High-Dimensional Structured Bandits: Minimax Regret, Estimation and Inference

Sai Li [*1], Elynn Y. Chen [† 2], Tony T. Cai [‡3], and Michael I. Jordan [§ 4]

[1,3]University of Pennsylvania.
[2,4]University of California, Berkeley.

December 5, 2020

Abstract

In this paper, we consider the structured bandits problem with high-dimensionality. The structure of the environment class of a bandits problem enables the learner to obtain information about some actions while never playing them. Exploration is necessary in most bandits settings to achieve minimax regret. However, we show that, in the structured bandits setting, the proposed *exploration-free* algorithm can achieve minimax regret as long as the the *margin* condition is satisfied. To facilitate decision making under uncertainty and high-dimensionality, we further propose an online inference algorithm that is able to correct the biases incurred by using penalty for the high-dimensional problems.

1 Introduction

The multi-armed bandit (MAB) (Robbins, 1952; Lai and Robbins, 1985) problem is prototypical in the area of data-driven sequential making in an unknown environment. In an unstructured environment where playing action a_i cannot inform anything about the distributions of action $a_j \neq a_i$, a decision maker needs to address the dilemma between "exploration" (to generate information about the unknown environment parameters needed for

[*]Corresponding author. Email: sai.li@pennmedicine.upenn.edu
[†]Equal contribution. Supported in part by NSF Grants DMS-1803241. Email: elynn.chen@berkeley.edu
[‡]Email: tcai@wharton.upenn.edu
[§]Email: jordan@cs.berkeley.edu

efficient decision making) and "exploitation" (to set the action that attempt to maximize the expected rewards from the system outputs) (Lai and Robbins, 1985). However, exploration may be prohibitively costly or infeasible in a variety of practical applications Bird et al. (2016). For example, in marketing applications, testing out an inappropriate ad on a potential customer may result in the costly, permanent loss of the customer; in dynamic pricing, incorrect pricing may result in permanent closure of a business; in medical decision making, it may be unethical to choose a treatment on a specific patient merely for the purpose of exploration.

At the same time, practical applications commonly possess intrinsic structures: the distribution class of different arms are not independent and can be predicted from observable covariates. Such applications include the mobile health system Tewari and Murphy (2017), advertisement/news article placement (Langford and Zhang, 2007; Li et al., 2010), revenue management (Ferreira et al., 2018), dynamic pricing (Javanmard and Nazerzadeh, 2019a; Javanmard et al., 2020), marketing (Schwartz et al., 2017) and recommendation systems. Relaxing the requirement that the environment class is a product set makes structured bandit problems, an important feature of which is that the learner can often obtain information about some actions while never playing them (Lattimore and Szepesvári, 2020). Thus a natural question is whether exploration is needed to achieve minimax regret in the structured bandits.

In this paper, we characterize a broad class of structured bandits problem where an exploration-free algorithm can be designed to achieve minimax regrets. At each time t, the decision maker observes an context feature $\boldsymbol{x}_t \in \mathbb{R}^d$. Upon choosing action a_t, she observes a random response y_t with mean $f\left(a_t, \boldsymbol{\beta}^\top \boldsymbol{x}_t\right)$. The realized reward at time t is a known function $r\left(a_t, y_t\right)$. The dynamic pricing, stochastic linear and non-linear contextual bandits are shown to be special instances of this class of structured bandit problems.

For this general class of structured bandits, We derive the minimax lower bound of the regret, allowing high-dimensional context feature $\boldsymbol{x}$. An exploration-free algorithm, namely *Phased LASSO Bandit*, is proposed to simultaneously estimate the unknown environment parameters and set the action to maximize the expected rewards. We show that the upper regret bound of the Phased LASSO Bandit achieves minimax optimality in the general setting and that our bounds are tighter then those in the literature when applied to the special instances of dynamic pricing and stochastic linear bandits.

To facilitate decision making under uncertainty, we further introduce an online inference algorithm for the estimated parameters and predicted rewards obtained by the Phased LASSO Bandit, which fills the blank of the statistical inference in high-dimensional bandits. The inference in high-dimensional bandits is delicate because a direct application of results

from low-dimensional setting inevitably introduce bias and thus is incorrect. We adapt the debias technique to the online bandits and obtain a tighter confidence interval for both estimated parameters and predicted rewards. The efficacy of the proposed algorithm and the validity of theoretical results are further confirmed with empirical studies on synthetic and real data.

Our contributions are from several folds. First, we characterize a general class of structured bandit problems where exploration-free algorithms are possible to achieve the minimax regret. We establish the minimax lower regret bound for such a general class. Second, we design the Phased LASSO Bandit algorithm that is exploration-free and achieves the minimax lower bound without enforced exploration. When applied to the special instances of dynamic pricing and stochastic linear bandits, the Phased LASSO Bandit algorithm is shown to have tighter bounds than those in the related work. Third, all the related work in dynamic decision focuses on the analysis of regrets. But none of them studied the statistical inference of the parameter estimator or the predicted rewards, which is an important problem for online decisions in medicine, business or finance. As Tewari and Murphy (2017) discuss in their survey that work should be done to assess the "uncertainty of the decisions". Our novel online inference algorithm fills this blank in the high-dimensional decision making.

1.1 Related works

The bandit problem has been extensively studied in the computer science, operation, and statistics literature. See Bubeck et al. (2012); Lattimore and Szepesvári (2020) and references therein for an informative review. The most relevant literature to us are high dimensional stochastic contextual bandits, dynamic pricing and statistical inference in bandits.

Contextual bandits incorporate additional contextual information that may help predict the quality of the actions (Auer, 2002; Dani et al., 2008; Li et al., 2010; Chu et al., 2011). Under the *adversarial* setting, Abbasi-Yadkori et al. (2011) proved an upper bound of $\mathcal{O}\left(d\sqrt{T}\right)$ regret after T time periods when contexts are d-dimensional. We consider the *stochastic* setting where rewards are generated stochastically from some unknown distribution. The stochastic setting is well suited to many applications, such as news recommendation and clinical trials on treatments for a noninfectious disease and allows for improved regret bounds in T from $\sqrt{T}$ to $\log(T)$. For low dimensional context features, Goldenshluger and Zeevi (2013) derive an upper bound of $\mathcal{O}\left(d^3\log(T)\right)$ regret. Clearly, the exponential dependence on the covariate dimension d makes the algorithm unfavorable in high dimensions.

