Learning in Generalized Linear Contextual Bandits with Stochastic Delays

Zhengyuan Zhou[1,2]*, **Renyuan Xu**[3]* **and Jose Blanchet**[4]
[1] Department of Electrical Engineering, Stanford University
[2] Bytedance Inc.
[3] Department of Industrial Engineering and Operations Research, UC Berkeley
[4] Department of Management Science and Engineering, Stanford University

Abstract

In this paper, we consider online learning in generalized linear contextual bandits where rewards are not immediately observed. Instead, rewards are available to the decision maker only after some delay, which is unknown and stochastic, even though a decision must be made at each time step for an incoming set of contexts. We study the performance of upper confidence bound (UCB) based algorithms adapted to this delayed setting. In particular, we design a delay-adaptive algorithm, which we call Delayed UCB, for generalized linear contextual bandits using UCB-style exploration and establish regret bounds under various delay assumptions. In the important special case of linear contextual bandits, we further modify this algorithm and establish a tighter regret bound under the same delay assumptions. Our results contribute to the broad landscape of contextual bandits literature by establishing that UCB algorithms, which are widely deployed in modern recommendation engines, can be made robust to delays.

1 Introduction

The growing availability of user-specific data has welcomed the exciting era of personalized recommendation, a paradigm that uncovers the heterogeneity across individuals and provides tailored service decisions that lead to improved outcomes. Such heterogeneity is ubiquitous across a variety of application domains (including online advertising, medical treatment assignment, product/news recommendation (Li et al. (2010), Bubeck et al. (2012),Chapelle (2014),Bastani and Bayati (2015),Schwartz et al. (2017))) and manifests itself as different individuals responding differently to the recommended items. Rising to this opportunity, contextual bandits have emerged to be the predominant mathematical formalism that provides an elegant and powerful formulation: its three core components, the features (representing individual characteristics), the actions (representing the recommendation), and the rewards (representing the observed feedback), capture the salient aspects of the problem and provide fertile ground for developing algorithms that balance exploring and exploiting users' heterogeneity.

As such, the last decade has witnessed extensive research efforts in developing effective and efficient contextual bandits algorithms. In particular, two types of algorithms–upper confidence bounds (UCB) based algorithms (Li et al. (2010); Filippi et al. (2010); Chu et al. (2011); Jun et al. (2017); Li et al. (2017)) and Thompson sampling (TS) based algorithms (Agrawal and Goyal (2013a,b); Russo and Van Roy (2014, 2016); Abeille et al. (2017))–stand out from this flourishing and fruitful line of work: their theoretical guarantees have been analyzed in many settings, often yielding (near-)optimal regret bounds; their empirical performance have been thoroughly validated, often providing insights into

[*] These two authors contributed equally.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

their practical efficacy (including the consensus understanding that TS-based algorithms often suffer from intensive computation for posterior updates but can leverage a correctly specified prior and have superior empirical performance; UCB-based algorithms can often achieve tight theoretical regret bounds but are often sensitive to hyper-parameter tuning in empirical performance). To a large extent, these two family of algorithms have been widely deployed in many modern recommendation engines.

However, a key assumption therein–both the algorithm design and their analyses–is that the reward is immediately available after an action is taken. Although useful as a first-step abstraction, this is a stringent requirement that is rarely satisfied in practice, particularly in large-scale systems where the time-scale of a single recommendation is significantly smaller than the time-scale of a user's feedback. For instance, in E-commerce, a recommendation is typically made by the engine in milliseconds, whereas a user's response time (i.e. to buy a product or conversion) is typically much larger, ranging from hours to days, sometimes even to weeks. Similarly, in clinical trials, it is infeasible to immediately observe and hence take into account the medical outcome after applying a treatment to a patient–collecting medical feedback can be a time-consuming and often random process; and in general, it is common to have applied trial treatments to a large number of patients, with individual medical outcomes only available much later at different, random points in time. In both the E-commerce (Kannan et al. (2001); Chapelle (2014); Vernade et al. (2017))and the clinical trials cases (Chow and Chang (2011)), a random and often significantly delayed reward is present, thereby requiring adjustments in classical formulations to understand the impact of delays.

1.1 Related Work

The problem of learning on bandits with delays has recently been studied in different settings in the existing literature, where most of the efforts have concentrated on the multi-armed bandits setting, including both the stochastic and the adversarial multi-armed bandits. For stochastic multi-armed bandits with delays, Joulani et al. (2013) show a regret bound $O(\log T + \mathbb{E}[\tau] + \sqrt{\log T \mathbb{E}[\tau]})$ where $\mathbb{E}[\tau]$ is the mean of the **iid** delays. Desautels et al. (2014) consider Gaussian Process bandits with a bounded stochastic delay. Mandel et al. (2015) follow the work of Joulani et al. (2013) and propose a queue-based multi-armed bandit algorithm to handle delays. Pike-Burke et al. (2017) match the same regret bound as in Joulani et al. (2013) when feedback is not only delayed but also anonymous.

For adversarial multi-armed bandits with delays, Neu et al. (2010) establish the regret bound of $\mathbb{E}[R_T] \leq O(\tau_{\text{const}}) \times \mathbb{E}[R'_T(\frac{T}{\tau_{\text{const}}})]$ for Markov decision process, where τ_{const} is the constant delay and R'_T is the regret without delays. Cesa-Bianchi et al. (2019) consider adversarial bandits with fixed constant delays on the network graph, with a minimax regret of the order $\tilde{O}\left(\sqrt{(K+\tau_{\text{const}})T}\right)$, where K is the number of arms. Another related line of work is adversarial learning with full information (all arms' rewards are observed), where its different variants in the delayed setting have been studied by Weinberger and Ordentlich (2002), Mesterharm (2005), Quanrud and Khashabi (2015) and Garrabrant et al. (2016). Very recently, Bistritz et al. (2019) studied adversarial bandits learning under arbitrary delays using Exp3 and established finite-sample delay-adaptive regret bounds.

