$

and bounded the l_2 estimation error by $O(\phi^{-1} t^{-1/2})$ resulting in an $O(d\sqrt{T})$ regret bound in cases when $\phi^{-1} = O(d)$. This bound has an additional $\sqrt{d}$ compared to (7) because $\|\widehat{\beta}_{t-1}^{DR} - \beta^*\|_2$ is greater than $\max_{i \in [N]} \mathcal{D}_i(t)$ because of Cauchy-Schwartz inequality and using a less accurate imputation estimators. To obtain faster rates on d, we propose the DDR estimator and directly bound the $\texttt{regret}(t)$ with $\mathcal{D}_i(t)$ as in (7) without using Cauchy-Schwartz inequality.

Lemma 6. *(Prediction error bound for DDR estimator) Suppose Assumptions 1-5 hold and the event in* (6) *holds. Then for each $t > \mathcal{T}_*$, with probability at least $1 - 8\delta/T$*

$$\mathcal{D}_i(t) \le \mu'\left(X_{i,t}^T \beta^*\right)(2 + 4L_1) \sqrt{3N \log \frac{2NT}{\delta}} \|X_{i,t}\|_{W_{t-1}^{-1}} + \frac{D_{\mu,B,\lambda,S}}{\phi^3 \kappa^3 (t-1)} \log \frac{4T}{\delta}. \tag{8}$$

for all $i \in [N]$, where $D_{\mu,B,\lambda,S}$ is a constant defined in Section C.2 of Kim, Lee, and Paik (2022).

Proof. For each $t \in (\mathcal{T}_*, T]$, and $i \in [N]$,

$$\mathcal{D}_i(t) \le \mu'\left(X_{i,t}^T \beta^*\right) \left| \sum_{\tau=1}^{t-1} \sum_{j=1}^{N} \eta_{j,\tau}^{\breve{\beta}_{t-1}} X_{i,t}^T W_{t-1}^{-1} X_{j,\tau} \right| + \frac{O\left(\phi^{-3}\kappa^{-3} \log T\right)}{t-1},$$

where $\eta_{j,\tau}^{\beta} := Y_{j,\tau}^{\beta} - \mu(X_{j,\tau}^T \beta^*)$ is the residual for pseudo-rewards. Now for each $\tau \in [t]$, define the filtration as $\mathcal{F}_\tau = \mathcal{H}_\tau \cup \{\mathcal{X}_1, \ldots, \mathcal{X}_t\}$ and $\mathcal{F}_0 := \{\mathcal{X}_1, \ldots, \mathcal{X}_t\}$. Then the random variable $\eta_{j,\tau}^{\beta} X_{j,t}^T W_{t-1}^{-1} X_{j,\tau}$ is $\mathcal{F}_{\tau+1}$-measurable and

$$\begin{aligned} \mathbb{E}\left[\eta_{j,\tau}^{\beta} X_{i,t}^T W_{t-1}^{-1} X_{j,\tau} \middle| \mathcal{F}_\tau \right] &= \mathbb{E}\left[\eta_{j,\tau}^{\beta} \middle| \mathcal{F}_\tau \right] X_{i,t}^T W_{t-1}^{-1} X_{j,\tau} \\ &= \mathbb{E}\left[\eta_{j,\tau}^{\beta} \middle| \mathcal{H}_\tau \right] X_{i,t}^T W_{t-1}^{-1} X_{j,\tau} \\ &= 0. \end{aligned} \tag{9}$$

The first equality holds since $X_{i,t}^T W_{t-1} X_{j,\tau}$ depends only on $\mathcal{F}_0$. The second equality holds due to the independence between $\eta_{i,\tau}^{\beta}$ and $\{\mathcal{X}_u\}_{u=\tau+1}^{t}$ induced by Assumption 4. The remainder of the proof follows from the Azuma-Hoeffding inequality. For details, see Section C.2 in Kim, Lee, and Paik (2022). □