For high dimensional setting where the number of relevant covariates is $s \ll d$, Bastani and Bayati (2020) design a LASSO bandit that is proven to achieve $\mathcal{O}\left(s^2\left(\log(T)+\log(d)\right)^2\right)$

when the margin condition is satisfied. Their algorithm is of exploration-then-commit type and needs an enforced exploration stage of $\mathcal{O}\left(s^2 \log(d) \log(T)\right)$ rounds. Kim and Paik (2019) designed a doubly-robust LASSO bandit with uniform exploration stage of $\mathcal{O}\left(\sqrt{T \log(dT) \log(T)}\right)$ rounds and with regret upper bound at $\mathcal{O}\left(s \log(dT)\sqrt{T}\right)$ All the aforementioned algorithms require the knowledge of the sparsity index s since it is needed to determine the length of the enforced exploration. Oh et al. (2020) propose a Sparse-Agnostic LASSO bandit that does not require any uniform exploration stage and thus does not require the knowledge of s. They show that the regret upper bound is $\mathcal{O}\left(s^2 \log(d) + s\sqrt{T \log(dT)}\right)$ under the relaxed symmetry assumption on the context distribution. Their regret bound can be improved to $\mathcal{O}\left(\sqrt{sT \log(dT)}\right)$ with an additional restricted eigenvalue condition. Ariu et al. (2020) propose a Threshold LASSO bandit with regret scales at most as $\mathcal{O}\left(\log(d) + \sqrt{T \log(T)}\right)$ in general and $\mathcal{O}\left(\log(d) + \log(T)\right)$ under the additional margin condition.

Our regret bound in Theorem 4.1 applies to a broader class of structured bandits. When applied to the special instance of high-dimensional stochastic linear bandits, our algorithm is shown to incur regret upper bounded at $\mathcal{O}\left(s\sqrt{\log(d)T}\right)$ in general and $\mathcal{O}\left(s \log(d) \log(T)\right)$ under the additional margin condition, which are better than the aforementioned upper bounds. In addition, we establish the minimax lower bound of the regret for the class of structured bandits and thus establish the optimality of the proposed Phased LASSO Bandit algorithm.

All the contextual bandit literature mentioned above focuses on the analysis of regret. But none of them studied the asymptotic distribution of the parameter estimators or the reward predictors. Recently, Chen et al. (2020a,b) establish the asymptotic normality of the parameter estimator produced by ϵ-greedy algorithm. They focus on the statistical inference of online decision making and do not offer the upper or lower bounds on the regret. Besides, they do not deal with the high-dimensional setting.

2 Stochastic Structured Bandits

A *stochastic contextual bandit* is a collection of distributions $\nu = (\mathcal{P}_a : a \in \mathcal{A})$ where $\mathcal{A}$ is the set of available actions. The decision maker (or "agent" for brevity) and the environment interact sequentially over T rounds. In each round $t \in \{1, \cdots, T\}$, the agent observe environment covariates $\boldsymbol{x}_t$ and sets an action $a_t \in \mathcal{A}$, which is fed to the environment. The environment then samples a response $y_t \in \mathbb{R}$ from the distribution $\mathcal{P}_{a_t}$ and reveals y_t to the agent. The reward at time t is given by a known function $r\left(a_t, y_t\right)$. The interaction between the agent and the environment induce a history of observed covariates, actions and responses

before period t, that is $\mathcal{H}_{t-1} \equiv \{(\boldsymbol{x}_s, a_s, y_s), 1 \leq s \leq t-1\}$.

We characterize a class of *structured bandits* as follows. Given the covariate $\boldsymbol{x}_t$ and the action a_t, the response y_t and the reward r_t at time t are given, respectively, by

$$\begin{aligned} y_t &= f\left(\boldsymbol{a}_t, \boldsymbol{x}_t^\top \boldsymbol{\beta}\right) + \varepsilon_t, \\ r_t &= r_0\left(\boldsymbol{a}_t, \mathbb{E}[y_t|\boldsymbol{a}_t, \boldsymbol{x}_t]\right) = r\left(\boldsymbol{a}_t, \boldsymbol{x}_t^\top \boldsymbol{\beta}\right). \end{aligned} \tag{1}$$

where ε_t is the random noise and $\mathbb{E}[\varepsilon_t|x_t, \mathcal{H}_{t-1}] = 0$.

We first show that the dynamic pricing problem, stochastic contextual bandits are special instances of the structured bandits defined by (1).

Dynamic pricing Dynamic pricing (den Boer and Zwart, 2014) is a class of dynamic decision problems where a seller sequentially sets prices to maximize revenue while simultaneously learning the unknown demand function. At each time period $1 \leq t \leq T$, the agent observes environment covariates $\boldsymbol{x}_t$ consisting of product features, market size, macroeconomic indices, seasonality, or geographic indicators and the determines an action, i.e. selling price $a_t \in [a_l, a_h]$. After setting the price , the seller observes a realization y_t of the random demand and collects revenue $r(a_t, y_t) = a_t \cdot y_t$. The model corresponds to a special case of the structured bandits (1) by setting

$$\begin{aligned} y_t &= f\left(\boldsymbol{x}^\top \boldsymbol{\beta} - a\right) + \varepsilon_t, \\ r_t &= a_t \times y_t. \end{aligned} \tag{2}$$

where ε_t is the random noise and $\mathbb{E}[\varepsilon_t|x_t, \mathcal{H}_{t-1}] = 0$. The demand function $f(\cdot)$ can be a linear or logit function, corresponding to various random demands, such as Normal, Poisson, or Bernoulli distributed demands. The seller's expected single-period revenue function is

$$r(a, \boldsymbol{x}) = a \times f(\boldsymbol{x}^\top \boldsymbol{\beta} - a). \tag{3}$$

The optimal pricing policy that maximizes the expected single-period revenue is defined by function

$$a^*(\boldsymbol{\beta}, \boldsymbol{x}) = \arg \max_{a \in [a_l, a_u]} a \times f(\boldsymbol{x}^\top \boldsymbol{\beta} - a) \tag{4}$$

Stochastic contextual bandits At the beginning of period t, the agent observes contextual information $\boldsymbol{X}_t = \{(\boldsymbol{x}_t)_1, \ldots, (\boldsymbol{x}_t)_K\} \in \mathbb{R}^{K \times d}$. Having observe the context, the agent chooses their action $a_t \in \{1, \cdots, K\}$ based on the information available. The response and

reward function are defined by the same function

$$f\left(\boldsymbol{a}_t,\ \boldsymbol{X}_t^\top \boldsymbol{\beta}\right) = r\left(\boldsymbol{a}_t,\ \boldsymbol{X}_t^\top \boldsymbol{\beta}\right) = \boldsymbol{a}_t^\top \boldsymbol{X}_t^\top \boldsymbol{\beta}, \tag{5}$$

where the action chosen is represented as $\boldsymbol{a}_t \in \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_K\}$ and $\{\boldsymbol{e}_k\}_{1\le k\le K}$ is the unit standard basis in $\mathbb{R}^K$.