On the other hand, learning in contextual bandits with delays are much less explored. Joulani et al. (2013) consider learning on adversarial contextual bandits with delays and establish an expected regret bound $\mathbb{E}\left[R_T\right] \leq (1+\mathbb{E}[M_T^*]) \times \mathbb{E}\left[R'_T\left(\frac{T}{1+\mathbb{E}[M_T^*]}\right)\right]$ by using a black-box algorithm, where M_T^* is the running maximum number of delays up to round T. Dudik et al. (2011) consider stochastic contextual bandits with a fixed constant delay. The reward model they consider is general (i.e. not necessarily parametric); however, they require the policy class to be finite. In particular, they obtain the regret bound $O(\sqrt{K \log N}(\tau_{\text{const}} + \sqrt{T}))$, where N is the number of policies and τ_{const} is again the fixed constant delay. On a related front, Grover et al. (2018b) studied the problem of best-arm identification under delayed feedback. There, the objective is to identify the best arm using as few samples as possible, without taking into account the cost incurred along the way (i.e. a different objective from regret). In closing, we also mention that there is also a growing literature in offline contextual bandits learning Swaminathan and Joachims (2015); Kitagawa and Tetenov (2018); Zhou et al. (2018a). In this domain, delay is typically not a concern as all the data has already been collected in a single batch before any learning/decision-making takes place.

1.2 Our Contributions

In this paper, we consider learning on generalized linear (stochastic) contextual bandits with stochastic delays. More specifically, we design a delay-adaptive algorithm for generalized linear contextual bandits using UCB-style exploration, which we call Delayed UCB (DUCB, as given in Algorithm 1). DUCB requires a carefully designed delay-adaptive confidence parameter, which depends on how many rewards are missing up to the current time step. Next, we give regret characterizations of DUCB under independent stochastic, unbounded delays. In particular, as a special case of our results, when the delays are **iid** with mean μ_D, we establish a high-probability regret bound of $\tilde{O}\left(\left(\sqrt{\mu_D d}+\sqrt{\sigma_G d}+d\right)\sqrt{T}\right)$ on DUCB, where σ_G is a parameter characterizing the tail bound of the delays and d is the feature/context dimension. For comparison, the state-of-the-art regret bound of UCB on generalized linear contextual bandits without delays is $\tilde{O}\left(d\sqrt{T}\right)$ (Filippi et al. (2010); Li et al. (2017)). Regret bounds for more general delays are also given. Note that our analysis here does not assume the number of actions to be finite, and hence these regret bounds apply to infinite-action setting as well.

Finally, we consider the important special case of linear contextual bandits with finitely many actions. In this setting, we provide a different UCB-based algorithm that estimates the underlying parameters using a biased estimator (as opposed to the unbiased estimator employed in the generalized linear contextual bandits setting) and provide a more refined analysis that achieves regret bounds which are a factor of $O(\sqrt{d})$ tighter. More specifically in this setting, as a direct comparison, when the delays are again **iid** with mean μ_D, we establish a high-probability regret bound[2] of $\tilde{O}\left((1+\mu_D+\sigma_G)\sqrt{dT}\right)$.

To the best of our knowledge, these regret bounds provide the first theoretical characterizations in (generalized) linear contextual bandits with large delays and contribute to the broad landscape of contextual bandits literature by delineating the impact of delays on performance.

2 Problem Setup

In this section, we describe the formulation for learning in generalized linear contextual bandits (GLCB) in the presence of delays. We start by reviewing the basics of generalized linear contextual bandits, followed by a description of the delay model. Before proceeding, we first fix some notation.

For a vector $x \in \mathbb{R}^d$, we use $\|x\|$ to denote its l_2-norm and x' its transpose. $\mathbb{B}^d := \{x \in \mathbb{R}^d : \|x\| \leq 1\}$ is the unit ball centered at the origin. The weighted l_2-norm associated with a positive-definite matrix A is defined by $\|x\|_A := \sqrt{x'Ax}$. The minimum and maximum singular values of a matrix A are written as $\lambda_{\min}(A)$ and $\|A\|$ respectively. For two symmetric matrices A and B the same dimensions, $A \succeq B$ means that A-B is positive semi-definite. For a real-valued function f, we use $\dot{f}$ and $\ddot{f}$ to denote its first and second derivatives. Finally, $[n] := \{1, 2, \cdots, n\}$.

2.1 Generalized Linear Contextual Bandits

Decision procedure. We consider the generalized linear contextual bandits problem with K arms. At each round t, the agent observes a context consisting of a set of K feature vectors $x_t := \{x_{t,a} \in \mathbb{R}^d | a \in [K]\}$, which is drawn **iid** from an unknown distribution γ with $\|x_{t,a}\| \leq 1$. Each feature vector $x_{t,a}$ is associated with an unknown stochastic reward $y_{t,a} \in [0, 1]$. If the agent selects one action a_t, there is a reward $y_{t,a_t} \in [0, 1]$ associated with the selected arm a_t and the associated x_{t,a_t}. Under the classic setting, the reward is immediately observed after the decision and the information can be utilized to make decision in the next round.

Relationship between reward Y and context X. In terms of the relationship between y_{t,a_t} and x_{t,a_t} $(t \geq 1)$, we follow the standard generalized linear contextual bandits literature (Filippi et al.

