Batched Learning in Generalized Linear Contextual Bandits with General Decision Sets

Zhimei Ren* Zhengyuan Zhou† Jayant R. Kalagnanam‡

Abstract

In real-world adaptive personalized decision making, due to physical and/or resource constraints, a decision maker often does not have the luxury of immediately incorporating the feedback from the previous individual into forming new policies for future individuals. This is an important aspect that has largely been abstracted away from the traditional online learning/decision making literature. In this paper, we study the problem of batched learning in generalized linear contextual bandits where the decision maker, unlike in traditional online learning, can only access feedback at the end of a limited number of batches, and when selecting actions within a batch, can only use information from prior batches. We provide a lower bound that characterizes the fundamental limit of performance in this setting and then give a UCB-based batched learning algorithm whose regret bound, obtained using a self-normalized martingale style analysis, nearly matches this lower bound. Our results provide a novel inquiry into generalized linear contextual bandits with arbitrary action sets, which include several bandits setting as special cases and thus shed light on batch-constrained decision making in general.

1 Introduction

The growing availability of user-specific data has welcomed the exciting era of personalized decision making: a paradigm of tailoring decisions according to each individual's distinct characteristics, thereby exploiting heterogenity in a population so as to achieve better outcomes. Such heterogeneity (i.e. best decisions often vary for different individuals) is ubiquitous across a variety of application domains, including medical treatment selection, product recommendation and online advertising (Bertsimas and Mersereau, 2007; Li et al., 2010; Chapelle, 2014; Bastani and Bayati, 2020; Schwartz et al., 2017). Rising to this opportunity, contextual bandits have emerged to be the predominant mathematical framework that is at once elegant and powerful: its three core components—the contexts (representing individual characteristics), the actions (representing the recommended items), and the rewards (representing the observed feedback)—capture the salient aspects of the problem and provide fertile ground for developing algorithms that contribute to making quality decisions.

Roughly speaking, the existing contextual bandits literature can be divided into two categories: online contextual bandits and offline contextual bandits. In online contextual bandits (Bubeck et al., 2012; Lattimore and Szepesvári, 2018; Slivkins et al., 2019), individual data points arrive sequentially and the decision

*Department of Statistics, Stanford University. Email: zren@stanford.edu.
†Stern School of Business, New York University. Email: zzhou@stern.nyu.edu.
‡IBM Research. Email: jayant@us.ibm.com

maker can actively interact with the data-collection process and adaptively make his next decision by incorporating feedback from the current data point. This has motivated an extensive line of work (e.g. Li et al. (2010); Rusmevichientong and Tsitsiklis (2010); Filippi et al. (2010); Rigollet and Zeevi (2010); Chu et al. (2011); Goldenshluger and Zeevi (2013); Agrawal and Goyal (2013); Russo and Van Roy (2016); Jun et al. (2017); Li et al. (2017); Abeille et al. (2017)) that has developed online contextual bandits algorithms that effectively trade off exploration and exploitation, a key challenge therein. On the other hand, in offline contextual bandits (Dudík et al., 2011; Zhang et al., 2012; Zhao et al., 2014; Swaminathan and Joachims, 2015; Rakhlin and Sridharan, 2016; Kallus, 2018; Kitagawa and Tetenov, 2018; Joachims et al., 2018), the decision maker is given a full batch dataset that has already been collected, and his goal is to learn from that dataset an effective policy which will be deployed in the future. Here, learning a policy is offline and one-shot as the decision maker does not have the opportunity to adapt one's decisions based on future outcomes. Offline contextual bandits have also attracted great attention in machine learning, statistics and casual inference, leading to several algorithms that perform efficient exploitation.

However, in practical situations, the reality often stands somewhere in between. In particular, the decision maker is often able to perform active learning, but such adaptation is sometimes limited to a fixed number of rounds of interaction. For instance, in medical treatment selection in clinical trials (Robbins, 1952), each trial involves applying medical treatment to a group of patients, with the medical outcomes for the entire group of patients available at the end of this trial. This batch of data is then analyzed and utilized to design the next trial. Here, the decision maker does have the flexibility in adaptive learning since there will be subsequent trials after the first one. However, such flexibility is limited since the number of patients tend to far exceed the number of trials that is feasible to conduct in standard medical practice. Another example is product recommendation (Bertsimas and Mersereau, 2007; Schwartz et al., 2017), where the marketer, when running a promotion campaign for its product, usually sends out product offers to a batch of customers at once. Customers' feedback will then be collected in the aggregate at the end and analyzed to design the next round of promotions in the targeted customer population.

Consequently, motivated by the above considerations, an important and practical question—unanswered by existing online or offline contextual bandits literature—immediately arises: given a learning horizon T (corresponding to T individuals) and at most M batches where the decision maker can partition the T individuals into, how should the size of each batch be designed and what to do within each batch? Recognizing the significance of this direction, recent works (Perchet et al., 2016; Gao et al., 2019) have made progress by studying the batched learning problem in the simple multi-armed bandit (MAB) problems, where Perchet et al. (2016) focused on two-armed MABs and Gao et al. (2019) generalized the results to K-armed MAB settings. However, although pioneering in this direction, MAB models are simply too simple as they do not capture individuals' characteristics. Consequently, in MABs, the decisions can only be made at a population level (selecting a single action for the entire population), rather than personalized at an individual level, which severly limits its practical applicability.

Very recently, Han et al. (2020) has made the pioneering attempt to incorporate personalized decision making problems in this context. Although their results are interesting, fruitful and inspiring, the setting is limited in several ways. First, it considered linear contextual bandits, which do not capture the class of logistic contextual bandits (a widely used framework in the binary-outcome based applications, e.g. click/no-click, purchase/no-purchase, recovery/no-recovery). Second, it only considered finite decision sets (with a restriction on maximum number of actions). This is inadequate in applications where the decision is continuous, including for instance the dosage of a drug for a patient, the coupon promotion amount and/or

price for a customer, news articles and/or videos (characterized by a continuous vector) for a user. Third, the proof techniques in Han et al. (2020) rely on these setting-specific assumptions, and cannot be directly extended to the more general setting studied here. As such, in this paper, we fill in this gap by providing the first broad inquiry into batched learning where individual characteristics provide the basis for personalized decision making under the broad class of generalized linear models with general decision sets and generic contexts and aim to delineate, in this much more general and practical setting, how the batch constraints impact the performance of decision making. We do point out, however, that our results here do not imply the results in Han et al. (2020) either (including the remarkable doubly-exponential regret bounds), as the specific assumptions made there allow for different learning limits and hence different bounds (both upper and lower bounds). Finally, we would like to note that in our paper, we consider a setting where no assumption is made about the generating mechanism of the contexts, i.e., x_t can be generated either adversarially or stochastically; there is another regime where we can resrtict our attention to a subset of context generating mechanisms (x_t generated in an i.i.d. fashion for example) and potentially achieve a regret of $\log T$ scale—we leave that for future work.

1.1 Our Contributions

We provide two main contributions. First, we study batched learning in generalized linear contextual bandits (GLCB) with general decision sets and provide a regret lower bound of $c \cdot \max\left\{\sqrt{\min(d,M)} \cdot T/M, d\sqrt{T} \cdot \left(T/d^2\right)^{\frac{1}{2(2^M-1)}}\right\}$, where T is the horizion (i.e. number of individuals), d is the feature dimension, M is the maximum number of batches allowed and c is some universal numerical constant (that is independent of d, M, T). This lower bound—which is established through a novel information-theoretic argument—characterizes the fundamental decision making limits and indicates that at least $O(\min(\lceil\sqrt{T/d}\rceil, \lceil T/d^2\rceil))$ batches[1] are needed to achieve the corresponding online minimax optimal regret rate $O(d\sqrt{T})$ where no batch constraints are present (Rusmevichientong and Tsitsiklis, 2010). Second, by adapting an upper confidence bound (UCB) style exploration to our setting, we propose the *Batched Upper Confidence Bound Algorithm for Generalized Linear Contexual Bandit* (BUCB-GLCB), and establish an $\tilde{O}(dT/M + d\sqrt{T})$ upper bound for the expected regret. This upper bound is established using a novel application of self-normalized martingales without any distributional assumption on the contexts. Further, this upper bound indicates that at most $O(\lceil\sqrt{T}\rceil)$ batches are needed to achieve the online minimax optimal regret rate $O(d\sqrt{T})$. Together with our lower bound above, we see that when the number of batches M is larger than $O(\sqrt{T})$, the upper bound is tight; when the number of batches M is less than $O(\sqrt{T})$, the upper bound is loose by at most a factor of $\max(\sqrt{d}, d/\sqrt{M})$. Put together, we see that BUCB-GLCB achieves the minimax-optimal performance, up to at most a factor of $O(\max(\sqrt{d}, d/\sqrt{M}))$;[2] further, in the low dimensional regime (when d is considered a constant to the growing T), it is in fact minimax optimal.