Remark 1. *A common stochastic linear bandits setup (Lattimore and Szepesvári, 2020) assumes that the learner observes a context $C_t \in \mathcal{C}$ and the reward satisfies*

$$r\left(C_t, \boldsymbol{a}_t\right) = \left\langle \boldsymbol{\beta}, \boldsymbol{\phi}(C_t, \boldsymbol{a}_t)\right\rangle, \quad \forall\ (C_t, \boldsymbol{a}_t) \in \mathcal{C} \times [K],$$

where $\boldsymbol{\phi}(\cdot)$ is a feature map from $\mathcal{C} \times [K] \mapsto \mathbb{R}^d$. Formulation (5) explicitly rewrites $\boldsymbol{\phi}(C_t, a_t) = \boldsymbol{X}_t \boldsymbol{a}_t$ where $\boldsymbol{X}_t = \left[\boldsymbol{\phi}(C_t, 1)^\top, \cdots, \boldsymbol{\phi}(C_t, K)^\top\right]^\top$.

Remark 2. *In the literature of contextual linear bandits, there are two different definitions. Our definition (5) is equivalent to one common setting where $\boldsymbol{\beta}$ is the same across different actions (Lattimore and Szepesvári, 2020; Auer, 2002; Dani et al., 2008; Li et al., 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011). The other setting consider finite arm set $\mathcal{A} = \{1, \cdots, K\}$ and treats the unknown parameter defining the reward function as arm-specific (Bastani and Bayati, 2020; Chen et al., 2020a). Specifically, the reward of choosing an action $a_t = k$ is $r_t = \boldsymbol{x}_t^\top \boldsymbol{\beta}_k + \varepsilon_t$ with $\boldsymbol{\beta}_k$ being different for different actions. This setting is unstructured in that playing an action k_1 does not provide any information on $\boldsymbol{\beta}_{k_2}$ for a different action $k_2 \neq k_1$. In fact, arm-specific coefficients setting is no easier than the multi-armed bandits (MAB) in that the latter only needs to estimate K different means, while the former needs to estimate K different regressions. We will show that this setting with arm-specific coefficients does not satisfy the essential condition for existence of an exploration-free algorithm that can achieve minimax regret.*

Remark 3. *Certain forms of games (Salant and Cherry, 2020) and linear quadratic regulator (Dean et al., 2018) with one stage can also be cast into the structured bandits defined by (1).*

The goal of data-driven decision making is to construct a sequential policy π to maximize the expected total reward while learning the environment parameters $\boldsymbol{\beta}$ that are essential for decision making. The performance of π is measured by its *cumulative regret* or *cumulative expected regret*, which is the standard metric in the analysis of bandit algorithms (Lai and Robbins, 1985; Auer, 2002). In particular, we compare ourselves to the oracle optimal policy π^* which is obtained with the knowledge of the true values of $\boldsymbol{\beta}$. Let

$a_t^* = \arg\max_{\boldsymbol{a}\in\mathcal{A}} r\left(\boldsymbol{a}, \boldsymbol{x}_t^\top \boldsymbol{\beta}\right)$ be the optimal action given the true model. The *cumulative regret* of a policy $\pi = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_T)$ is defined as

$$\mathrm{R}_\pi(T) = \sum_{t=1}^{\top} (r_t^* - r_t) = \sum_{t=1}^{\top} \left(r\left(a_t^*, \boldsymbol{x}_t^\top \boldsymbol{\beta}\right) - r\left(a_t, \boldsymbol{x}_t^\top \boldsymbol{\beta}\right)\right).$$

The *cumulative expected regret* is defined as $\mathbb{E}_{\boldsymbol{x}_t \sim \mathcal{P}_X}[\mathrm{R}_\pi(T)]$.

2.1 Conditions for the existence of exploration-free algorithms

We now define a property of the observed response function $f(\cdot)$ that determines whether an exploration is necessary for achieving minimax regrets.

Definition 2.1 (Uniformly informative)**.** *We say that the actions are* uniformly informative *to $\boldsymbol{\beta}$ with respect to $f(\cdot)$ if, for any $\boldsymbol{a}_t \in \mathcal{A}$,*

$$\frac{1}{n}\sum_{t=1}^{n} (\boldsymbol{x}_t^\top \boldsymbol{u})\{f(\boldsymbol{a}_t, \boldsymbol{x}_t^\top(\boldsymbol{\beta}+\boldsymbol{u})) - f(\boldsymbol{a}_t, \boldsymbol{x}_t^\top \boldsymbol{\beta})\} \geq c\|\boldsymbol{u}\|_2^2 \quad \text{for } \|\boldsymbol{u}\|_1 \leq 1 \tag{6}$$

with probability at least $1 - \exp(-n)$. That is, whichever action is taken, the variation in the response y will provide information on the unknown parameter $\boldsymbol{\beta}$.

The condition (6) is essentially the restricted strong convexity condition (RSC) of f on $\boldsymbol{\beta}$. If f has a generalized linear model (GLM) representation, then (6) holds with mild conditions on $\boldsymbol{x}_t$. In Section 4 we will show that (6) holds for the dynamic pricing model and K-arm contextual bandits under mild conditions on $\boldsymbol{x}_t$.

To see the key role that (6) plays in requiring exploration, we list some examples for which (6) *does not* hold.

Example 1. Bandits with arm-specific reward functions (Bastani and Bayati, 2020; Chen et al., 2020a).

$$f(\boldsymbol{a}_t, \boldsymbol{x}_t^\top \boldsymbol{\beta}) = \boldsymbol{x}_t^\top \boldsymbol{B} \boldsymbol{a}_t,$$

where $\boldsymbol{B} \in \mathbb{R}^{p\times K}$ and $\boldsymbol{a}_t \in \{e_1, \ldots, e_K\}$. Hence, $\boldsymbol{\beta} = vec(\boldsymbol{B})$ in this example. Here, different arms have distinct parameters and information of k-th arm can be obtained only when $a_t = e_k$. Hence, condition 6 is not fulfilled.

Example 2. K-products dynamic pricing with different sensitivities.

$$\mathbb{E}[y_t|\boldsymbol{x}_t] = f((\boldsymbol{a}_t^\top, \boldsymbol{x}_t^\top)\boldsymbol{\beta}),$$

where $\boldsymbol{\beta} \in \mathbb{R}^{K+p}$. The parameter $\boldsymbol{\beta}_{1:K}$ denote the sensitivity to prices for different products. Revenue is $r((\boldsymbol{a}_t^\top, \boldsymbol{x}_t^\top)\boldsymbol{\beta})$.