[2] In this case, the number of actions being finite is important. In particular, the regret bound has a $O(\log K)$ dependence. Consequently, strictly speaking, if K is not viewed as a constant, we would also need K to not be too large compared to d in order to retain the same regret bound of $\tilde{O}\left((\mu_D+\sigma_G+1)\sqrt{dT}\right)$. A common (and rather weak) assumption is the K is polynomial in d.

(2010); Li et al. (2017)). Define $\mathcal{H}_t^0 = \{(s, x_s, a_s, y_{s,a_s}), s \leq t-1\} \cup \{x_t\}$ as the information at the beginning of round t. The agent maximizes the cumulative expected rewards over T rounds with information $\mathcal{H}_t^0$ at each round t ($t \geq 1$). Suppose the agent takes action a_t at round t. Denote by $X_t = x_{t,a_t}$, $Y_t = y_{t,a_t}$ and we assume the conditional distribution of Y_t given X_t is from the exponential family. Therefore its density is given by

$$\mathbb{P}_{\theta^*}(Y_t|X_t) = \exp\left(\frac{Y_t X_t'\theta^* - m(X_t'\theta^*)}{h(\eta)} + A(Y_t, \eta)\right). \tag{1}$$

Here, θ^* is an unknown number under the frequentist setting; $\eta \in \mathbb{R}^+$ is a given parameter; A, m and h are three normalization functions mapping from $\mathbb{R}$ to $\mathbb{R}$.

For exponential families, m is infinitely differentiable, $\dot{m}(X'\theta^*) = \mathbb{E}[Y|X]$, and $\ddot{m}(X'\theta^*) = \mathbb{V}(Y|X)$. Denote $g(X'\theta^*) = \mathbb{E}[Y|X]$, one can easily verify that $g(x'\theta) = x'\theta$ for linear model, $g(x'\theta) = \frac{1}{1+\exp(-x'\theta)}$ for logistic model and $g(x'\theta) = \exp(x'\theta)$ for Poisson model. In the generalized linear model (GLM) literature (Nelder and Wedderburn (1972); McCullagh (2018)), g is often referred to as the *inverse link function*. Note that (1) can be rewritten as the GLCB form,

$$Y_t = g(X_t'\theta^*) + \epsilon_t, \tag{2}$$

where $\{\epsilon_t, t \in [T]\}$ are independent zero-mean noise, $\mathcal{H}_t^0$-measurable with $\mathbb{E}[\epsilon_t|\mathcal{H}_t^0] = 0$. Data generated from (1) automatically satisfies the sub-Gaussian condition:

$$\mathbb{E}\left[\exp(\lambda\epsilon_t)|\mathcal{H}_t^0\right] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right). \tag{3}$$

Throughout the paper, we denote $\sigma > 0$ as the sub-Gaussian parameter of the noise ϵ_t.

Remark 1. *In this paper, we focus on the GLM with exponential family* (1)*. In general, one can work with model* (2) *under the sub-Gaussian assumption* (3)*. Our analysis will still hold by considering maximum quasi-likelihood estimator for* (2)*. See more explanations in the appendix.*

2.2 The Delay Model

Unlike the traditional setting where each reward is immediately observed, here we consider the case where stochastic and unbounded *delays* are present in revealing the rewards. Let T be the number of total rounds. At round t, after the agent takes action a_t, the reward y_{t,a_t} may not be available immediately. Instead, it will be observed at the end of round $t + D_t$ where D_t is the delay at time t. We assume D_t is a non-negative random number which is independent of $\{D_s\}_{s \leq t-1}$ and $\{x_s, y_{s,a_s}, a_s\}_{s \leq t}$. First, we define the available information for the agent at each round.

Information structure under delays. At any round t, if $D_s + s \leq t-1$ (reward occurred in round s is available at the beginning of round t), then we call (s, x_s, y_{s,a_s}, a_s) the *complete information tuple* at round t. If $D_s + s \geq t$, we call (s, x_s, a_s) the *incomplete information tuple* at the beginning of round t. Define

$$\mathcal{H}_t = \{(s, x_s, y_{s,a_s}, a_s) \mid s + D_s \leq t-1\} \cup \{(s, x_s, a_s) \mid s \leq t-1, s + D_s \geq t\} \cup \{x_t\},$$

then $\mathcal{H}_t$ is the information (filtration) available at the beginning of round t for the agent to choose action a_t. In other words, $\mathcal{H}_t$ contains all the incomplete and complete information tuples up to round $t-1$ and the content vector x_t at round t.

Moreover define

$$\mathcal{F}_t = \{(s, x_s, a_s, y_{s,a_s}) \mid s + D_s \leq t\}. \tag{4}$$

Then $\mathcal{F}_t$ contains all the complete information tuples (s, x_s, a_s, y_{s,a_s}) up to the end of round t. Denote $\mathcal{I}_t = \mathcal{F}_t - \mathcal{F}_{t-1}$, $\mathcal{I}_t$ is the new complete information tuples revealed at the end of round t.

Performance criterion. Under the frequentist setting, assume there exists an unknown true parameter $\theta^* \in \mathbb{R}^d$. The agent's strategy can be evaluated by comparing her rewards to the best reward. To do so, define the optimal action at round t by $a_t^* = \arg\max_{a \in [K]} g(x_{t,a}'\theta^*)$. Then, the agent's total regret of following strategy π can be expressed as follows

$$R_T(\pi) := \sum_{t=1}^{T}\left(g\left(x_{t,a_t^*}'\theta^*\right) - g\left(x_{t,a_t}'\theta^*\right)\right),$$

where $a_t \sim \pi_t$ and policy π_t maps $\mathcal{H}_t$ to the probability simplex $\Delta^K := \{(p_1, \cdots, p_K) \mid \sum_{i=1}^K p_i = 1, p_i \geq 0\}$. Note that $R_T(\pi)$ is in general a random variable due to the possible randomness in π.