2 Problem Formulation

We formalize adaptive batched learning in generalized linear contextual bandits with general decision sets.

[1] $\lceil x\rceil$ denotes the smallest integer that is not less than x.

[2] To put this in the perspective of contextual bandits literature, state-of-the-art regret bounds of the widely used Thompson sampling algorithm are at least a factor of $\Omega(d)$ loose from the minimax optimal regret rate (Hamidi and Bayati, 2020).

2.1 Notation

For a positive integer n, let $[n] \triangleq \{1, \cdots, n\}$. For a vector v, $\|v\|_A$ denotes the norm induced by a positive semi-definite matrix A: $\|x\|_A = \sqrt{x^T A x}$. For two square matrices A, B, $A \preceq B$ denote that the difference $B - A$ is symmetric and positive semi-definite. We write $\mathsf{Poly}(n)$ and $\mathsf{Polylog}(n)$ for terms that grow polynomially and polylogarithmically with n.

2.2 Generalized Linear Contextual Bandits

Let T be the time horizon of the problem, $\mathcal{X}$ be a context set and $\mathcal{A}$ be an action set for the decision maker: both $\mathcal{X}$ and $\mathcal{A}$ are arbitrary sets, although a standard setting that suffices for practical purposes is that $\mathcal{X}$ and $\mathcal{A}$ are subsets of some finite-dimensional vector spaces (i.e. $\mathcal{X} \subset \mathbf{R}^{d_1}$ and $\mathcal{A} \subset \mathbf{R}^{d_2}$). At each time $t \in [T]$ in the online decision making process, a context x_t arrives, and if the decision maker selects action $a \in \mathcal{A}$ at this time, the resulting reward at time t is:

$$r_{t,a} = \mu(\langle \phi(x_t, a), \theta^\star \rangle) + \xi_t,$$

where $\phi : \mathcal{X} \times \mathcal{A} \to \mathbf{R}^d$ is a given feature map that featurizes a given context-action pair into some d-dimensional Euclidean space, $\theta^\star \in \mathbf{R}^d$ is the underlying parameter vector for the linear contextual bandits, the noises $\{\xi_t\}_{t=0}^{\infty}$ is a sequence of zero-mean independent 1-sub-Gaussian[3] random variables, and $\mu(\cdot)$ is an *inverse link function*. GLCB is quite general: when $\mu(\cdot)$ is the identity function, it reduces to linear contextual bandits; when $\mu(x) = 1/(1 + \exp(-x))$, it reduces to logistic contextual bandits.

Throughout the paper, we adopt the standard regularity assumptions (Li et al., 2017) on the inverse link function: $\mu(\cdot)$ is monotonically increasing, Lipschitz with Lipschitz constant L_μ, twice differentiable with its first derivative μ' satisfies that $\inf_{\|\theta - \theta^\star\|_2 \le 2, \|\phi\|_2 \le 1} \mu'(\phi^\top \theta) \ge \kappa > 0$ and second dereivative $\mu'' \ge 0$. Additionally, we work in the standard low-dimensional regime (Chu et al., 2011): $d^2 \le T$. We assume the underlying parameters $\theta^\star$ and the feature map ϕ are bounded, and for simplicity, we normalize everything into unit scale and assume $\|\theta^\star\|_2 \le 1$; $\|\phi(x, a)\|_2 \le 1, \forall x \in \mathcal{X}, \forall a \in \mathcal{A}$ (the general case can be easily derived and is presented later as well). Throughout, we assume T, d, M and κ are known a priori. Finally, we use a_t to denote the action selected at time t (note that a_t is a random variable) and r_{t,a_t} to denote the reward realized at time t under action a_t: when there is no confusion, we also write r_t in its replacement. As noted, generalized linear contextual bandits include both logistic and linear contextual bandits as special cases. It's also important to point out that the generality of our framework also stems from the general decision sets we are working. In particular, *even* in the subclass of linear contextual bandits, the general decision set structure contains many classes of well-known bandits problems as special cases. Below we briefly highlight a few:

1. **K-armed Linear Contextual Bandits** (Chu et al., 2011; Goldenshluger and Zeevi, 2013; Bastani and Bayati, 2020; Bastani et al., 2017) This is a setting where a finite set $\mathcal{A} = \{a_1, a_2, \ldots, a_K\}$ of K actions are available. Each action a_i has an associated parameter vector $\beta_i \in \mathbf{R}^d$. When a context vector $x_t \in \mathbf{R}^d$ arrives, if action a_i is chosen, then a reward $r_{t,a_i} = \langle x_t, \beta_i \rangle + \xi_t$ is obtained. To see that this is a special case of our setting, note that given a K-armed linear contextual bandit, we can take $\theta^* = [\beta_1, \beta_2, \ldots, \beta_K]$ and the following feature map: $\phi(x_t, a_i) = [\mathbf{0}, \ldots, \mathbf{0}, x_t, \mathbf{0}, \ldots, \mathbf{0}]$: that is, a Kd-dimensional vector with all zeros except at the i-th block. It is easy to see that $\langle x_t, \beta_i \rangle = \langle \phi(x_t, a_i), \theta^* \rangle$.

[3] A random variable $X \in \mathbb{R}$ is said to be σ^2-sub-Gaussian if $\mathbb{E}[X] = 0$ and $\mathbb{E}[\exp(sX)] \le \exp(\sigma^2 s^2/2)$, for any $s \in \mathbb{R}$.

2. **Linear Bandits** (Rusmevichientong and Tsitsiklis, 2010; Bubeck et al., 2012; Abeille et al., 2017) The reward at time t is given by $r_t = \langle a_t, \theta^* \rangle + \xi_t$, where $a_t \in \mathcal{A}$, typically a convex and compact subset of $\mathbf{R}^d$ (often normalized to be a unit ball). To see that this is a special case of our setting, simply specialize our setting to the case where contexts are absent and take $\phi(x_t, a_t) = a_t$.

3. **Combinatorial Bandits** (Cesa-Bianchi and Lugosi, 2012; Durand and Gagné, 2014; Combes et al., 2015) In a combinatorial bandit, actions come from a combinatorial set: $a_t \in \{0,1\}^d$ and the key feature in a combinatorial bandit is that there is a finite but large number of actions. Further the reward is $r_t = \langle a_t, x_t \rangle + \xi_t$. Depending on whether x_t is stochastically generated or adversarially generated, a combinatorial bandit further divides into stochastic combinatorial bandits and adversarial combinatorial bandits. In either case, one can see that a combinatorial bandit is a special case of our setting by taking $\phi(x_t, a_t) = \langle a_t, x_t \rangle$ and $\theta^* = 1$.

2.3 Batched Learning

In standard online learning, the decision maker immediately observes the reward r_{t,a_t} after selecting action a_t at time t. In contrast, here we consider a *batched learning* setting, where the decision maker is only allowed to partition the T units into (at most) M batches, and the reward corresponding to each unit in a batch can only be observed at the end of the batch. More specifically, given M, the decision maker needs to choose a batched learning algorithm $\mathbf{Alg} = (\mathcal{T}, \pi)$, where 1) a *grid* $\mathcal{T} = \{t_1, t_2, \cdots, t_M\}$ (with $0 = t_0 < t_1 < \cdots < t_M = T$) partitions the T units into M batches (the first batch contains units $t_0 + 1$ to t_1; the second batch contains units $t_1 + 1$ to t_2 and so on); 2) a sequential batch policy $\pi = (\pi_1, \pi_2, \ldots, \pi_T)$ such that each π_t can only use reward information from all the prior batches as well as all the contexts that are observed up to and including t.