The purpose of the this example is to show that even if different actions share same parameter, exploration may still be needed. This is because inference for the unknown parameter $\boldsymbol{\beta}$ depends on sufficient variation in $\boldsymbol{a}_t$. If an algorithm keeps choosing the same action, then the estimated $\boldsymbol{\theta}$ has poor accuracy which will affect the estimation of reward.

In Section 4 we will show that (6) holds for the dynamic pricing model and K-arm contextual bandits under mild conditions on $\boldsymbol{x}_t$. Knowing the sufficient condition, we are confident to present an exploration free algorithm with optimality in regret.

3 An exploration-free algorithm

The proposed algorithm is a phase-based greedy method. The updates in the model parameter estimation only occurs at the beginning of each phase, with using only the samples collected in the previous episode. Therefore, the actions chosen in each phase are independent from the model noises in that episode. This independence is not a mere serendipity, rather it holds because of the specific design of the algorithm. Using this property, we could decouple the dependence between the action and the noises, which leads to tighter concentration bounds. The lengths of each phase in our algorithm increase geometrically ($\tau_k = 2^k$), allowing for more accurate estimate of as the phase index grows. The algorithm terminates

at the end of the horizon (period T), but note that it does not need to know T in advance.

Algorithm 1: Master Algorithm

Input : (at $t = 0$) Functions $f(\cdot)$ and $r(\cdot)$

Input : (arrives over time) $\{\boldsymbol{x}_t\}_{t\geq 0}$

Output: Action a_t and confidence interval for a_t^* and f_t

Initialize $\tau_0 \leftarrow 1$, $\boldsymbol{a}_0 \leftarrow a_{\min}$, $\boldsymbol{\beta}_0 \leftarrow 0$.

for *each episode* $k = 1, 2, 3, \cdots$ **do**

Set the length of k-th episode: $\tau_k \leftarrow 2^k$.

At $t = \tau_k + 1$, update

$$\widehat{\boldsymbol{\beta}}_e = \arg\min_{\beta\in\mathbb{R}^d} \|\boldsymbol{\beta}\|_1 \quad \text{subject to} \left\| \frac{1}{\tau_e} \sum_{t=\tau_{e-1}+1}^{\tau_e} \boldsymbol{x}_t^\top (y_t - f(\boldsymbol{a}_t, \boldsymbol{x}_t^\top \boldsymbol{\beta})) \right\|_\infty \leq \lambda_e \tag{7}$$

for $t = \tau_e + 1, \ldots, \tau_{e+1} - 1$ **do**

$$\boldsymbol{a}_t = \arg\max_{\boldsymbol{a}\in\mathcal{A}} r(\boldsymbol{a}, \boldsymbol{x}_t^\top \widehat{\boldsymbol{\beta}}_e).$$

end

end

4 Regret analysis

Condition 4.1 (sub-Gaussian errors and bounded contexts)**.** *ε_t are independent sub-Gaussian with mean zero and variance σ^2. The contexts $\sup_{t\leq T} \|\boldsymbol{x}_t\|_\infty \leq C < \infty$*

Condition 4.2 (Restricted strong convexity)**.** *f satisfies (6) for a bounded support $\mathcal{A}$.*

We now introduce a Lipschitz condition on the reward function. Let

$$M_e(\kappa, S) = \sup_{u:\|u\|_2\leq\kappa, u\in\mathcal{C}(3,S)} \mathbb{E}_n[r(\boldsymbol{a}_t^*, \boldsymbol{x}_t^\top\boldsymbol{\beta}) - r(\boldsymbol{a}_t(\boldsymbol{\beta}+u), \boldsymbol{x}_t^\top\boldsymbol{\beta})].$$

Condition 4.3 (Lipschitz condition on reward function)**.** *For any $a \in \mathcal{A}$, $r(\boldsymbol{a}, \boldsymbol{x}^\top\boldsymbol{\beta}) \leq R < \infty$. For any $n \geq s \log d$,*

$$\mathbb{E}||M_e(\kappa, S)|] \leq L_p \kappa^\alpha \quad (0 \leq \alpha \leq 2).$$

Condition 4.3 says that $\mathbb{E}_n[r(\boldsymbol{a}_t^*(u), \boldsymbol{x}_t^\top\boldsymbol{\beta})]$ is α-order Lipschitz in u around $u = 0$ in

expectation.

Theorem 4.1 (Regret analysis). *Under Conditions 4.1, 4.2, and 4.3,*

$$Reg_{\widehat{\pi}}(T) \leq c_1 \begin{cases} Rs\log d + L_d(s\log d)\log T & \text{for } \alpha = 2 \\ Rs\log d + L_d(s\log d)^{\alpha/2}T^{1-\alpha/2} & \text{for } \alpha \in [0,2). \end{cases}$$

4.1 Corollary for dynamic pricing

a linear model (Mills, 1959; Petruzzi and Dada, 1999) and a logit model (Chen and Simchi-Levi, 2012).

Corollary 4.1. *Assume Condition 4.1 and $\mathcal{A} = [c_*, c^*]$ for bounded positive constants c_* and c^*. For some constant $C_1 > 0$,*

$$Reg_{\widehat{\pi}}(T) \leq C_1 s\log d\log T.$$

4.2 Corollary for contextual bandits

Margin condition:

$$\max_{k\neq k^*} \mathbb{P}\left(|\Delta_{t,k}| \leq \rho\right) < c\rho \quad \forall \rho \leq c_0 \leq 1. \tag{8}$$

Corollary 4.2 (Lasso contextual bandits with the margin condition). *Assume Condition 4.1 and (8). For some constant $C_1 > 0$,*

$$Reg_{\widehat{\pi}}(T) \leq C_1 s\log d\log T.$$

Corollary 4.3 (Lasso contextual bandits without the margin condition). *Assume Condition 4.1. For some constant $C_1 > 0$,*

$$Reg_{\widehat{\pi}}(T) \leq C_1 s\sqrt{(\log K + \log d)T}.$$

4.3 Minimax optimality

We provide a lower bound on the achievable regret by any policy in the structured bandits characterized by (1).

Consider the parameter space

$$\Omega = \{\|\beta\|_0 \leq s, c_1 I \preceq \ \Sigma \preceq c_2 I\}.$$

Theorem 4.2 (Lower bound on the regret)**.** *Assume that for any* $\boldsymbol{\beta} \in \Omega$*,*

$$|r(\boldsymbol{a}_t^*, \boldsymbol{x}_t^\top \boldsymbol{\beta}) - r(\boldsymbol{a}_t(\boldsymbol{\beta}'), \boldsymbol{x}_t^\top \boldsymbol{\beta})| \geq \min\{L_d \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_2^2, R\}.$$

Then it holds that

$$\inf_{\pi} \sup_{\beta} Reg(T) \geq \min\{Cs \log d \log T, RT\}.$$

5 Online Inference

In applications, it is important to provide uncertainty quantifications in our estimates at each time t. In dynamic pricing, for example, $f(\boldsymbol{a}_t, \boldsymbol{x}_t^\top \boldsymbol{\beta})$ is the expected demand at time t given the price $\boldsymbol{a}_t$. Providing confidence intervals for the expected demand is crucial for practical needs. In the following, we propose an inference procedures for the true conditional mean function $f(\boldsymbol{a}_t, \boldsymbol{x}_t^\top \boldsymbol{\beta})$ and the reward function $r(\boldsymbol{a}_t, \boldsymbol{x}_t^\top \boldsymbol{\beta})$.