Assumptions. Through out the paper, we assume the following assumption on distribution γ and function g, which is standard in the generalized linear bandit literature (Filippi et al. (2010); Li et al. (2017); Jun et al. (2017)).

Assumption 1 (GLCB).

- *$\lambda_{\min}(\mathbb{E}[\frac{1}{K}\sum_{a\in[K]} x_{t,a}x'_{t,a}]) \geq \sigma_0^2$ for all $t \in [T]$.*
- *$\kappa := \inf_{\{\|x\|\leq 1, \|\theta-\theta^*\|\leq 1\}} \dot{g}(x'\theta) > 0$.*
- *g is twice differentiable. $\dot{g}$ and $\ddot{g}$ are upper bounded by L_g and M_g, respectively.*

In addition, we assume the delay sequence $\{D_t\}_{t=1}^T$ satisfies the following assumption.

Assumption 2 (Delay). *Assume $\{D_t\}_{t=1}^T$ are independent non-negative random variables with tail-envelope distribution (ξ_D, μ_D, M_D). That is, there exists a constant $M_D > 0$ and a distribution ξ_D with mean $\mu_D < \infty$ such that for any $m \geq M_D$ and $t \in [T]$,*

$$\mathbb{P}(D_t \geq m) \leq \mathbb{P}(D \geq m),$$

where $D \sim \xi_D$ and $\mathbb{E}[D] = \mu_D$. Furthermore, assume there exists $q > 0$ such that

$$\mathbb{P}(D - \mu_D \geq x) \leq \exp\left(\frac{-x^{1+q}}{2\sigma_D^2}\right).$$

Note that when $q = 1$, D is sub-Gaussian with parameter σ_D. When $q \in (0,1)$, D has near-heavy tail distribution. When D_i's are **iid**, the following condition guarantees Assumption 2:

$$\mathbb{P}(D_i - \mathbb{E}[D_i] \geq x) \leq \left(\frac{-x^{1+q}}{2\tilde{\sigma}_D^2}\right),$$

with some $\tilde{\sigma}_D > 0$ and $q > 0$.

For ease of reference (as there are many floating parameters in this paper), we summarize all the parameter definitions in Table 1.

Notation	Definition	Notation	Definition
K	number of arms	ξ_D	tail-envelope distribution for the delays
d	feature dimension	q	parameter of ξ_D
κ	$\inf_{\{\|x\|\leq 1, \|\theta-\theta^*\|\leq 1\}} \dot{g}(x'\theta)$	μ_D	expectation of ξ_D
θ^*	unknown true parameter	M_D	parameter of ξ_D
σ	sub-Gaussian parameter for ϵ_t	σ_D	parameter of ξ_D
L_g	upper bound on $\dot{g}$	σ_G	sub-Gaussian parameter of G_t
M_g	upper bound on $\ddot{g}$	μ'_D	expectation of **iid** delays
σ_0^2	lower bound on $\lambda_{\min}(\mathbb{E}[\frac{1}{K}\sum_{a\in[K]} x_{t,a}x'_{t,a}])$	D_{max}	upper bound on bounded delays

Table 1: Parameters in the GLCB model with delays.

3 Delayed Upper Confidence Bound (DUCB) for GLCB

In this section, we propose a UCB type of algorithm for GLCB, adapting the delay information in an online version. Let us first define some variables and state the main algorithm.

3.1 Algorithm: DUCB-GLCB

Denote $G_t = \sum_{s=1}^{t-1} \mathbb{I}\{s + D_s \geq t\}$ as the number of missing reward when the agent is making a prediction at round t. Denote $T_t = \{s : s \leq t-1, D_s + s \leq t-1\}$ as the set containing timestamps with complete information tuples at the beginning of round t. Further denote $W_t = \sum_{s\in T_t} X_s X'_s$ as the matrix consisting feature information with timestamps in T_t and $V_t = \sum_{s=1}^{t-1} X_s X'_s$ as the matrix consisting all available features at the end of round $t-1$. The main algorithm is given below.

Algorithm 1 DUCB-GLCB

```
1: Input: the total rounds $T$ , model parameters $d$ and $\kappa$, and tuning parameters $\tau$ and $\delta$.
2: Initialization: randomly choose $\alpha_t \in [K]$ for $t \in [\tau]$, set $V_{\tau+1} = \sum_{i=1}^{\tau} X_s X_s'$, $T_{\tau+1} := \{s : s \leq \tau, s + D_s \leq \tau\}$, $G_{\tau+1} = \tau - |T_{\tau+1}|$ and $W_{\tau+1} = \sum_{s \in T_{\tau+1}} X_s X_s'$
3: for $t = \tau + 1, \tau + 2, \cdots , T$ do
4:     Update Statistics: calculate the MLE $\hat{\theta}_t$ by solving $\sum_{s \in T_t} (Y_s - g(X_s'\theta)) X_s = 0$
5:     Update Parameter: $\beta_t = \frac{\sigma}{\kappa} \sqrt{\frac{d}{2} \log\left(1 + \frac{2(t-G_t)}{d}\right) + \log(\frac{1}{\delta})} + \sqrt{G_t}$
6:     Select Action: choose $a_t = \arg\max_{a \in [K]} \left( x_{t,a}' \hat{\theta}_t + \beta_t \|x_{t,a}\|_{V_t^{-1}} \right)$
7:     Update Observations: $X_t \leftarrow x_{t,a_t}$, $V_{t+1} \leftarrow V_t + X_t X_t'$ and $T_{t+1} \leftarrow T_t \cup \{s : s + D_s = t\}$, $G_{t+1} = t - |T_{t+1}|$, and $W_{\tau+1} = W_\tau + \sum_{s: s+D_s = t} X_s X_s'$
8: end for
```