We highlight that the key part of the proof is in (9) which holds because the Gram matrix W_{t-1} contains all contexts. Let $A_t := \sum_{\tau=1}^{t} \mu'(X_{a_\tau,\tau}^T \beta^*) X_{a_\tau,\tau} X_{a_\tau,\tau}^T + I_d$ be the Gram matrix consists of the selected contexts only and $\eta_\tau := Y_\tau - \mu(X_{a_\tau,\tau}^T \beta^*)$ be the error. In general,

$$\mathbb{E}\left[\eta_\tau X_{i,t}^T A_{t-1}^{-1} X_{a_\tau,\tau} \middle| \mathcal{H}_\tau \right] \neq \mathbb{E}\left[\eta_\tau \middle| \mathcal{H}_\tau \right] X_{i,t}^T A_{t-1}^{-1} X_{a_\tau,\tau},$$

due to the dependency between A_{t-1} and η_τ through the actions $\{a_\tau, a_{\tau+1}, \ldots, a_{t-1}\}$. To avoid this dependency, Li, Lu, and Zhou (2017) and Jun et al. (2021) used the approach of Auer (2002) by devising a_τ to be independent of η_τ. Based on this independence, they invoked the equality analogous to the first equality in (9), achieving an $\tilde{O}(\sqrt{dT})$ regret bound. However, the crafted independence negatively affects inefficiency by ignoring dependent samples during estimation. In contrast, we achieve independence using all contexts, generating the Gram matrix free of $a_1, \ldots, a_T$.

From (7) and (8), the $\texttt{regret}(t)$ is bounded by,

$$\begin{aligned} &\texttt{regret}(t) \\ &\le 2 \left\{ (2 + 4L_1) w_{i,t} \sqrt{3N \log \frac{2NT}{\delta}} + \sqrt{\kappa L_1} \right\} \max_{i \in [N]} s_{i,t} \\ &\quad + \frac{D_{\mu,B,\lambda,S}}{\phi^3 \kappa^3 (t-1)} \log \frac{4T}{\delta} \\ &\le 4 (2 + 4L_1) \sqrt{3L_1 N \log \frac{2NT}{\delta}} \max_{i \in [N]} \sqrt{w_{i,t}} s_{i,t} \\ &\quad + \frac{D_{\mu,B,\lambda,S}}{\phi^3 \kappa^3 (t-1)} \log \frac{4T}{\delta}, \end{aligned} \tag{10}$$

where $w_{i,t} := \mu'(X_{i,t}^T \beta^*)$, and $s_{i,t} := \|X_{i,t}\|_{W_{t-1}^{-1}}$. Now we need a bound for the weighted sum of $s_{i,t}$ over $t \in (\mathcal{T}_*, T]$. In this point, many existing regret analyses have used a version of the *elliptical potential lemma*, i.e., Lemma 11 in Abbasi-Yadkori, Pál, and Szepesvári (2011), yielding an $O(\sqrt{dT \log T})$ bound for $\sum_{t=1}^{T} \|X_{a_t,t}\|_{A_{t-1}^{-1}}$. This bound cannot be applied to our case because we have different Gram matrix composed of the contexts of all arms. Therefore we develop the following lemma to prove a bound for the weighted sum.

Lemma 7. *(Maximal elliptical potential lemma) Suppose Assumptions 1-5 hold. Set $w_{i,t} := \mu'(X_{i,t}^T \beta^*)$, and $s_{i,t} := \|X_{i,t}\|_{W_{t-1}^{-1}}$. Then with probability at least $1 - \delta$,*

$$\sum_{t=\mathcal{T}_*}^{T} \max_{i \in [N]} \sqrt{w_{i,t}} s_{i,t} \leq \sqrt{\frac{8 L_1 T}{\phi N \kappa}}. \tag{11}$$

Now the regret bound of `DDRTS-GLM` is proved by applying (11) to (10).

A Logarithmic Cumulative Regret Bound Under a Margin Condition

In this subsection, we present an $O(\kappa^{-1} d \log NT \log T)$ regret bound for `DDRTS-GLM` under the margin condition stated as follows.

Assumption 6 (Margin condition)**.** For all t, there exist *unknown* $\rho_0 > 0$ and $h \geq 0$ such that

$$\mathbb{P}\left(\mu(X_{a_t^*,t}^T \beta^*) \leq \max_{j \neq a_t^*} \mu(X_{j,t}^T \beta^*) + \rho\right) \leq h\rho, \tag{12}$$

for all $\rho \in (0, \rho_0]$.