The pipeline is to first construct an asymptotically normal estimator of $\boldsymbol{x}_t^\top \boldsymbol{\beta}$. Second, we construct confidence intervals for $f(\boldsymbol{a}_t, \boldsymbol{x}_t^\top \boldsymbol{\beta})$ using the confidence intervals for $\boldsymbol{x}_t^\top \boldsymbol{\beta}$. Although f is injective, the second step is complicated because $\boldsymbol{a}_t$ is a random quantity estimated based on the previous episode. Hence, conditioning on $\boldsymbol{x}_t$, our inference target

$f(\boldsymbol{a}_t, \boldsymbol{x}_t^\top \boldsymbol{\beta})$ is random.

Algorithm 2: Online inference in episode e when $\tau_e \lesssim p$

for $t = \tau_e + 1, \ldots, \tau_{e+1}$ **do**

Estimate the debiased score $\widehat{\boldsymbol{w}}_t$ via

$$\widehat{\boldsymbol{w}}_t = \arg\min_{\boldsymbol{w} \in \mathbb{R}^d} \|\boldsymbol{w}\|_1 \quad \text{subject to} \quad \begin{cases} \left\| \widehat{\Sigma}_{\widehat{\beta},e-1} \boldsymbol{w}_t - \boldsymbol{x}_t \right\|_\infty \le \|\boldsymbol{x}_t\|_2 \lambda_e \\ \left| \boldsymbol{x}_t^T \widehat{\Sigma}_{\widehat{\beta},e-1} \boldsymbol{w}_t - \|\boldsymbol{x}_t\|_2^2 \right| \le \|\boldsymbol{x}_t\|_2^2 \lambda_k \\ \max_{\tau_{e-1} < j \le \tau_e} |\boldsymbol{x}_j^T \boldsymbol{w}| \le C \|\boldsymbol{x}_j\|_2 \sqrt{\log \tau_{e-1}} \end{cases} \tag{9}$$

for $\lambda_e \asymp \sqrt{2 \log d / \tau_{e-1}}$ and $\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}},e-1} = \frac{1}{\tau_{e-1}} \sum_{j=\tau_{e-1}+1}^{\tau_e} \boldsymbol{x}_j \boldsymbol{x}_j^T \dot{f}(\boldsymbol{a}_j, \boldsymbol{x}_j^T \widehat{\boldsymbol{\beta}}_{e-1})$.

Debiased estimate of $\mu_t = \boldsymbol{x}_t^\top \boldsymbol{\beta}$ based on $\widehat{\boldsymbol{w}}_t$:

$$\widehat{\mu}_t = \boldsymbol{x}_t^\top \widehat{\boldsymbol{\beta}}_e + \frac{\sum_{j=\tau_{e-1}+1}^{\tau_e} \left\langle \boldsymbol{x}_j^\top \widehat{\boldsymbol{w}}_t, y_j - f(\boldsymbol{a}_j, \boldsymbol{x}_j^T \widehat{\boldsymbol{\beta}}_e) \right\rangle}{\tau_{e-1}}. \tag{10}$$

CI for μ_t:

$$\widehat{\mu}_t \pm z_{1-\alpha/2} \sqrt{\frac{\widehat{V}_t}{\tau_{e-1}}} \tag{11}$$

where $\widehat{V}_t$ is the estimated variance of $\widehat{\mu}_t$. Specifically,

$$\widehat{V}_t = \frac{1}{\tau_{e-1}} \sum_{j=\tau_{e-1}+1}^{\tau_e} (\boldsymbol{x}_j^\top \widehat{\boldsymbol{w}}_t)^2 \{y_j - f(\boldsymbol{a}_j, \boldsymbol{x}_j^T \widehat{\boldsymbol{\beta}}_{e-1})\}^2.$$

end

Theorem 5.1 (Asymptotic normality for Algorithm 2)**.** *Assume Conditions 4.1 and 4.2. If* $s \log d = o(\tau_k)$*, then for any fixed* $t \ge \tau_k$*,*

$$\sqrt{\frac{\tau_{e-1}}{V_t}} (\widehat{\mu}_t - \boldsymbol{x}_t^T \boldsymbol{\beta}) = Z_t + O_P\left(\frac{\sqrt{\log \tau_{e-1}} s \log d}{\tau_{e-1}} \right).$$

where $Z_t \xrightarrow{D} N(0,1)$ *and*

$$V_t = \widehat{\boldsymbol{w}}_t^T \widehat{\Sigma}_{\boldsymbol{\beta},e-1} \widehat{\boldsymbol{w}}_t. \tag{12}$$

We avoid the sample splitting by using $\widehat{\boldsymbol{\beta}}_{e-1}$ in (9) rather than the current $\widehat{\boldsymbol{\beta}}_e$.

6 Numerical Studies

In this section, we study the performance of the proposed algorithm using simulated data from dynamic pricing and stochastic contextual bandits.

6.1 Dynamic pricing

We evaluate the performance of our method in a dynamic pricing settings where the noisy demand is generated according to a logit model (Chen and Simchi-Levi, 2012). The generative model of market data is specified as follows. The demand y_t is generated such that $y_t = f\left(\boldsymbol{x}_t^\top \boldsymbol{\beta} - a_t\right) + \varepsilon_t$ where $f(\cdot)$ is the logit function and $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon)$ with $\sigma_\varepsilon = 0.5$. Covariates $\boldsymbol{x}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ with $d = 100$. We set the sparsity index as $\|\boldsymbol{\beta}\|_0 \leq s$, $s = 10$. Specifically, $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_{10}, 0, \cdots, 0]$, $\beta_i \sim \mathcal{N}(\mu_\beta, \sigma_\beta)$ is randomly generated with $\mu_\beta = 5$ and $\sigma_\beta = 0.5$. The revenue r_t is generated according to (2).

Figure 1 presents the average regrets and the ratios of total regrets over $\log(T)$ and $\sqrt{T}$ for the dynamic pricing with the logit demand.