Remark 2. *In step 4, we use Maximum Likelihood Estimators (MLEs) for the parameter estimation step at each round t. For more details on the derivation and explanation, we refer to the appendix.*

Remark 3 (Comparison to UCB-GLM Algorithm in Li et al. (2017))**.** *We make several adjustments to the UCB-GLM Algorithm in Li et al. (2017). First, in step 4 (statistics update), we only use data with timestamps in T_t to calculate the estimator using MLE. In this step, using data without reward will cause bias in the estimation. Second, when selecting the action in step 5, parameter β_t is updated adaptively at each round whereas in Li et al. (2017), the corresponding parameter is constant over time. Moreover, in step 4, we choose to use V_t to normalize the context vector $X_{t,a}$ instead of W_t.*

3.2 Preliminary Analysis

Denote $G_t^* = \max_{1 \leq s \leq t} G_s$ as the running maximum number of missing reward up to round t. The property of G_t and G_t^* is the key to analyze the regret bound for UCB algorithm. We next characterize the tail behavior of G_t and G_t^*.

Proposition 1 (Properties of G_t and $G_t^\star$)**.** *Assume Assumption 2. Denote $\sigma_G = \sqrt{\frac{I}{4} + \frac{\sigma_D^2(1+q)}{q}}$ with $I = \max\left\{ \sqrt[1+q]{2\log(2)\sigma_D^2}, \sqrt[q]{\frac{2\sigma_D^2}{1+q}} + 1 \right\}$. Then,*

1. *G_t is sub-Gaussian. Moreover, for all $t \geq 1$,*

$$\mathbb{P}\left(G_t \geq 2(\mu_D + M_D) + x\right) \leq \exp\left(\frac{-x^2}{2\sigma_G^2}\right). \tag{5}$$

2. *With probability $1 - \delta$,*

$$G_T^* \leq 2(\mu_D + M_D) + \sigma_G \sqrt{2\log(T)} + \sigma_G \sqrt{2\log\left(\frac{1}{\delta}\right)}, \tag{6}$$

where $G_T^* = \max_{1 \leq s \leq T} G_s$.

3. *Define $W_t = \sum_{s \in T_t} X_s X_s'$ where X_t is drawn* ***iid****. from some distribution γ with support in the unit ball $\mathbb{B}_d$. Furthermore, let $\Sigma := \mathbb{E}[X_t X_t']$ be the second moment matrix, and B and $\delta > 0$ be two positive constants. Then there exist positive, universal constants C_1 and C_2 such that $\lambda_{\min}(W_t) \geq B$ with probability at least $1 - 2\delta$, as long as*

$$t \geq \left(\frac{C_1\sqrt{d} + C_2\sqrt{\log(\frac{1}{\delta})}}{\lambda_{\min}(\Sigma)} \right)^2 + \frac{2B}{\lambda_{\min}(\Sigma)} + 2(\mu_D + M_D) + \sigma_G \sqrt{2\log\left(\frac{1}{\delta}\right)}. \tag{7}$$

The proof of Proposition 1 is deferred to the appendix. This Note that G_t is sub-Gaussian even when D has near-heavy tail distribution when $p \in (0, 1)$.

3.3 Regret Bounds

Theorem 2. *Assume Assumptions 1-2. Fix any δ. There exists a universal constant $C := C(C_1, C_2, M_D, \mu_D, \sigma_0, \sigma_G, \sigma, \kappa) > 0$, such that if we run DUCB-GLCB with $\tau := C\left(d + \log(\frac{1}{\delta})\right)$ and $\beta_t = \frac{\sigma}{\kappa}\sqrt{\frac{d}{2}\log\left(1 + \frac{2(t-G_t)}{d}\right) + \log(\frac{1}{\delta})} + \sqrt{G_t}$, then, with probability at least $1 - 5\delta$, the regret of the algorithm is upper bounded by*

$$\begin{aligned} R_T \quad \leq \quad & \tau + L_g \left[4\sqrt{\mu_D + M_D}\sqrt{Td\log\left(\frac{T}{d}\right)} + 2^{7/4}\sqrt{\sigma_G}\left(\log\left(\frac{1}{\delta}\right)\right)^{1/4}\sqrt{d\log\left(\frac{T}{d}\right)T} \right. \\ & \left. + 2^{7/4}\sqrt{\sigma_G}\left(\log\left(T\right)\right)^{1/4}\sqrt{d\log\left(\frac{T}{d}\right)T} + \frac{2d\sigma}{\kappa}\log\left(\frac{T}{d\delta}\right)\sqrt{T}\right]. \end{aligned} \tag{8}$$

For parameter definition, we refer to Table 1.The proof of Theorem 2 is deferred to the appendix.

Corollary 3 (Expected regret)**.** *Assume Assumptions 1-2. The expected regret is bounded by*

$$\mathbb{E}[R_T] = O\left(d\sqrt{T}\log(T) + \sqrt{\mu_D + M_D}\sqrt{Td\log\left(T\right)} + \sqrt{\sigma_G}\sqrt{Td}\left(\log(T)\right)^{3/4}\right). \tag{9}$$

Given the result in (8), (9) holds by choosing $\delta = \frac{1}{T}$ and using the fact that $R_T \leq T$.

The highest order term $O(d\sqrt{T}\log(T))$ does not depend on delays. Delay impacts the expected regret bound in two folds. First, the sub-Gaussian parameter σ_G appears in the second-highest order term. Second, the mean-related parameter $\mu_D + M_D$ appears in the third-order term. Note that here we include the log factors in deciding the highest order term, the second higest order term and so on. If we exclude the log terms, then both delay parameters impact the regret bound multiplicatively.