This margin condition guarantees a probabilistic positive gap between the expected rewards of the optimal arm and the other arms. In the margin condition, $h > 0$ represents how heavy the tail probability is for the margin. Bastani and Bayati (2020) and Bastani, Bayati, and Khosravi (2021) adopted this assumption and proved $O(\log T)$ regret bounds with contextual bandit problems when contexts are the same for all arms and coefficients are arm-specific. When coefficients are the same for all arms and contexts are arm-specific, Dani, Hayes, and Kakade (2008); Abbasi-Yadkori, Pál, and Szepesvári (2011) and Russac et al. (2021) used a *deterministic* margin condition, which is a special case of Assumption 6 when $h = 0$. Now we show that `DDRTS-GLM` has a logarithmic cumulative regret bound under the margin condition.

Theorem 8. *Suppose Assumptions 1-6 hold. Then with probability at least $1 - 10\delta$, the cumulative regret of `DDRTS-GLM` is bounded by*

$$\begin{aligned} R(T) \leq & 2 L_1 S^* \mathcal{T}_0 + O\left(\phi^{-2} \kappa^{-2} \log NT\right) \\ & + \frac{192 h L_1^2 (2 + 4L_1)^2}{\kappa \phi} \log T \log \frac{2NT}{\delta} \end{aligned} \tag{13}$$

for $\mathcal{T}_0 = \rho_0^{-2} \Omega\left(\phi^{-4} \kappa^{-4} \log NT\right)$.

Since $\mathcal{T}_0$ has only $\log NT$, the first term is not the main order term. In the practical cases of $\phi^{-1} = O(d)$ (see Lemma 2), the order of the regret bound is $O(d\kappa^{-1} \log NT \log T)$. To our knowledge, our work is the first to derive a logarithmic cumulative regret bound for `LinTS` variants. We defer the challenges and the intuition deriving the regret bound (13) to Appendix D of Kim, Lee, and Paik (2022).

Experiment Results

In this section, we compare the performance of the five algorithms: (i) `GLM-UCB` (Filippi et al. 2010), (ii) `TS (GLM)` (Abeille and Lazaric 2017), (iii) `SupCB-GLM` (Li, Lu, and Zhou 2017), (iv) `SupLogistic` (Jun et al. 2021), and (v) the proposed `DDRTS-GLM` using simulation data (Section) and the two real datasets (Section and).

Simulation Data

To generate data, we set the number of arms as $N = 10$ and 20 and the dimension of contexts as $d = 20$ and 30. For $j \in [d]$, the j-th elements of the contexts, $[X_{1,t}^{(j)}, \ldots, X_{N,t}^{(j)}]$ are sampled from the normal distribution $\mathcal{N}(\mu_N, V_N)$ with mean $\mu_{10} = [-5, -4, \ldots, -1, 1, \ldots, 4, 5]^T$, and $\mu_{20} = [-10, -9, \ldots, -1, 1, \ldots, 9, 10]^T$. The covariance matrix $V_N \in \mathbb{R}^{N \times N}$ has $V(i,i) = 1$ for every i and $V(i,k) = 0.5$ for every $i \neq k$. The sampled contexts are truncated to satisfy $\|X_i(t)\|_2 \leq 1$. For rewards, we sample Y_t independently from $\text{Ber}(\mu(X_{a_t,t}^T \beta^*))$, where $\mu(x) := 1/(1 + e^{-x})$. Each element of β^* follows a uniform distribution, $\mathcal{U}(-1, 1)$.

As hyperparameters of the algorithms, `GLM-UCB`, `SupCB-GLM`, and `SupLogistic` have α as an exploration parameter. For `TS (GLM)` and the proposed method, v controls the variance of $\tilde{\beta}_i(t)$. In each algorithm, we choose the best hyperparameter from $\{0.001, 0.01, 0.1, 1\}$. The proposed method requires a positive threshold γ for resampling; however, we do not tune γ but fix the value to be $1/(N+1)$. Figure 1 shows the mean cumulative regret $R(T)$, and the proposed algorithm, as represented by the solid line, outperforms four other candidates in all four scenarios.