Note that since M is the maximum number of batches, the decision maker is technically allowed to choose a grid that partitions the T units into less than M batches. However, since doing so will only decrease the amount of information available to the decision maker, we can without loss of generality assume the decision maker will always choose a grid of M batches. Additionally, $M = T$ corresponds exactly to standard online learning. In this case, the decision maker need not select the grid. Consequently, the adaptive batched learning setting has a more complex decision space—one that entails selecting both the grid and the policy. Finally, the performance of a batched learning algorithm **Alg** can be measured by regret, which compares the cumulative reward obtained by **Alg** to that obtained by an **optimal** policy (i.e. an oracle that knows $\theta^\star$ and hence the optimal action for each unit):

Definition 2.1. *The regret of **Alg** is defined as:* $R_T(\textbf{\textit{Alg}}) \triangleq \sum_{t=1}^{T} \left(\max_{a \in \mathcal{A}} \mu(\langle \phi(x_t, a), \theta^\star \rangle) - \mu(\langle \phi(x_t, a_t), \theta^\star \rangle)\right)$, *where* $a_1, a_2, \ldots, a_T$ *are actions generated by **Alg**.*

Remark 2.1. *Although the form of regret defined here is the same as that in standard online learning, the problem here is much more difficult: in batched learning, batches induce delays in obtaining reward feedback, and hence the decision maker cannot immediately incorporate the feedback into her subsequent decision making process. Nevertheless, we pursue an ambitious research agenda by still using the same regret metric as in standard online learning.*

3 A Regret Lower Bound

In this section we characterize the fundamental learning limits of the batched decision making problem.

Theorem 3.1. *For any batched learning algorithm* ***Alg****,*

$$\sup_{\substack{\theta^\star:\|\theta^\star\|_2\le 1,\ \mathcal{D}\left(\{x_t\}_{t=1}^T\right)\in\mathbf{D}\\ \mathcal{A},\ \phi(\cdot,\cdot):\|\phi(\cdot,\cdot)\|_2\le 1.}} \mathbf{E}_{\theta^\star}[R_T(\boldsymbol{Alg})] \ge c\cdot\max\left\{\frac{\sqrt{\min(d,M)}T}{M}, d\sqrt{T}\left(\frac{T}{d^2}\right)^{\frac{1}{2(2^M-1)}}\right\}, \tag{3.1}$$

where $c>0$ is a numerical constant independent of (T,M,d), $\mathcal{D}\left(\{x_t\}_{t=1}^T\right)$ refers to the generating distribution of $\{x_t\}_{t=1}^T$ and $\mathbf{D}$ denotes the set of all the possible choices of the context generating mechanism.

Note that here the supreme is taken over all $\theta^\star$ such that $\|\theta^\star\|_2\le 1$, and all possible choices of the generating mechanism of x_t, the action set $\mathcal{A}$, the feature map ϕ and distribution of the noise. It suffices to show that there *exists* a distribution of x_t, a form of the action set and feature map and a noise distribution such that the above inequality holds. We consider two cases to obtain the above two terms separately. For both of the cases, the feature map is the pointwise multiplication: $\phi(x_t,a)=(x_{t,j}a_j)_{j\in[d]}$; the inverse link function $\mu(\cdot)$ is the identity function.

Proof of the first term We here consider a case where, roughly speaking, the context in each batch is orthogonal to each other (so that the observations in one batch do not contribute to the learning in the subsequent batches). We construct (a prior for) $\theta^\star$ in the following way: let $d_M=\min(M,d)$; for $j\le d_M$, let $\theta_j^\star=-1/\sqrt{d_M}$ or $1/\sqrt{d_M}$ with equal probability, and for $j>d_M$, let $\theta_j^\star=0$. We further claim by an averaging argument that for *any* grid design $\mathcal{T}$, there must exist d_M batches $\{m_1,\ldots,m_{d_M}\}$ such that $\sum_{k=1}^{d_M}(t_{m_k}-t_{m_k-1})\ge d_MT/M$.

We then pick the context x_t sampled from a distribution where suppose t is in batch m_k—that is, $t_{m_k-1}<t\le t_{m_k}$—the k_{th} coordinate of x_t takes value in -1 and 1 with equal probability and the other coordinates take value 0. In the remaining batches (that are not in $\{m_1,\ldots,d_{d_M}\}$), we simply set $x_t=0$. The action set $\mathcal{A}$ is set to be $\{-1,1\}^d$. Then in batch m_k, there is no available information on the k_{th} coordinate of $\theta^\star$ from the previous batches, and a wrong decision incurs a regret at least $2/\sqrt{d_M}$. Averaging over $\theta^\star$, we obtain that the average regret is at least $\sum_{k=1}^{d_M}(t_{m_k}-t_{m_k-1})/\sqrt{d_M}\ge\sqrt{d_M}T/M$. We finish the proof by noting the worst-case regret is always lower bounded by the average regret.

Proof of the second term We now consider the case where x_t is generated *iid* from a distribution ν such that for all $j\in[d]$, $\mathbf{E}_\nu[|x_{t,j}|]\ge c_{\min}$ and $|x_{t,j}|\le 1$ (e.g. $x_{t,j}\sim$Uniform $[-1,1]$ and $c_{\min}=1/2$). We further let $\mathcal{A}=[-\Delta,\Delta]^d$, where $0<\Delta\le 1/\sqrt{d}$ is determined later (by construction $\|\phi(x_t,a)\|_2\le\sqrt{d\Delta^2}\le 1$). The parameter $\theta^\star$ is (to be) chosen from $\{-1/\sqrt{d},1/\sqrt{d}\}^d$ and the noise follows a standard normal distribution. Let $\mathbf{x}^t$ denote the vector of context up to time t, i.e., $(x_1,\ldots,x_t)$. For any algorithm **Alg**, we have:

$$\begin{aligned}
\mathbf{E}_{\theta^\star}[R_T(\mathbf{Alg})] &= \sum_{m=1}^{M}\sum_{t=t_{m-1}+1}^{t_m}\mathbf{E}_{\theta^\star}\left[\sup_{a\in[-\Delta,\Delta]^d}\langle\phi(x_t,a),\theta^\star\rangle-\langle\phi(x_t,a_t),\theta^\star\rangle\right]\\
&= \sum_{m=1}^{M}\sum_{t=t_{m-1}+1}^{t_m}\sum_{j=1}^{d}\mathbf{E}_{\theta^\star}\left[\sup_{a_j\in[-\Delta,\Delta]}x_{t,j}a_j\theta_j^\star-x_{t,j}a_{t,j}\theta_j^\star\right]\\
&\ge \Delta\sum_{m=1}^{M}\sum_{t=t_{m-1}+1}^{t_m}\sum_{j=1}^{d}\mathbf{E}\left[|x_{t,j}\theta_j^\star|\mathbf{1}_{\left\{\mathrm{sgn}(x_{t,j}\theta_j^\star)\ne\mathrm{sgn}(a_{t,j})\right\}}\right]
\end{aligned} \tag{3.2}$$

Taking expectation conditional on $\mathbf{x}^t$, we have

$$(3.2) \geq \frac{\Delta}{\sqrt{d}} \sum_{m=1}^{M} \sum_{t=t_{m-1}+1}^{t_m} \sum_{j=1}^{d} \mathbf{E}_{\theta^\star} \left[|x_{t,j}| p_{j,t}^{\theta^\star}(\mathbf{x}^t) \right], \tag{3.3}$$

where $p_{j,t}^{\theta^\star}(\mathbf{x}^t) = \mathbb{P}_{\theta^\star} \left(\mathrm{sgn}(x_{t,j}\theta_j^\star) \neq \mathrm{sgn}(a_{t,j}) \mid \mathbf{x}^t \right)$. Lemma 3.2 then establishs a lower bound for $p_{j,t}^{\theta^\star}(\mathbf{x}^t)$, whose proof can be found in supplementary material.