3.4 Tighter Regret Bounds for Special Cases

When the sequence $\{D_s\}_{s=1}^T$ satisfies some specific assumptions, we are able to provide tighter high probability bounds on the regret.

Proposition 4. *Under Assumption 1, we have:*

1. *If there exists a constant $D_{\max} > 0$ such that $\mathbb{P}(D_s \leq D_{\max}) = 1$ for all $s \in [T]$. Fix δ. There exists a universal constant $C > 0$ such that by taking $\tau = D_{\max} + C(d + \log(\frac{1}{\delta}))$, with probability $1 - 3\delta$, the regret of the algorithm is upper bounded by*

$$R_T \leq \tau + L_g\left(2\sqrt{D_{\max}}\sqrt{2Td\log\left(\frac{T}{d}\right)} + \frac{2d\sigma}{\kappa}\log\left(\frac{T}{d\delta}\right)\sqrt{T}\right). \tag{10}$$

   Therefore, $\mathbb{E}[R_T] = O\left(\sqrt{D_{\max}}\sqrt{dT\log(T)} + d\sqrt{T}\log(T)\right)$.

2. *Assume $D_1, \cdots, D_T$ are* ***iid*** *non-negative random variables with mean μ'_D that satisfy Assumption (2). There exists $C > 0$ such that by taking $\tau := C\left(d + \log(\frac{1}{\delta})\right)$, with probability $1 - 5\delta$, the regret of the algorithm is upper bounded by*

$$\begin{aligned} R_T \quad \leq \quad & \tau + L_g\left[4\sqrt{\mu'_D}\sqrt{Td\log\left(\frac{T}{d}\right)} + 2^{7/4}\sqrt{\sigma_G}\left(\log\left(\frac{1}{\delta}\right)\right)^{1/4}\sqrt{d\log\left(\frac{T}{d}\right)T}\right. \\ & \left. + 2^{7/4}\sqrt{\sigma_G}\left(\log\left(T\right)\right)^{1/4}\sqrt{d\log\left(\frac{T}{d}\right)T} + \frac{2d\sigma}{\kappa}\log\left(\frac{T}{d\delta}\right)\sqrt{T}\right]. \end{aligned}$$

   Therefore, $\mathbb{E}[R_T] = O\left(\left(\sqrt{\mu'_D} + \sqrt{\sigma_G}\log\left(T\right)^{3/4}\right)\sqrt{Td} + d\log\left(T\right)\sqrt{T}\right)$

4 Tighter Regret Bounds on Linear Contextual Bandits with Finite Actions

We now consider the important special case of linear contextual bandits. and tighten the $O(d)$ dependence from previous bounds to $O(\sqrt{d})$. This requires two new elements that we incorporate into DUCB-GLCB in Algorithm 1. First, instead of using MLE which is unbiased, here we use an unbiased estimator that incorporates all the contexts (including those contexts for which the rewards have not been received). In the linear contextual bandits setting, one can obtain analytical formulas for the estimation procedure. Second, we extend the Sup-Base UCB decomposition framework (first devised in Auer (2002) and subsequently adapted in Chu et al. (2011); Li et al. (2017)) to the current setting in order to resolve the reward dependency issue. This framework is a commonly used one in the literature that deals with the dependency issue, and provides a $O(\sqrt{dT})$ regret bound instead of a $O(d\sqrt{T})$ regret bound. Here we adapt this framework in the delayed reward setting.

In summary, the algorithm has two components, Delayed BaseLinUCB (Algorithm 2) and Delayed SupLinUCB (Algorithm 3). Delayed BaseLinUCB performs estimation and the confidence bound computation, using a subset Ψ_t of the past time steps as opposed to the set of all past time steps (note that when $t = 1$, the chosen subset Ψ_t is necessarily the empty set). This subset is carefully chosen in Delayed SupLinUCB to make sure rewards are indepenent when conditioned on the past selected contexts. Delayed SupLinUCB is further responsible for selecting an action at each time step.

Algorithm 2 Delayed BaseLinUCB at Step t

```
Input: $\Psi_t \subset \{1, 2, \cdots, t-1\}$.
$A_t = I_d + \sum_{\tau \in \Psi_t} x_{t,a_\tau} x'_{t,a_\tau}$
$c_t = \sum_{\tau \in \Psi_t} \mathbf{1}(D_\tau + \tau \leq t-1) y_{\tau,a_\tau} x_{\tau,a_\tau}$
$\theta_t = A_t^{-1} c_t$
Observe $K$ arm features, $x_{t,1}, x_{t,2}, \cdots, x_{t,K} \in \mathbb{R}^d$
for $a \in [K]$ do
    $w_{t,a} = \alpha_t \sqrt{x_{t,a}^T A_t^{-1} x_{t,a}}$
    $\hat{y}_{t,a} \leftarrow \theta_t^T x_{t,a}$
end for
```