Forest Cover Type Dataset

We use the Forest Cover Type dataset from the UCI Machine Learning repository (Blake, Keogh, and Merz 1999), as used by Filippi et al. (2010). The dataset contains 581,021 observations, where the response variable is the label of the dominant species of trees of each region, and covariates include ten continuous cartographic variables characterizing features of the forest. We divide the dataset into 32 clusters using the k-means clustering algorithm, and the resulting clusters of the forest represent arms. We repeat the experiment 10 times with $T = 10,000$. Each centroid of each cluster is set to be a 10-dimensional context vector of the corresponding arm, and by introducing an intercept, we obtain $d = 11$. In this example, context vectors remain unchanged in each round. We dichotomize the reward based on whether the forest's dominant class is Spruce/Fir. The goal is to find arms with Spruce/Fir as the dominant class. We execute a

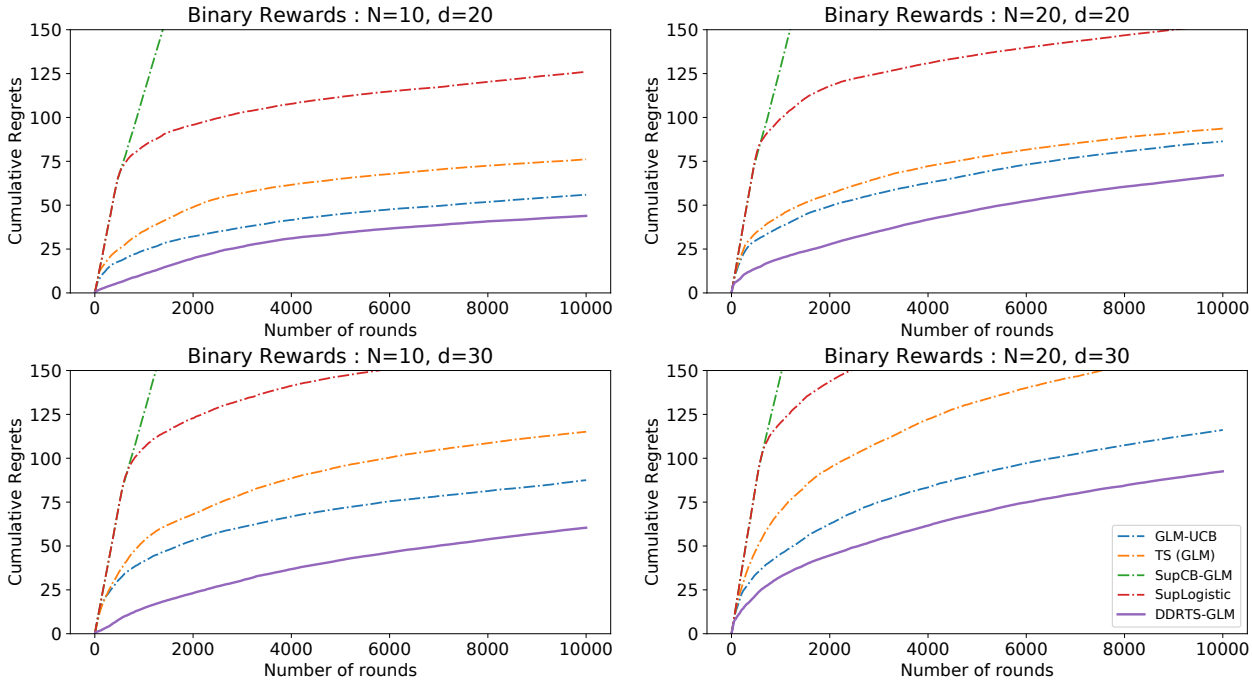

Figure 1: Comparison of the average cumulative regret on synthetic dataset over 5 repeated runs with $T = 10000$.

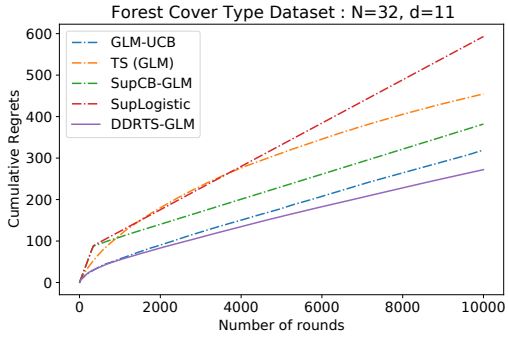

Figure 2: Comparison of the average cumulative regret on forest cover type dataset over 10 repeated runs.

32-armed 11-dimensional contextual bandit with binary rewards. Figure 2 shows that `DDRTS-GLM` outperforms other algorithms.