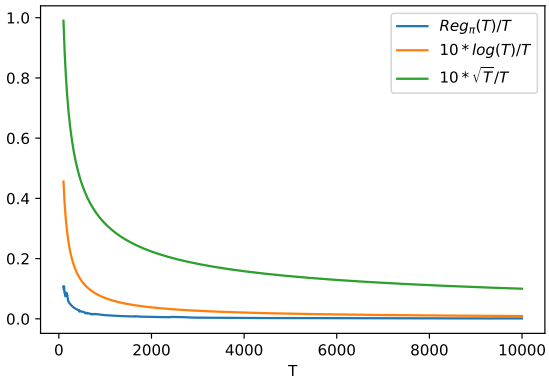
6.2 Stochastic linear contextual bandits

We now present the the numerical results for stochastic linear contextual bandits. Under this setting the response y_t is the same as the reward r_t. The action set $\mathcal{A} = [K]$ with $K = 10$. We generate covariates $(\boldsymbol{x}_t)_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ with $d = 100$ for $k \in [10]$. We set the sparsity index as $\|\boldsymbol{\beta}\|_0 \leq s$, $s = 10$. Model coefficients are generated as $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_{10}, 0, \cdots, 0]$, $\beta_i \sim \mathcal{N}(\mu_\beta, \sigma_\beta)$ is randomly generated with $\mu_\beta = 5$ and $\sigma_\beta = 0.5$. The reward y_t associated with action a_t is generated as $y_t = (\boldsymbol{x}_t)_{a_t}^\top \boldsymbol{\beta} + \varepsilon_t$ where noise $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon)$ with $\sigma_\varepsilon = 0.5$.

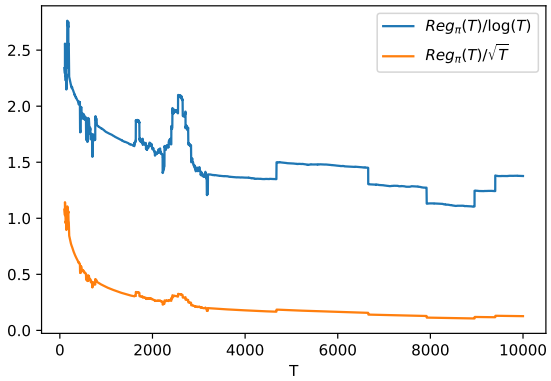
Figure 1 plots the average regrets and the ratios of total regrets over $\log(T)$ and $\sqrt{T}$ for stochastic linear contextual bandits.

7 Summary

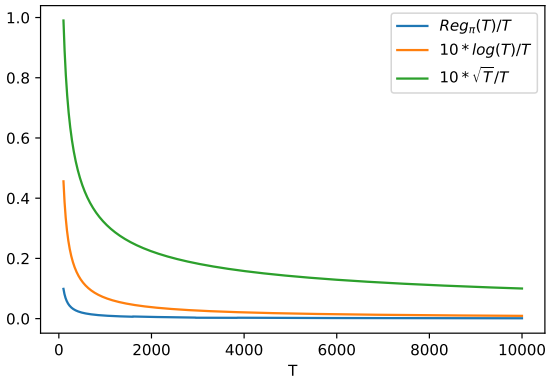
In this paper, we consider the structured bandits problem with high-dimensionality. The structure of the environment class of a bandits problem enables the learner to obtain information about some actions while never playing them. Exploration is necessary in most bandits settings to achieve minimax regret. However, we show that, in the structured bandits setting, the proposed *exploration-free* algorithm can achieve minimax regret as long as the the *margin* condition is satisfied. To facilitate decision making under uncertainty and



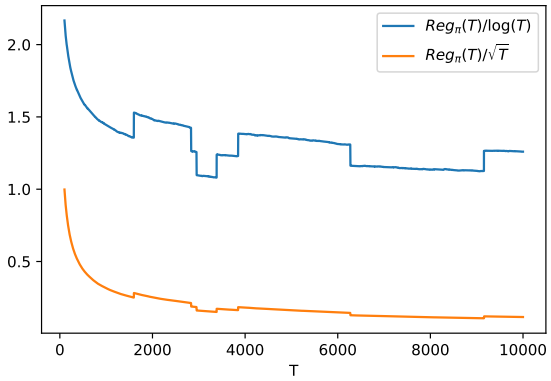
(a) Average regrets, sparse setting $p = 100$, $s = 10$.



(b) Ratio of total regrets over $\log(T)$ and $\sqrt{T}$, the sparse setting $p = 100$, $s = 10$.

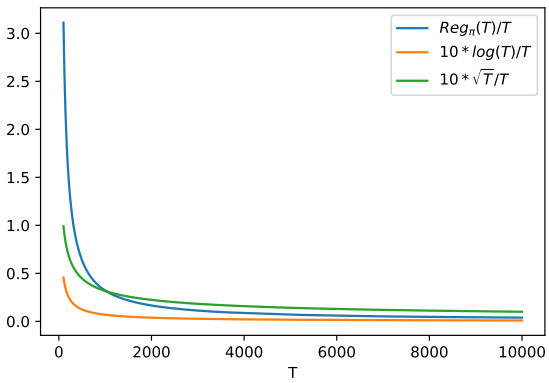


(c) Average regrets, non-sparse setting $p = 10$.

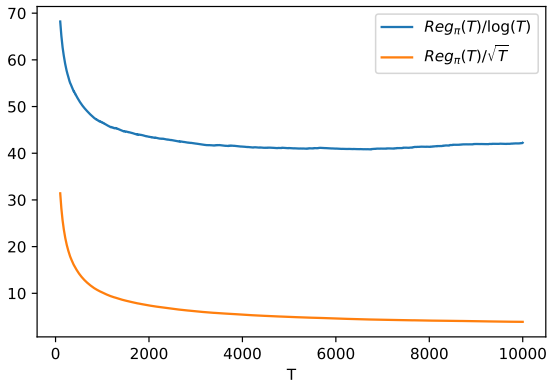


(d) Ratio of total regrets over $\log(T)$ and $\sqrt{T}$, non-sparse setting $p = 10$.

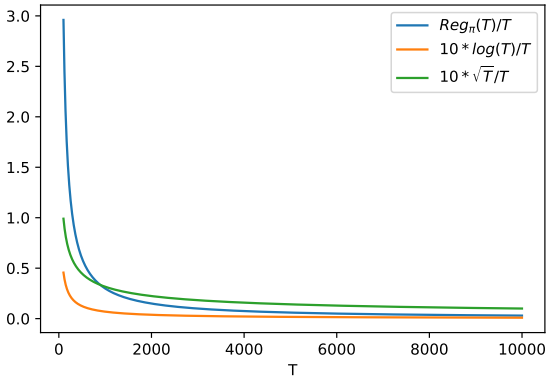
Figure 1: Regrets in dynamic pricing with logit demand.