Algorithm 3 Delayed SupLinUCB

```
Input: $T \in \mathbb{N}$, $S \leftarrow \log(T)$
$\Psi_1^s \leftarrow \emptyset$ for all $s \in [T]$
for $t = 1, 2, \cdots, T$ do
    $s \leftarrow 1$ and $\hat{A}_1 \leftarrow [K]$
    repeat
        Use Delayed BaseLinUCB with $\Psi_t^s$ to calculate the width, $w_{t,a}^s$, and upper confidence bound, $\hat{y}_{t,a}^s + w_{t,a}^s$, for all $a \in \hat{A}_s$
        if $w_{t,a}^s \leq 1/\sqrt{T}$ for all $a \in \hat{A}_s$ then
            Choose $a_t = \arg\max_{a \in \hat{A}_s} \left(\hat{y}_{t,a}^s + w_{t,a}^s\right)$, Update $\Psi_{t+1}^{s'} \leftarrow \Psi_t^{s'}$ for all $s' \in [S]$.
        else if $w_{t,a}^s \leq 2^{-s}$ for all $a \in \hat{A}_s$ then
            $\hat{A}_{s+1} \leftarrow \{a \in \hat{A}_s \mid \hat{y}_{t,a}^s + w_{t,a}^s \geq \max_{a' \in \hat{A}_s} (\hat{y}_{t,a'}^s + w_{t,a'}^s) - 2^{1-s}\}$, $s \leftarrow s + 1$
        else
            Choose $a_t \in \hat{A}_s$ such that $w_{t,a_t}^s > 2^{-s}$, Update
                $\Phi_{t+1}^{s'} \leftarrow \begin{cases} \Phi_t^{s'}, \cup\{t\} & \text{if } s = s' \\ \Phi_t^{s'}, & \text{otherwise} \end{cases}$
        end if
    until an action $a_t$ is found.
end for
```

Remark 4. *There are two modifications compared to Algorithm 2 in Chu et al. (2011). First, the estimator θ_t (in step 4) is a biased estimator. We use all the features in matrix A_t and only use features with observed rewards in vector c_t. In particular, when the indicator $\mathbf{1}(D_\tau + \tau \leq t-1)$ evaluates to 1, the reward corresponding to the action taken at time step τ has been received by the end of (and possibly prior to) $t-1$ (and hence available at the beginning of t); all the other rewards (i.e. those that have not been received by $t-1$) are excluded. In comparison, Chu et al. (2011) construct an unbiased estimator in each time step. Second, the width parameter α_t (in step 7) is time-dependent and adapts to new information (based on the delays) in each round. In comparison, the width parameter is constant in Chu et al. (2011) that only depends on the horizon T.*

Theorem 5 (Regret on Delayed SupLinUCB-BaseLinUCB)**.** *If Delayed SupLinUCB is run with $\alpha_t = \bar{\alpha} + G_t + 1$, where $\bar{\alpha} = \sqrt{\frac{1}{2}\ln\left(\frac{2TK\log(T)}{\delta}\right)}$, then with probability at least $1-2\delta$, the regret of the algorithm is*

$$O\left(\sqrt{Td}\left((\sigma_G+1)\log^{3/2}(\frac{TK\log T}{\delta}) + \log(\frac{TK\log T}{\delta})(1+\mu_D+M_D+\sigma_G\sqrt{\log\frac{1}{\delta}})\right)\right). \quad (11)$$

The proof of Theorem 5 requires of modification of two lemmas in Chu et al. (2011). Lemma 6 is a modification of (Chu et al., 2011, Lemma 1) and Lemma 7 is a modification of (Chu et al., 2011, Lemma 6). We defer the detailed proofs of Lemmas 6-7 to the appendix. Proof of Theorem 5 is also given in the appendix.

In the regret bound (11), the delay parameters (μ_D, M_D, σ_D) appear on the highest order term $\sqrt{Td}$. Although the highest order term $\sqrt{Td}$ is removed from (8), the delay on order $O(\sqrt{Td})$ is essential and this is also true for (8).

Lemma 6. *Suppose the input index set Φ_t in Delayed BaseLinUCB is constructed so that for fixed x_{τ,a_τ} with $\tau \in \Phi_t$, the rewards y_{τ,a_τ} are independent random variables with means $\mathbb{E}[y_{\tau,a_\tau}] = x'_{\tau,a_\tau}\theta^*$. Suppose $\{G_t\}$ is fixed and given. Then, with probability at least $1-\delta/T$, we have for all $a \in [K]$ that*

$$|\hat{y}_{t,a} - x'_{t,a}\theta^*| \leq \left(1 + \sqrt{\frac{1}{2}\ln\left(\frac{2TK}{\delta}\right)} + G_t\right) s_{t,a}.$$

Lemma 7. *Assume G_T^* is fixed and given. For all $s \in [S]$,*

$$\Psi_{T+1}^s \leq 5 \cdot 2^s \left(\sqrt{2}\bar{\alpha}(G_T^* + \bar{\alpha})\right)\sqrt{d|\Psi_{T+1}^s|}$$

Remark 5 (Why Assumption 1 can be dropped in Theorem 5)**.** *There are essentially two methods to guarantee a positive lower bound $\lambda_{\min}(\sum_{s=1}^t X_s X_s')$. One method is to randomly sample actions for τ rounds. In this way, (Li et al., 2017, Proposition 1) guarantees a positive lower bound on $\lambda_{\min}(\sum_{s=1}^t X_s X_s')$. This is the method adopted in Algorithm 1 and Theorem 2. The other method adds a regularization term. This is adopted in the definition of A_t (See Algorithm 2 and Theorem 5). This method corresponds to the Ridge regression when estimating parameter θ_t.*

5 Conclusion

Beyond contextual bandits and looking at the broader landscape of data-driven decision making, delays have emerged to be an important phenomenon in several domains, including, among other things, distributed stochastic optimization (Bertsekas and Tsitsiklis (1997); Zhou et al. (2018b)), multi-agent game-theoretical and reinforcement learning (Zhou et al. (2017); Grover et al. (2018a); Guo et al. (2019); Mertikopoulos and Zhou (2019)), real-time scheduling in large-scale systems (Pinedo; Mehdian et al. (2017); Mahdian et al. (2018)). Data-driven decision making with imperfect information is an emerging research paradigm and much remains to be understood in regards to how decision-making needs to be adapted in the presence of delays.

References

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

Agrawal, S. and Goyal, N. (2013a). Further optimal regret bounds for thompson sampling. In *Artificial intelligence and statistics*, pages 99–107.