Yahoo! News Article Recommendation Log Data

The Yahoo! Front Page Today Module User Click Log Dataset (Yahoo! Webscope 2009) contains 45,811,883 user click logs for news articles on Yahoo! Front Page from May 1st, 2009, to May 19th, 2009. Each log consists of a randomly chosen article from $N = 20$ articles and the binary reward Y_t which takes the value 1 if a user clicked the article and 0 otherwise. For each article $i \in [20]$, we have a context vector $X_{i,t} \in \mathbb{R}^{11}$ which comprising 10 extracted features of user-article information and an intercept using a dimension reduction method, as in Chu et al. (2009).

In log data, when the algorithm chooses an action not selected by the original logger, it cannot observe the reward, and the regret is not computable. Instead, we use the click-through rate (CTR): the percentage of the number of clicks. We tune the hyperparameters using the log data from May 1st and run the algorithms on the randomly sampled 10^6 logs in each run. We evaluate the algorithms on the basis of the method by Li et al. (2011), counting only the rounds in which the reward is observed in T. Thus, T is not known a priori and `SupCB-GLM` and `SupLogistic` are not applicable. As a baseline, we run a uniform random policy to observe the CTR lift of each algorithm. Table 2 presents the average/first quartile/third quartile of CTR of each algorithm (the higher is the better) over 10 repetitions, showing that `DDRTS-GLM` achieves the highest CTR.

CTR	1st Q	average	3rd Q
`DDRTS-GLM`	**0.0410**	**0.0449**	**0.0476**
`GLM-UCB`	0.0356	0.0420	0.0468
`TS (GLM)`	0.0393	0.0438	0.0467
uniform random	0.0337	0.0344	0.0351

Table 2: Average/first quartile/third quartile of CTR of news articles over 10 repeated runs for each algorithm.

Acknowledgments

This work is supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT, No.2020R1A2C1A01011950). Wonyoung Kim is also supported by Hyundai Chung Mong-koo scholarship foundation.

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.; and Lazaric, A. 2017. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, 176–184. PMLR.

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. 2014. Thompson Sampling for Contextual Bandits with Linear Payoffs. arXiv:1209.3352.

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

Bastani, H.; and Bayati, M. 2020. Online decision making with high-dimensional covariates. *Operations Research*, 68(1): 276–294.

Bastani, H.; Bayati, M.; and Khosravi, K. 2021. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3): 1329–1349.

Blake, C.; Keogh, E.; and Merz, C. 1999. UCI repository of machine learning databases (Machinereadable data repository). Irvine, CA: Department of Information and Computer Science, University of California at Irvine.

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.

Chu, W.; Park, S.-T.; Beaupre, T.; Motgi, N.; Phadke, A.; Chakraborty, S.; and Zachariah, J. 2009. A case study of behavior-driven conjoint analysis on Yahoo! Front Page Today module. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1097–1104.

Dani, V.; Hayes, T.; and Kakade, S. 2008. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory*, 355–366.

Dimakopoulou, M.; Zhou, Z.; Athey, S.; and Imbens, G. 2019. Balanced linear contextual bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3445–3453.

Faury, L.; Abeille, M.; Calauzènes, C.; and Fercoq, O. 2020. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, 3052–3060. PMLR.

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.

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.

Jun, K.-S.; Jain, L.; Mason, B.; and Nassif, H. 2021. Improved Confidence Bounds for the Linear Logistic Model and Applications to Bandits. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 5148–5157. PMLR.

Kim, G.-S.; and Paik, M. C. 2019. Doubly-Robust Lasso Bandit. In *Advances in Neural Information Processing Systems*, 5869–5879.

Kim, W.; Kim, G.-S.; and Paik, M. C. 2021. Doubly Robust Thompson Sampling with Linear Payoffs. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

Kim, W.; Lee, K.; and Paik, M. C. 2022. Double Doubly Robust Thompson Sampling for Generalized Linear Contextual Bandits. *arXiv preprint arXiv:2209.06983*.

Li, L.; Chu, W.; Langford, J.; and Wang, X. 2011. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 297–306.

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.

Li, Y.; Wang, Y.; and Zhou, Y. 2019. Nearly minimax-optimal regret for linearly parameterized bandits. In *Conference on Learning Theory*, 2173–2174. PMLR.

Russac, Y.; Faury, L.; Cappé, O.; and Garivier, A. 2021. Self-Concordant Analysis of Generalized Linear Bandits with Forgetting. In Banerjee, A.; and Fukumizu, K., eds., *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, 658–666. PMLR.

Yahoo! Webscope. 2009. Yahoo! Front Page Today Module User Click Log Dataset, version 1.0. Accessed 07-April-2021.