Lemma 3.2. *There exists $\theta^\star \in \{-1/\sqrt{d}, 1/\sqrt{d}\}^d$ s.t.,*

$$\frac{\Delta}{\sqrt{d}} \sum_{m=1}^{M} \sum_{t=t_{m-1}+1}^{t_m} \sum_{j=1}^{d} \mathbf{E}[|x_{t,j}| p_{j,t}^{\theta^\star}(\mathbf{x}^t)] \geq \frac{c_{\min}\Delta\sqrt{d}}{4} \cdot \sum_{m=1}^{M} \left[(t_m - t_{m-1}) \exp(-2t_{m-1}\Delta^2/d) \right]. \tag{3.4}$$

With the result of Lemma 3.2, we have that (3.3) is lower bounded by

$$(c_{\min}\Delta\sqrt{d}/4) \cdot \sum_{m=1}^{M} \left[(t_m - t_{m-1}) \exp(-2t_{m-1}\Delta^2/d) \right],$$

for any $\Delta \in [0, 1/\sqrt{d}]$. we consider $\Delta \in \{1/\sqrt{d}, \sqrt{d/t_s}, \sqrt{d/t_{s+1}}, \ldots, \sqrt{d/t_M}\}$, where $s = \inf\{m \in [M] : t_m \geq d^2\}$—such s exists due to the assumption $T \geq d^2$. Then the worst-case regret is lower bounded by $\frac{c_{\min}}{4e^2} \cdot \max\{t_s, \frac{dt_{s+1}}{\sqrt{t_s}}, \ldots, \frac{dt_M}{\sqrt{t_{M-1}}}\} \geq c \cdot d\sqrt{T}(T/d^2)^{\frac{1}{2(2^M-1)}}$, QED.

4 Batched UCB for GLCB

In this section, we describe a batched learning algorithm for GLCB. We then highlight the main theoretical results that characterize the performance of the algorithm. The overall idea of the algorithm is that, at the end of every batch, the learner computes an estimate $\hat{\theta}$ of the unknown parameter $\theta^\star$ via ridge regression, based on which it contructs a confidence set that contains $\theta^\star$ with high probability. When entering the new batch, at each time t the learner simply picks the action with the largest value, which is determined by the sum of the estimated reward value (using the current estiamte) and the upper confidence score. Finally, we choose the uniform grid, i.e., $t_m = \lfloor \frac{mT}{M} \rfloor$ for each $m \in [M]$. See Algorithm 1.

Remark 4.1. *There are two important features that distinguish the above algorithm BUCB-GLCB from the standard LinUCB algorithm in Chu et al. (2011). First, as already mentioned before, BUCB-GLCB is run on a grid of batches, and within each batch m, the confidence score is formed using the same sample variance matrix A_{m-1} at the end of the previous batch. Further, as it turns out, the simple uniform grid already attains near-optimal performance, and, from the subsequent analysis, it is unlikely that any other choice of grid can do better. Second, and perhaps more importantly, a particular choice of ridge regularization constant is needed here to achieve the sharp performance characterized later. Specifically, while in the standard online setting (Chu et al., 2011), A_0 can be directly initialized to I_d, therefore corresponding to a ridge regularization constant $\lambda = 1$, in the current adaptive batched learning setting, it is crucial that $\lambda = \tilde{\Theta}(d)$, which comes as a result of two competing components of regret: one from standard online learning, the other from the batches. See Remark 5.1 for a detailed discussion. In particular, as it turns out, using a standard choice of $\lambda = 1$ would yield a regret bound that is a factor of $\sqrt{d}$ worse.*

Remark 4.2. *In the (canonical) generalized linear model, solving equation (4.1) is equivalent to computing*

Algorithm 1 Batched UCB for GLCB (BUCB-GLCB)

Input: time horizon T; feature dimension d; number of batches M; inverse function derivative lower bound κ; regularization tuning parameter $\lambda = 4d\log(1+T/d^2)/\kappa^2$; confidence tuning parameter sequence $\gamma_m = \sqrt{\lambda} + \sqrt{\log T + d\log\left(1 + t_m/(d\lambda)\right)}$.
Grid choice: $\mathcal{T} = \{t_1, \cdots, t_M\}$ with $t_m = \lfloor \frac{mT}{M} \rfloor$.
Initialization:
$A_0 = \lambda I_d \in \mathbf{R}^{d\times d}$; $\hat{\theta}_0 = \mathbf{0} \in \mathbf{R}^d$; $t_0 = 0$.
for $m \leftarrow 1$ to M **do**
  for $t \leftarrow t_{m-1} + 1$ to t_m **do**
    $a_t \leftarrow \arg\max_{a\in\mathcal{A}} \left\{ \langle \phi(x_t, a), \hat{\theta}_{m-1}\rangle + \gamma_{m-1}\|\phi(x_t, a)\|_{A_{m-1}^{-1}} \right\}$
  end for
  Receive rewards in the m-th batch: $\{r_{t,a_t}\}_{t_{m-1}+1\le t\le t_m}$.
  $A_m \leftarrow A_{m-1} + \sum_{t=t_{m-1}+1}^{t_m} \phi(x_t, a_t)\phi(x_t, a_t)^\top$.
  $\hat{\theta}_m$ is the solution to the equation:

$$\sum_{t=1}^{t_m} (r_{t,a_t} - \mu(\langle \phi(x_t, a_t), \theta\rangle))\phi(x_t, a_t) = \kappa\lambda\theta. \tag{4.1}$$

end for

the penalized maximum likelihood estimator (MLE) under independence, and the existence and uniqueness—in fact our analysis does not require uniqueness—is guaranteed when $\lambda > 0$ and the parameter space is closed and convex. In fact, the existence, uniqueness and the computation of a penalized MLE have been well studied in previous works (see e.g., Friedman et al. (2010)).

4.1 Main Upper Bound for Expected Regret

The expected regret achieved by BUCB-GLCB is formalized by the following result.

Theorem 4.1. *Assume the reward is generated from the generalized linear model, the inverse link function μ is monotonically increasing, Lipschitz with Lipschitz constant L_μ and twice differentiable with its derivative μ' satisfies that $\inf_{\|\theta-\theta^\star\|_2\le 2, \|\phi\|_2\le 1} \mu'(\phi^\top\theta) \ge \kappa > 0$ and second derivative $\mu'' \ge 0$. Then we have*

$$\sup_{\|\theta^\star\|_2\le 1} \mathbf{E}_{\theta^\star}\left[R_T(\mathbf{Alg})\right] \le \mathsf{Polylog}(T, d)\frac{L_\mu \max(\kappa, 1)}{\kappa}\left(d\sqrt{T} + \frac{dT}{M}\right),$$

*where **Alg** is the BUCB-GLCB algorithm.*

5 Regret Analysis

In this section we sketch the proof of Theorem 4.1, whereas the complete and detailed proof (and corresponding lemmas) are provided in the supplementary material.

The first step of the the BUCB-GLCB algorithm is to construct confidence intervals for $\theta^\star$ at t_m for all $m \in [M]$, as formalized in Lemma 5.1.

Lemma 5.1. *Under the assumption in Theorem 4.1, with probability at least* $1-\delta$*, for any* $m \in [M]$*,*

$$\|\hat{\theta}_m - \theta^\star\|_{A_m} \le \gamma_m(\delta).$$

where $\gamma_m(\delta) = \sqrt{\lambda} + \sqrt{2\log(1/\delta) + d\log\left(1 + t_m/(d\lambda)\right)}/\kappa$.