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

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. (2015). Online decision-making with high-dimensional covariates.

Bertsekas, D. P. and Tsitsiklis, J. N. (1997). Parallel and distributed computation: Numerical methods.

Bistritz, I., Zhou, Z., Chen, X., Bambos, N., and Blanchet, J. (2019). Online exp3 learning in adversarial bandits with delayed feedback. In *Advances in Neural Information Processing Systems*.

Bubeck, S., Cesa-Bianchi, N., et al. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122.

Cesa-Bianchi, N., Gentile, C., and Mansour, Y. (2019). Delay and cooperation in nonstochastic bandits. *The Journal of Machine Learning Research*, 20(1):613–650.

Chapelle, O. (2014). Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1097–1105. ACM.

Chen, K., Hu, I., Ying, Z., et al. (1999). Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *The Annals of Statistics*, 27(4):1155–1163.

Chow, S.-C. and Chang, M. (2011). *Adaptive design methods in clinical trials*. Chapman and Hall/CRC.

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*, pages 208–214.

Desautels, T., Krause, A., and Burdick, J. W. (2014). Parallelizing exploration-exploitation tradeoffs in gaussian process bandit optimization. *The Journal of Machine Learning Research*, 15(1):3873–3923.

Dudik, M., Hsu, D., Kale, S., Karampatziakis, N., Langford, J., Reyzin, L., and Zhang, T. (2011). Efficient optimal learning for contextual bandits. *arXiv preprint arXiv:1106.2369*.

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

Garrabrant, S., Soares, N., and Taylor, J. (2016). Asymptotic convergence in online learning with unbounded delays. *arXiv preprint arXiv:1604.05280*.

Grover, A., Al-Shedivat, M., Gupta, J. K., Burda, Y., and Edwards, H. (2018a). Learning policy representations in multiagent systems. *arXiv preprint arXiv:1806.06464*.

Grover, A., Markov, T., Attia, P., Jin, N., Perkins, N., Cheong, B., Chen, M., Yang, Z., Harris, S., Chueh, W., et al. (2018b). Best arm identification in multi-armed bandits with delayed feedback. *arXiv preprint arXiv:1803.10937*.

Guo, X., Hu, A., Xu, R., and Zhang, J. (2019). Learning mean-field games. In *Advances in Neural Information Processing Systems*.

Joulani, P., Gyorgy, A., and Szepesvári, C. (2013). Online learning under delayed feedback. In *International Conference on Machine Learning*, pages 1453–1461.

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*, pages 99–109.

Kannan, P., Chang, A.-M., and Whinston, A. B. (2001). Wireless commerce: marketing issues and possibilities. In *Proceedings of the 34th Annual Hawaii International Conference on System Sciences*, pages 6–pp. IEEE.

Kitagawa, T. and Tetenov, A. (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616.

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*, pages 661–670. ACM.

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*, pages 2071–2080. JMLR. org.

Mahdian, S., Zhou, Z., and Bambos, N. (2018). Robustness of join-the-shortest-queue scheduling to communication delay. In *2018 Annual American Control Conference (ACC)*, pages 3708–3713. IEEE.

Mandel, T., Liu, Y.-E., Brunskill, E., and Popović, Z. (2015). The queue method: Handling delay, heuristics, prior data, and evaluation in bandits. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

McCullagh, P. (2018). *Generalized linear models*. Routledge.

Mehdian, S., Zhou, Z., and Bambos, N. (2017). Join-the-shortest-queue scheduling with delay. In *2017 American Control Conference (ACC)*, pages 1747–1752. IEEE.

Mertikopoulos, P. and Zhou, Z. (2019). Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173(1-2):465–507.

Mesterharm, C. (2005). On-line learning with delayed label feedback. In *International Conference on Algorithmic Learning Theory*, pages 399–413. Springer.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

Neu, G., Antos, A., György, A., and Szepesvári, C. (2010). Online markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems*, pages 1804–1812.

Ostrovsky, E. and Sirota, L. (2014). Exact value for subgaussian norm of centered indicator random variable. *arXiv preprint arXiv:1405.6749*.

Pike-Burke, C., Agrawal, S., Szepesvari, C., and Grunewalder, S. (2017). Bandits with delayed, aggregated anonymous feedback. *arXiv preprint arXiv:1709.06853*.

Pinedo, M. *Scheduling*, volume 29. Springer.

Quanrud, K. and Khashabi, D. (2015). Online learning with adversarial delays. In *Advances in neural information processing systems*, pages 1270–1278.

Russo, D. and Van Roy, B. (2014). Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243.

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

Schwartz, E. M., Bradlow, E. T., and Fader, P. S. (2017). Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522.

Swaminathan, A. and Joachims, T. (2015). Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16(52):1731–1755.

Vernade, C., Cappé, O., and Perchet, V. (2017). Stochastic bandit models for delayed conversions. *arXiv preprint arXiv:1706.09186*.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Weinberger, M. J. and Ordentlich, E. (2002). On delayed prediction of individual sequences. *IEEE Transactions on Information Theory*, 48(7):1959–1976.

Zhou, Z., Athey, S., and Wager, S. (2018a). Offline multi-action policy learning: Generalization and optimization. *arXiv preprint arXiv:1810.04778*.

Zhou, Z., Mertikopoulos, P., Bambos, N., Glynn, P., Ye, Y., Li, L.-J., and Fei-Fei, L. (2018b). Distributed asynchronous optimization with unbounded delays: How slow can you go? In *International Conference on Machine Learning*, pages 5965–5974.

Zhou, Z., Mertikopoulos, P., Bambos, N., Glynn, P. W., and Tomlin, C. (2017). Countering feedback delays in multi-agent learning. In *Advances in Neural Information Processing Systems*, pages 6171–6181.