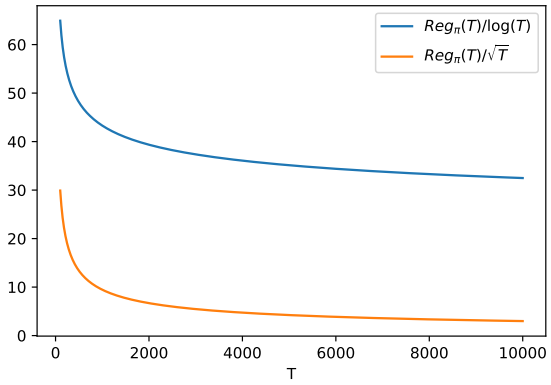
(a) Average regrets, sparse setting $p = 100$, $s = 10$.



(b) Ratio of total regrets over $\log(T)$ and $\sqrt{T}$, the sparse setting $p = 100$, $s = 10$.



(c) Average regrets, non-sparse setting $p = 10$.



(d) Ratio of total regrets over $\log(T)$ and $\sqrt{T}$, non-sparse setting $p = 10$.

Figure 2: Regrets in linear contextual bandits setting.

high-dimensionality, we further propose an online inference algorithm that is able to correct the biases incurred by using penalty for the high-dimensional problems.

8 Proof

Lemma 8.1. *Under Conditions 4.1 -4.3, if $c_1 s \log d \le \tau_{e-1} \le d^2/c_1$, then for any constant $\alpha \le 2$,*

$$\sum_{t \in e\text{-}th\ episode} \mathbb{E}[reg_t] \le L_d (s \log d)^{\alpha/2} \tau_e^{\alpha/2}.$$

Proof of Lemma 8.1. Let

$$\mathcal{E}_{1,e} = \left\{ \text{RSC holds for } \widehat{\Sigma}_e,\ \left\| \frac{1}{\tau_{e-1}} \sum_{t \in e\text{-th episode}} \boldsymbol{x}_t \varepsilon_t \right\|_\infty \le \lambda_e/2 \right\}.$$

$$\sum_{t \in e\text{-th episode}} \mathbb{E}[reg_t] = \underbrace{\mathbb{E}[\sum_{t \in e\text{-th episode}} reg_t \mathbb{1}(\mathcal{E}_{1,e})]}_{R_{1,e}} + \underbrace{\mathbb{E}[\sum_{t \in e\text{-th episode}} reg_t \mathbb{1}(\mathcal{E}_{1,e}^c)]}_{R_{2,e}}.$$

For $R_{1,e}$, standard analysis of the Lasso implies that

$$\mathcal{E}_{1,e} \subseteq \left\{ \widehat{\boldsymbol{\beta}}_e - \boldsymbol{\beta} \in \mathcal{C}(3, S),\ \|\widehat{\boldsymbol{\beta}}_e - \boldsymbol{\beta}\|_2 \le C \sqrt{\frac{s \log d}{\tau_{e-1}}} \right\}. \tag{13}$$

Hence, we can use Condition 4.3 in event $\mathcal{E}_{1,e}$.

$$\begin{aligned} R_{1,e} &\le \mathbb{E}[| \sum_{t \in e\text{-th episode}} r(a_t^*, \boldsymbol{x}_t^T \boldsymbol{\beta}) - r(a_t(\widehat{\boldsymbol{\beta}}_e), \boldsymbol{x}_t^T \boldsymbol{\beta})| \mathbb{1}(\mathcal{E}_{1,e})] \\ &\le C L_d \tau_e (\frac{s \log p}{\tau_e})^{\alpha/2} \le C L_d (s \log d)^{\alpha/2} \tau_e^{\alpha/2}, \end{aligned}$$

where the second last step is by the Lipschitz condition in Condition 4.3.

For $R_{2,e}$, under Condition 4.3, $reg_t \le 2R$ and hence

$$R_{2,e} \le 2R \tau_e \mathbb{P}(\mathcal{E}_{1,e}^c \cup \mathcal{E}_e^c).$$

Using the sub-Gaussian property of $\boldsymbol{x}_t$, we have

$$\mathbb{P}(\mathcal{E}_{1,e}^c) \le \exp\{-c_1 \tau_e\} + \exp\{-c_2 \log d\}.$$

As $\tau_e \ge c_1 s \log d \ge c_1 \log d$, we have $\mathbb{P}(\mathcal{E}_{1,e}^c) \ge 2 \exp\{-c_3 \log d\}$. Hence,

$$R_{2,e} \le 2R(\exp\{-c_1 \tau_e + \log \tau_e\} + \exp\{-c_2 \log d + \log \tau_e\}) \le C,$$

where the last step is due to $\tau_e \to \infty$ and $\tau_e \le p^2$. □

Lemma 8.2. *Under Conditions 4.1- 4.3, if $d^2/c_1 \le \tau_{e-1}$, then for any constant $\alpha \le 2$,*

$$\sum_{t \in e\text{-}th\ episode} \mathbb{E}[reg_t] \le c_1 L_d (s \log d)^{\alpha/2} (\tau_e)^{1-\alpha/2}.$$

Proof of Lemma 8.2. we define

$$\mathcal{E}_{2,e} = \left\{ \Lambda_{\min}(\widehat{\Sigma}_e) \ge c_1 \right\} \text{ and } \mathcal{E}_{3,e} = \left\{ \left\| \frac{1}{\tau_{e-1}} \sum_{t \in e\text{-th episode}} \boldsymbol{x}_t \varepsilon_t \right\|_\infty \le \lambda_e/2 \right\}.$$

We know that $\mathbb{P}(\mathcal{E}_{3,e}) \ge 1 - \exp(-c_2 \tau_{e-1})$ for $\tau_{e-1} \ge p^2$.

$$\begin{aligned} &\mathrm{Reg}_{\widehat{\pi}}(e\text{-th episode}) \\ &= \sum_{t=\tau_e+1}^{\tau_{e+1}} \mathbb{E}[\mathrm{reg}_t \mathbb{1}(\mathcal{E}_{2,e} \cap E_{3,e})] + \sum_{t=\tau_e+1}^{\tau_{e+1}} \mathbb{E}[\mathrm{reg}_t \mathbb{1}(\mathcal{E}_{2,e} \cap \mathcal{E}_{3,e}^c)] + \sum_{t=\tau_e+1}^{\tau_{e+1}} \mathbb{E}[\mathrm{reg}_t \mathbb{1}(\mathcal{E}_{2,e}^c)]. \end{aligned} \tag{14}$$

For the first term in (14), we use similar arguments as for $R_{1,e}$ in the proof of Lemma 8.1. As a result,

$$\sum_{t=\tau_e+1}^{\tau_{e+1}} \mathbb{E}[\mathrm{reg}_t \mathbb{1}(\mathcal{E}_{2,e} \cap E_{3,e})] \le c_1 \tau_e^{1-2\alpha} (s \log d)^{\alpha/2}.$$