To proceed, let E denote the "good" event on which for any $m \in [M]$, $\|\theta^\star - \hat{\theta}_m\|_{A_m} \le \gamma_m(\delta)$. By Lemma 5.1 $\mathbb{P}(E) \ge 1-\delta$. On E, for any $a \in \mathcal{A}$, any $m \in [M]$ and $t_{m-1} < t \le t_m$,

$$\begin{aligned}\langle \phi(x_t,a), \theta^\star\rangle &\overset{(a)}{\le} \langle \phi(x_t,a), \hat{\theta}_{m-1}\rangle + \gamma_{m-1}(\delta)\|\phi(x_t,a)\|_{A_{m-1}^{-1}} \\ &\overset{(b)}{\le} \langle \phi(x_t,a_t), \hat{\theta}_{m-1}\rangle + \gamma_{m-1}(\delta)\|\phi(x_t,a_t)\|_{A_{m-1}^{-1}},\end{aligned}$$

where step (a) is due to Cauchy-Schwarz inequality and step (b) the choice of a_t. By assumption, $\mu(\cdot)$ is monotonically increasing and Lipschitz,

$$\begin{aligned}\max_{a\in\mathcal{A}} \mu(\langle \phi(x_t,a),\theta^\star\rangle) - \mu(\langle \phi(x_t,a_t),\theta^\star\rangle) \le& \max_{a\in\mathcal{A}} L_\mu\left(\langle \phi(x_t,a),\theta^\star\rangle - \langle \phi(x_t,a_t),\theta^\star\rangle\right) \\ \le& L_\mu\left[\langle \phi(x_t,a_t), \hat{\theta}_{m-1} - \theta^\star\rangle + \gamma_{m-1}(\delta)\|\phi(x_t,a_t)\|_{A_{m-1}^{-1}}\right] \\ \overset{(a)}{\le}& L_\mu\left[\|\phi(x_t,a_t)\|_{A_{m-1}^{-1}}\|\hat{\theta}_{m-1} - \theta^\star\|_{A_{m-1}} + \right. \\ & \left.\gamma_{m-1}(\delta)\|\phi(x_t,a_t)\|_{A_{m-1}^{-1}}\right] \le 2L_\mu\gamma_{m-1}(\delta)\|\phi(x_t,a_t)\|_{A_{m-1}^{-1}},\end{aligned}$$

where step (a) follows from Cauchy-Schwarz inequality. Summing over t, we have:

$$\begin{aligned}R_T(\mathbf{Alg}) \overset{(a)}{\le}& 2L_\mu \sum_{m=1}^{M}\left[\gamma_{m-1}(\delta)\sum_{t=t_{m-1}+1}^{t_m}\left(\|\phi(x_t,a_t)\|_{A_{m-1}^{-1}}\right)\right] \\ \le& 2L_\mu\sum_{m=1}^{M}\left[\gamma_{m-1}(\delta)\sqrt{\frac{T}{M}}\sqrt{\sum_{t=t_{m-1}+1}^{t_m}\phi(x_t,a_t)^T A_{m-1}^{-1}\phi(x_t,a_t)}\right] \\ =& 2L_\mu\sum_{m=1}^{M}\left[\gamma_{m-1}(\delta)\sqrt{\frac{T}{M}}\sqrt{\mathsf{Tr}\left(A_{m-1}^{-1}D_m\right)}\right],\end{aligned}$$

where $D_m = A_m - A_{m-1}$; step (a) follows from Cauchy-Schwarz inequality and the choice of the grid. Lemma 5.2 then gives an upped bound on the quantity $\sum_{m=1}^{M}\sqrt{\mathsf{Tr}(A_{m-1}^{-1}D_m)}$.

Lemma 5.2. *For any* $m \in [M]$*,*

$$\sum_{m=1}^{M}\sqrt{\mathsf{Tr}(A_{m-1}^{-1}D_m)} \le \sqrt{10}\left(\sqrt{dM\log\left(\frac{\lambda+T}{\lambda}\right)} + \frac{d\sqrt{T}}{\sqrt{M\lambda}}\log\left(\frac{T+\lambda}{\lambda}\right)\right).$$

Note that γ_m is increasing in m. Then conditional on E,

$$R_T(\mathbf{Alg}) \le 8L_\mu\gamma_M(\delta)\cdot\left(\sqrt{dT\log\left(\frac{\lambda+T}{\lambda}\right)} + \frac{dT}{M\sqrt{\lambda}}\log\left(\frac{\lambda+T}{\lambda}\right)\right). \tag{5.1}$$

Finally we return to the expected regret of BUCB-GLCB,

$$\begin{aligned}&\mathbf{E}_{\theta^\star}[R_T(\mathbf{Alg})] = \mathbf{E}_{\theta^\star}[R_T(\mathbf{Alg})\mathbf{1}_{E^c}] + \mathbf{E}_{\theta^\star}[R_T(\mathbf{Alg})\mathbf{1}_E] \\ \leq & 2\delta T + 8L_\mu \gamma_M(\delta)\left[\sqrt{dT\log\left(\frac{\lambda+T}{\lambda}\right)} + \frac{dT}{M\sqrt{\lambda}}\log\left(\frac{\lambda+T}{\lambda}\right)\right]. \end{aligned} \tag{5.2}$$

Now let $\delta = 1/\sqrt{T}$ and $\lambda = 4d\log(1+T/d^2)/\kappa^2$. Then $\gamma_M \triangleq \gamma_M\left(1/\sqrt{T}\right) \leq \mathsf{Polylog}(T,d)\cdot\sqrt{d}/\kappa$, and consequently, $\mathbf{E}_{\theta^\star}[R_T(\mathbf{Alg})]$ is upper bounded by $\mathsf{Polylog}(T,d)(\max(\kappa,1)L_\mu/\kappa)\left(d\sqrt{T}+dT/M\right)$, completing the proof.

Remark 5.1. *In (5.2), the choice $\lambda = \tilde{\Theta}(d)$ is optimal. If the order of λ is less than d, then $\gamma_M = \tilde{\Theta}(d)$; the order of the upper bound is then $\tilde{\Theta}(d\sqrt{T} + (dT/M)\cdot\sqrt{d/\lambda})$, which is suboptimal. On the other hand, if the order of λ is greater than d, then $\gamma_M = \tilde{\Theta}(\lambda)$. The resulting upper bound is $\tilde{\Theta}(\sqrt{\lambda dT} + dT/M)$, which is again, suboptimal.*

As a direct by-product from the proof of Theorem 4.1, we present a high probability regret upper bound next.

Theorem 5.3. *Under the assumption in Theorem 4.1, for any $\theta^\star \in \mathbb{R}^d$ s.t. $\|\theta^\star\|_2 \leq 1$, with probability at least $1-\delta$, the regret incurred by adaptive batched UCB (Algorithm 1) is upper bounded by*

$$R_T(\boldsymbol{Alg}) \leq \mathsf{Polylog}(T,d)\frac{\max(\kappa,1)}{\kappa}L_\mu\left(d\sqrt{T}+\frac{dT}{M}\right), \tag{5.3}$$

Proof. By Equation (5.1), conditional on the "good" event E, with $\lambda = 4d\log(1+T/d^2)/\kappa$,

$$R_T(\mathbf{Alg}) \leq \mathsf{Polylog}(T,d)L_\mu\frac{\max(\kappa,1)}{\kappa}\left(d\sqrt{T}+dT/M\right).$$

By Lemma 5.1, $\mathbb{P}(E) \geq 1-\delta$, and we complete the proof. □

6 Conclusion

Our work provides a near-complete characterization of adaptive batched learning in generalized linear contextual bandits with general decision sets. In the future, it would be interesting to see whether one can further tighten the additional factor of $O(\max(\sqrt{d}, d/\sqrt{M}))$ in the regret bound.

7 Acknowledgment

The authors would like to thank Yanjun Han for helpful discussions, and the reviewers for the insightful comments that greatly improve the paper.

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

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. (2017). Mostly exploration-free algorithms for contextual bandits. *arXiv preprint arXiv:1704.09011*.

Bertsimas, D. and Mersereau, A. J. (2007). A learning approach for interactive marketing to a customer segment. *Operations Research*, 55(6):1120–1135.

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. and Lugosi, G. (2012). Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422.

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.

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.

Combes, R., Shahi, M. S. T. M., Proutiere, A., et al. (2015). Combinatorial bandits revisited. In *Advances in Neural Information Processing Systems*, pages 2116–2124.

Dudík, M., Langford, J., and Li, L. (2011). Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1097–1104.

Durand, A. and Gagné, C. (2014). Thompson sampling for combinatorial bandits and its application to online feature selection. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*.

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.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1.

Gao, Z., Han, Y., Ren, Z., and Zhou, Z. (2019). Batched multi-armed bandits problem. In *Advances in Neural Information Processing Systems*, pages 501–511.

Goldenshluger, A. and Zeevi, A. (2013). A linear response bandit problem. *Stochastic Systems*, 3(1):230–261.

Hamidi, N. and Bayati, M. (2020). On worst-case regret of linear thompson sampling. *arXiv preprint arXiv:2006.06790*.

Han, Y., Zhou, Z., Zhou, Z., Blanchet, J., Glynn, P. W., and Ye, Y. (2020). Sequential batch learning in finite-action linear contextual bandits. *arXiv preprint arXiv:2004.06321*.

Joachims, T., Swaminathan, A., and Rijke, M. d. (2018). Deep learning with logged bandit feedback. In *International Conference on Learning Representations*.

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.

Kallus, N. (2018). Balanced policy evaluation and learning. *Advances in Neural Information Processing Systems*.

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

Lattimore, T. and Szepesvári, C. (2018). Bandit algorithms. *preprint*, page 28.

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.

Perchet, V., Rigollet, P., Chassang, S., Snowberg, E., et al. (2016). Batched bandit problems. *The Annals of Statistics*, 44(2):660–681.

Rakhlin, A. and Sridharan, K. (2016). BISTRO: An efficient relaxation-based method for contextual bandits. In *Proceedings of the International Conference on Machine Learning*, pages 1977–1985.

Rigollet, P. and Zeevi, A. (2010). Nonparametric bandits with covariates. *arXiv preprint arXiv:1003.1630*.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535.

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.

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.

Slivkins, A. et al. (2019). Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286.

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.

Tsybakov, A. (2008). *Introduction to Nonparametric Estimation*. Springer-Verlag.

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M., and Laber, E. (2012). Estimating optimal treatment regimes from a classification perspective. *Stat*, 1(1):103–114.

Zhao, Y.-Q., Zeng, D., Laber, E. B., Song, R., Yuan, M., and Kosorok, M. R. (2014). Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102(1):151–168.

Supplementary material

S.1 Auxiliary results

Lemma S.1 (Tsybakov (2008), Chapter 2)**.** *Let P and Q be probability measures on the same measurable space $(\Omega, \mathcal{F})$ and let $A \in \mathcal{F}$ be an arbitrary event. Then,*

$$P(A) + Q(A^c) \geq \frac{1}{2}\exp(-D_{\mathrm{KL}}(P, Q))$$

where $A^c = \Omega \backslash A$ is the complement of A.

S.2 Proofs for the lower bound

S.2.1 Proof of Lemma 3.2

We prove the argument by an averaging argument. For any $\theta \in \{-1/\sqrt{d}, 1/\sqrt{d}\}^d$ and $j \in [d]$, define $h_j(\theta)$ as a d-dimensional vector such that $h_j(\theta) = (\theta_1, \ldots, -\theta_j, \ldots, \theta_d)$, and let $[h_j(\theta)]_k$ denote the k_{th} coordinate of $h_j(\theta)$. Averaging over all $\theta \in \{-1/\sqrt{d}, 1/\sqrt{d}\}^d$, one has

$$\frac{1}{2^d}\sum_{\theta\in\{-\frac{1}{\sqrt{d}},\frac{1}{\sqrt{d}}\}^d}\left[\frac{\Delta}{\sqrt{d}}\sum_{m=1}^{M}\sum_{t=t_{m-1}+1}^{t_m}\sum_{j=1}^{d}\mathbf{E}\left[|x_{t,j}|p_{j,t}^{\theta}(\mathbf{x}^t)\right]\right] \geq \frac{\Delta}{2^d\sqrt{d}}\sum_{m=1}^{M}\sum_{t=t_{m-1}+1}^{t_m}\sum_{j=1}^{d}\sum_{\theta:\theta_j=\frac{1}{\sqrt{d}}}\left[\mathbf{E}\left[|x_{t,j}|\left(p_{j,t}^{\theta}(\mathbf{x}^t) + p_{j,t}^{h_j(\theta)}(\mathbf{x}^t)\right)\right]\right]$$

$$\overset{(a)}{\geq} \frac{\Delta}{2^d\sqrt{d}}\sum_{m=1}^{M}\sum_{t=t_{m-1}+1}^{t_m}\sum_{j=1}^{d}\sum_{\theta:\theta_j=\frac{1}{\sqrt{d}}}\left[\frac{\mathbf{E}[|x_{t,j}|]}{2}\exp\left(-\frac{2t_{m-1}\Delta^2}{d}\right)\right] \geq \frac{c_{\min}\Delta\sqrt{d}}{4}\sum_{m=1}^{M}(t_m - t_{m-1})\exp\left(-\frac{2t_{m-1}\Delta^2}{d}\right),$$

where step (a) is because

$$\begin{aligned}
p_{j,t}^{\theta}(\mathbf{x}^t) + p_{j,t}^{h_j(\theta)}(\mathbf{x}^t) &= \mathbb{P}_{\theta}\left(\mathrm{sgn}(x_{tj}\theta_j) \neq \mathrm{sgn}(a_{t,j})|\mathbf{x}^t\right) + \mathbb{P}_{h_j(\theta)}\left(\mathrm{sgn}(x_{tj}[h_j(\theta)]_j) \neq \mathrm{sgn}(a_{t,j})|\mathbf{x}^t\right) \\
&= \mathbb{P}_{\theta}\left(\mathrm{sgn}(x_{tj}\theta_j) \neq \mathrm{sgn}(a_{t,j})|\mathbf{x}^t\right) + \mathbb{P}_{h_j(\theta)}\left(\mathrm{sgn}(x_{tj}\theta_j) = \mathrm{sgn}(a_{t,j})|\mathbf{x}^t\right) \\
&\overset{(a)}{\geq} \frac{1}{2}\exp\left(-D_{\mathrm{KL}}\left(\mathbb{P}_{\theta,\mathbf{x}}^t, \mathbb{P}_{h_j(\theta),\mathbf{x}}^t\right)\right),
\end{aligned}$$

where $\mathbb{P}_{\theta,\mathbf{x}}^t$ (resp. $\mathbb{P}_{h_j(\theta),\mathbf{x}}^t$) denotes the law of rewards observable up to time t with parameter θ (resp. $h_j(\theta)$) conditional on $\mathbf{x}^t$, and step (a) is a result of Lemma S.1. We now focus on upper bounding the KL divergence. We here similarly define $\mathbf{E}_{\theta,\mathbf{x}}^t$ (resp. $\mathbf{E}_{h_j(\theta),\mathbf{x}}^t$) to be the expectation taken w.r.t the law of the observable rewards up to time t with parameter θ (resp. $h_j(\theta)$) conditional on $\mathbf{x}^t$. Using the chain rule of the KL divergence,

$$D_{\mathrm{KL}}(\mathbb{P}_{\theta,\mathbf{x}}^t, \mathbb{P}_{\bar{\theta},\mathbf{x}}^t) \overset{(a)}{=} \sum_{\tau=1}^{t_{m-1}}\mathbf{E}_{\theta,\mathbf{x}}^{\tau}\left[\frac{1}{2}\langle\theta - h_j(\theta), \phi(x_\tau, a_\tau)\rangle^2\right] = \frac{2}{d}\sum_{\tau=1}^{t_{m-1}}\mathbf{E}_{\theta,\mathbf{x}}^{\tau}[(x_{\tau,j}a_{\tau,j})^2] \leq \frac{2\Delta^2}{d}\sum_{\tau=1}^{t_{m-1}}x_{\tau,j}^2 \leq \frac{2t_{m-1}\Delta^2}{d},$$

where step (a) is because up to time t, the learner can only observe rewards up to time t_{m-1}. Combining the above, we claim that there exists $\theta \in \{-1/\sqrt{d}, 1/\sqrt{d}\}^d$, such that

$$\frac{\Delta}{\sqrt{d}} \sum_{m=1}^{M} \sum_{t=t_{m-1}+1}^{t_m} \sum_{j=1}^{d} \mathbf{E}\left[|x_{t,j}| p_{j,t}^{\theta}(\mathbf{x}^t)\right] \geq \frac{c_{\min}\Delta\sqrt{d}}{4} \sum_{m=1}^{M} (t_m - t_{m-1}) \exp\left(-\frac{2t_{m-1}\Delta^2}{d}\right),$$

since otherwise the average would be less than the right-hand side of the above inequality.

S.3 Proofs for upper bound

S.3.1 Proof of Lemma 5.1

For any $t \in [T]$, define $V_t = \sum_{\tau=1}^{t} \phi(x_\tau, a_\tau)\phi(x_\tau, a_\tau)^\top$, $S_t = \sum_{\tau=1}^{t} \xi_\tau \phi(x_\tau, a_\tau)$; and for each $m \in [M]$ we define $A_m = \lambda I + V_{t_m}$. For any $z \in \mathbb{R}^d$, define $M_t(z) = \exp\left(\langle z, S_t\rangle - \frac{1}{2}\|z\|_{V_t}^2\right)$, with $M_0(z) = 1$. Let $\{\mathcal{F}^t\}_{0\leq t\leq T}$ denote a filtration of the past history, where for any $t \geq 0$,

$$\mathcal{F}^t = \sigma(x_1, a_1, r_{1,a_1}, \ldots, x_t, a_t, r_{t,a_t}, x_{t+1}, a_{t+1}).$$

By definition, $M_t(z)$ is adapted to the filtration $\{\mathcal{F}^t\}_{t\in[T]}$; additionally,

$$\begin{aligned}
\mathbf{E}\left[M_t \mid \mathcal{F}^{t-1}\right] =& \mathbf{E}\left[\exp\left(\langle z, S_t\rangle - \frac{1}{2}\|z\|_{V_t}^2\right) \mid \mathcal{F}^{t-1}\right] \\
=& \exp\left(\langle z, S_{t-1}\rangle - \frac{1}{2}\|z\|_{V_t}^2\right) \mathbf{E}\left[\exp\left(\langle z, \xi_t\phi(x_t, a_t)\rangle\right) \mid \mathcal{F}^{t-1}\right] \\
\overset{(a)}{\leq}& \exp\left(\langle z, S_{t-1}\rangle - \frac{1}{2}\|z\|_{V_t}^2\right) \exp\left(\frac{1}{2}\langle z, \phi(x_t, a_t)\rangle^2\right) \leq M_{t-1}(z),
\end{aligned}$$

where step (a) is due to the sub-gaussianity of ξ_t. Consequently, $M_t(z)$ is a supermartingale w.r.t. $\{\mathcal{F}^t\}_{0\leq t\leq T}$. Based on $M_t(x)$, we further define $\bar{M}_t = \int M_t(z)dh(z)$, where $h(\cdot)$ corresponds to the law of $\mathcal{N}(0, \frac{1}{\lambda}I)$. We claim that $\bar{M}_t$ as well is a supermartingale w.r.t. $\{\mathcal{F}^t\}_{0\leq t\leq T}$: it is first straightforward that $\bar{M}_t$ is adapted to $\{\mathcal{F}^t\}_{0\leq t\leq T}$; the conditional expectation

$$\mathbf{E}[\bar{M}_t \mid \mathcal{F}^{t-1}] = \mathbf{E}\left[\int M_t(z)dh(z) \mid \mathcal{F}^{t-1}\right] = \int \mathbf{E}\left[M_t(z) \mid \mathcal{F}^{t-1}\right] dh(z) \leq \int M_{t-1}(z)dh(z) = \bar{M}_{t-1}.$$

Hence $\bar{M}_t$ is a supermartingale w.r.t. $\{\mathcal{F}^t\}_{0\leq t\leq T}$. Invoking the maximal inequality, one has

$$\mathbb{P}\left(\sup_{t\geq 0} \bar{M}_t \geq \frac{1}{\delta}\right) \leq \delta \mathbf{E}\bar{M}_0 = \delta.$$

On the other hand, $\bar{M}_t$ can be computed as,

$$\begin{aligned}
\bar{M}_t =& \int M_t(z) \frac{1}{\sqrt{\det(\lambda^{-1}I)(2\pi)^d}} \exp(-\lambda\|z\|_2^2/2)dz = \int \exp\left(\langle z, S_t\rangle - \frac{1}{2}z^T V_t z\right) \left(\frac{\lambda}{2\pi}\right)^{d/2} \exp(-\lambda\|z\|_2^2/2)dz \\
=& \frac{\lambda^{\frac{d}{2}}}{\sqrt{\det(\lambda I + V_t)}} \exp\left(\frac{1}{2}\|S_t\|_{(\lambda I+Vt)^{-1}}^2\right).
\end{aligned}$$

Consequently, with probability at least $1-\delta$, for any $t \in [T]$,

$$\frac{\lambda^{\frac{d}{2}}}{\sqrt{\det(\lambda I + V_t)}} \exp\left(\frac{1}{2}\|S_t\|^2_{(\lambda I + Vt)^{-1}}\right) \leq \frac{1}{\delta}.$$

Rearranging yields $\|S_t\|^2_{(\lambda I+V_t)^{-1}} \leq 2\log(1/\delta) + \log\left(\det(\lambda I + V_t)/\lambda^d\right)$. We now focus on the grids $\{t_m\}_{m\in[M]}$. For $m \in [M]$, define $G_m(\theta) = \sum_{t=1}^{t_m} \left[\mu(\langle \phi(x_t, a_t), \theta\rangle) - \mu(\langle \phi(x_t, a_t), \theta^\star\rangle)\right] \cdot \phi(x_t, a_t)$. Then $G_m(\hat{\theta}_m) + \kappa\lambda\hat{\theta}_m = S_{t_m}$ and $G_m(\theta^\star) = 0$. By the mean-value theorem, there exists $0 < \alpha < 1$ and $\tilde{\theta} = \alpha\hat{\theta}_m + (1-\alpha)\theta^\star$, such that

$$G_m(\hat{\theta}_m) - G_m(\theta^\star) = \left[\sum_{t=1}^{t_m} \mu'(\langle \phi(x_t, a_t), \tilde{\theta}\rangle) \cdot \phi(x_t, a_t)\phi(x_t, a_t)^\top\right](\hat{\theta}_m - \theta^\star).$$

Note that

$$\begin{aligned}\|S_{t_m} - \kappa\lambda\theta^\star\|_{A_m^{-1}} =& \|G_m(\hat{\theta}_m) + \kappa\lambda\hat{\theta}_m - G_m(\theta^\star) - \kappa\lambda\theta^\star\|_{A_m^{-1}} \\ =& \left\|\left[\sum_{t=1}^{t_m} \mu'(\langle \phi(x_t, a_t), \tilde{\theta}\rangle)\phi(x_t, a_t)\phi(x_t, a_t)^T + \kappa\lambda I\right](\hat{\theta}_m - \theta^\star)\right\|_{A_m^{-1}}.\end{aligned}$$

A consequence of the above inequality is that

$$\inf_{\theta:\|\theta-\theta^\star\|_2=2} \|G_m(\theta) + \kappa\lambda\theta - G_m(\theta^\star) - \kappa\lambda\theta^\star\|_{A_m^{-1}} \geq 2\kappa\sqrt{\lambda},$$

where we use the fact that $\sum_{t=1}^{t_m}\left[\mu'(\langle \phi(x_t, a_t), \tilde{\theta}\rangle) \cdot \phi(x_t, a_t)\phi(x_t, a_t)^\top + \kappa\lambda I\right] \succeq \kappa A_m$ since $\|\tilde{\theta} - \theta^\star\|_2 \leq 2$. By Lemma A of Chen et al. (1999), we have that

$$\left\{\theta : \|G_m(\theta) + \kappa\lambda\theta - G_m(\theta^\star) - \kappa\lambda\theta^\star\|_{A_m^{-1}} \leq 2\kappa\sqrt{\lambda}\right\} \subset \{\theta : \|\theta - \theta^\star\| \leq 2\}.$$

On the other hand,

$$\|S_{t_m} - \kappa\lambda\theta^\star\|_{A_m^{-1}} \overset{(a)}{\leq} \|S_{t_m}\|_{A_m^{-1}} + \kappa\lambda\|\theta^\star\|_{A_m^{-1}} \leq \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(A_m)}{\lambda^d}\right)} + \kappa\lambda\|\theta^\star\|_{A_m^{-1}},$$

where step (a) follows from the triangular inequality. Since the eigenvalues of A_m are lower bounded by λ and $\|\theta^\star\|_2 \leq 1$,

$$\|\theta^\star\|_{A_m^{-1}} \leq \sqrt{(\theta^\star)^\top A_m^{-1}\theta^\star} \leq \frac{\|\theta^\star\|_2}{\sqrt{\lambda}} \leq \frac{1}{\sqrt{\lambda}}.$$

Meanwhile,

$$\det(A_m) = \det\left(\lambda I + \sum_{t=1}^{t_m} \phi(x_t, a_t)\phi(x_t, a_t)^\top\right) \overset{(a)}{\leq} \left(\frac{\mathsf{Tr}\left(\lambda I + \sum_{t=1}^{t_m} \phi(x_t, a_t)\phi(x_t, a_t)^\top\right)}{d}\right)^d \overset{(b)}{\leq} \left(\frac{\lambda d + t_m}{d}\right)^d,$$

where the step (a) uses the arithmetic mean-geometric mean inequality and step (b) uses the fact that $\mathsf{Tr}(\phi(x_t,a_t)\phi(x_t,a_t)^\top) = \|\phi(x_t,a_t)\|_2^2 \le 1$. Consequently,

$$\|S_{t_m} - \kappa\lambda\theta^\star\|_{A_m^{-1}} \le \sqrt{2\log(1/\delta) + d\log(1+t_m/(\lambda d))} + \kappa\sqrt{\lambda}.$$

Taking $\lambda = 4d\log(1+T/d^2)/\kappa^2$, we have that $\|S_{t_m} - \kappa\lambda\theta^\star\|_{A_m^{-1}} \le 2\kappa\sqrt{\lambda}$, and thus $\|\hat{\theta}_m - \theta^\star\|_2 \le 2$. By assumption, $\mu'\left(\langle\phi(x_t,a_t),\tilde{\theta}\rangle\right) \ge \kappa$, and we have

$$\|\hat{\theta}_m - \theta^\star\|_{A_m} \le \sqrt{2\log(1/\delta) + d\log\left(1+t_m/(\lambda d)\right)}/\kappa + \sqrt{\lambda}.$$

S.3.2 Proof of Lemma 5.2

Part of the proof is based on the technique in Han et al. (2020). To start, using Eqaution (2) in Han et al. (2020), we have

$$\sum_{m=1}^{M}\sqrt{\mathsf{Tr}(A_{m-1}^{-1}D_m)} \le \sum_{m=1}^{M}\sqrt{10\sum_{j=1}^{d}\frac{\nu_{m,j}-\nu_{m-1,j}}{\nu_{m-1,j}}},$$

where $\nu_{m,j}$ is the j_{th} eigenvalue of A_m arranged in a way such that $\nu_{m,j} \ge \nu_{m-1,j}$ and $\nu_{0,j} = \lambda$. Note that

$$\begin{aligned}\sum_{m=1}^{M}\sqrt{10\sum_{j=1}^{d}\frac{\nu_{m,j}-\nu_{m-1,j}}{\nu_{m,j}}} &\le \sqrt{10M}\sqrt{\sum_{j=1}^{d}\sum_{m=1}^{M}\frac{\nu_{m,j}-\nu_{m-1,j}}{\nu_{m,j}}} \le \sqrt{10M}\sqrt{\sum_{j=1}^{d}\int_{\nu_{0,j}}^{\nu_{M,j}}\frac{1}{x}dx} \\ &= \sqrt{10M\sum_{j=1}^{d}\log\left(\frac{\nu_{M,j}}{\nu_{0,j}}\right)} \le \sqrt{10Md\log\left(\frac{T+\lambda}{\lambda}\right)}. \end{aligned} \tag{S.1}$$

The last inequality is because $\nu_{0,j} = \lambda$ and for any $z \in \mathbb{R}^d$ such that $\|z\|_2 = 1$,

$$z^\top A_M z = z^\top\left(\lambda I + \sum_{t=1}^{T}\phi(x_t,a_t)\phi(x_t,a_t)^\top\right)z = \lambda\|z\|_2^2 + \sum_{t=1}^{T}\left(\phi(x_t,a_t)^\top z\right)^2 \le \lambda\|z\|_2^2 + T\|z\|_2^2.$$

Finally we bound the difference between $\sum_{m=1}^{M}\sqrt{10\sum_{j=1}^{d}\frac{\nu_{m,j}-\nu_{m-1,j}}{\nu_{m,j}}}$ and $\sum_{m=1}^{M}\sqrt{10\sum_{j=1}^{d}\frac{\nu_{m,j}-\nu_{m-1,j}}{\nu_{m-1,j}}}$.

$$\begin{aligned}&\sum_{m=1}^{M}\sqrt{10\sum_{j=1}^{d}\frac{\nu_{m,j}-\nu_{m-1,j}}{\nu_{m-1,j}}} - \sum_{m=1}^{M}\sqrt{10\sum_{j=1}^{d}\frac{\nu_{m,j}-\nu_{m-1,j}}{\nu_{m,j}}} \le \sqrt{10}\sum_{m=1}^{M}\frac{\sum_{j=1}^{d}\frac{\nu_{m,j}-\nu_{m-1,j}}{\nu_{m-1,j}} - \frac{\nu_{m,j}-\nu_{m-1,j}}{\nu_{m,j}}}{\sqrt{\sum_{j=1}^{d}\frac{\nu_{m,j}-\nu_{m-1,j}}{\nu_{m-1,j}}}} \\ =&\sqrt{10}\sum_{m=1}^{M}\frac{\sum_{j=1}^{d}\frac{(\nu_{m,j}-\nu_{m-1,j})^2}{\nu_{m,j}\nu_{m-1,j}}}{\sqrt{\sum_{j=1}^{d}\frac{\nu_{m,j}-\nu_{m-1,j}}{\nu_{m-1,j}}}} \overset{(a)}{\le} \sqrt{10}\sum_{m=1}^{M}\frac{\sqrt{\sum_{j=1}^{d}\frac{\nu_{m,j}-\nu_{m-1,j}}{\nu_{m-1,j}}}\sqrt{\sum_{j=1}^{d}\frac{(\nu_{m,j}-\nu_{m-1,j})^3}{\nu_{m,j}^2\nu_{m-1,j}}}}{\sqrt{\sum_{j=1}^{d}\frac{\nu_{m,j}-\nu_{m-1,j}}{\nu_{m-1,j}}}} \\ =&\sqrt{10}\sum_{m=1}^{M}\sqrt{\sum_{j=1}^{d}\frac{(\nu_{m,j}-\nu_{m-1,j})^3}{\nu_{m,j}^2\nu_{m-1,j}}}, \end{aligned} \tag{S.2}$$

where step (a) is by the Cauchy-Schwarz inequality. Also note that $\nu_{m-1,j} \geq \lambda$ and $\nu_{m,j} - \nu_{m-1,j} \leq \mathsf{Tr}(D_m) \leq t_m - t_{m-1} = T/M$. Then

$$\begin{aligned}(\text{S.2}) \leq & \sqrt{\frac{10T}{M\lambda}} \sum_{m=1}^{M} \sqrt{\sum_{j=1}^{d} \frac{(\nu_{m,j} - \nu_{m-1,j})^2}{\nu_{m,j}^2}} \leq \sqrt{\frac{10T}{M\lambda}} \sum_{m=1}^{M} \sum_{j=1}^{d} \frac{(\nu_{m,j} - \nu_{m-1,j})}{\nu_{m,j}} \\ \leq & \sqrt{\frac{10T}{M\lambda}} d \log\left(\frac{T+\lambda}{\lambda}\right), \end{aligned} \tag{S.3}$$

where the second inequality is because $\sum_{i=1}^{n} a_i^2 \leq (\sum_{i=1}^{n} a_i)^2$ for $a_i \geq 0$. Combining (S.1) and (S.3) completes the proof.