For the last term in (14), we have

$$\sum_{t=\tau_e}^{\tau_{e+1}-1} \mathbb{E}[\mathrm{reg}_t \mathbb{1}(\mathcal{E}_{2,e}^c)] \le \tau_e R \mathbb{P}(\mathcal{E}_{2,e}) \le R \tau_e \exp(-c_2 \tau_{e-1}) \le c_1.$$

For the second term, we use a different analysis of the Lasso. Let $\gamma_e = \left\| \frac{1}{\tau_{e-1}} \sum_{t=\tau_{e-1}+1}^{\tau_e} \boldsymbol{x}_t \varepsilon_t \right\|_\infty$. Specifically, we have in $\mathcal{E}_{2,e} \cap \mathcal{E}_{3,e}^c$,

$$\|\widehat{\boldsymbol{\beta}}_e - \boldsymbol{\beta}\|_2^2 \le (\gamma_e + \lambda_e) \|\widehat{\boldsymbol{\beta}}_e - \boldsymbol{\beta}\|_1 \le (\gamma_e + \lambda_e) \sqrt{p} \|\widehat{\boldsymbol{\beta}}_e - \boldsymbol{\beta}\|_2,$$

which gives $\|\widehat{\boldsymbol{\beta}}_e - \boldsymbol{\beta}\|_2^2 \le c_1 p\gamma_e^2$. Therefore,

$$\begin{aligned}
&\sum_{t=\tau_e+1}^{\tau_{e+1}} \mathbb{E}[\mathrm{reg}_t \mathbb{1}(\mathcal{E}_{2,e} \cap \mathcal{E}_{3,e}^c)] = \sum_{t=\tau_e+1}^{\tau_{e+1}} \int_0^\infty \mathbb{P}(\mathrm{reg}_t \mathbb{1}(\mathcal{E}_{2,e} \cap \mathcal{E}_{3,e}^c) \ge \alpha) d\alpha \\
&\le \sum_{t=\tau_e+1}^{\tau_{e+1}} \int_0^\infty \mathbb{P}(\mathrm{reg}_t \ge \alpha, \mathcal{E}_{2,e} \cap \mathcal{E}_{3,e}^c) d\alpha \le \sum_{t=\tau_e+1}^{\tau_{e+1}} \int_0^\infty \mathbb{P}(c_1 d\gamma_e^2 \ge \alpha, \mathcal{E}_{2,e} \cap \mathcal{E}_{3,e}^c) d\alpha \\
&\le \sum_{t=\tau_e+1}^{\tau_{e+1}} \int_0^{L_0} \mathbb{P}(\mathcal{E}_{2,e} \cap \mathcal{E}_{3,e}^c) d\alpha + \sum_{t=\tau_e+1}^{\tau_{e+1}} \int_{L_0}^\infty \mathbb{P}(c_1 d\gamma_k^2 \ge \alpha, \mathcal{E}_{2,e} \cap \mathcal{E}_{3,e}^c) d\alpha \\
&\le L_0 \sum_{t=\tau_e+1}^{\tau_{e+1}-1} \mathbb{P}(\mathcal{E}_{3,e}^c) + \sum_{t=\tau_e+1}^{\tau_{e+1}} \int_{L_0}^\infty \mathbb{P}(c_1 d\gamma_e^2 \ge \alpha) d\alpha,
\end{aligned}$$

where $L_0 = d\lambda_k^2$. It is easy to show that

$$\begin{aligned}
&L_0 \mathbb{P}(E_{3,e}^c) \le d\lambda^2 \exp(-c_2 \log d) \le c_2 \lambda^2. \\
&\mathbb{P}(c_1 d\gamma_e^2 \ge \alpha) \le \mathbb{P}(\gamma_e \ge \sqrt{\alpha/d}) \le \exp\{-c_2 \frac{\tau_{e-1}\alpha}{d} + \log d\} \quad \forall \alpha \ge L_0.
\end{aligned}$$

Hence, the second term can be upper bounded by

$$\begin{aligned}
&\tau_e \int_{L_n}^\infty \exp\{-c_2 \frac{\tau_{e-1}\alpha}{d} + \log d\} d\alpha \le \frac{d\tau_e}{\tau_{e-1}} \exp\{-\tau_{e-1}\lambda_e^2 + \log d\} \\
&\le d \exp\{-c_1 \log d\}.
\end{aligned}$$

To summarize,

$$\sum_{t=\tau_e+1}^{\tau_{e+1}} \mathbb{E}[\mathrm{reg}_t \mathbb{1}(\mathcal{E}_{2,e} \cap \mathcal{E}_{3,e}^c)] \le s \log d + C.$$

□

Proof of Theorem 4.1.

$$\mathrm{Reg}(T) = \sum_{e=1}^{\log_2 T} \sum_{t \in e\text{-th episode}} reg_t.$$

(i) If $\tau_{e-1} \le c_1 s \log d$, Condition 4.3 to arrive at

$$\sum_{t \in e\text{-th episode}} reg_t \le R\tau_{e-1}.$$

(ii) If $c_1 s \log d \geq \tau_{e-1} \geq d^2/c_1$, then we use Lemma 8.1 to conclude that

$$\sum_{t \in e\text{-th episode}} \mathbb{E}[reg_t] \leq c_1 L_d (s \log d)^{\alpha/2} \tau_e^{1-\alpha/2} = L_d (s \log d)^{\alpha/2} (2^{1-\alpha/2})^e.$$

(iii) If $c_1 \tau_{k-1} > d^2$, we use Lemma 8.2 to conclude that

$$\mathrm{Reg}_{\widehat{\pi}}(e\text{-th episode}) \leq C_2 L_d (s \log d)^{\alpha/2} (2^{1-\alpha/2})^e.$$

Hence,

$$\mathrm{Reg}_{\widehat{\pi}}(T) = \sum_{e=1}^{\log_2 T} \mathrm{Reg}(e\text{-th episode}) \leq Rs \log d + \sum_{e=1}^{\log_2 T} L_d (s \log d)^{\alpha/2} (2^{1-\alpha/2})^e.$$

To summarize,

$$\mathrm{Reg}_{\widehat{\pi}}(T) \leq \begin{cases} Rs \log d + L_d (s \log d) \log T & \text{for} \quad \alpha = 2 \\ Rs \log d + L_d (s \log d)^{\alpha/2} T^{1-\alpha/2} & \text{for} \quad \alpha \in [0, 2) \end{cases}$$

□

8.1 Proof of Corollary 4.1

Proof of Corollary 4.1. For dynamic pricing model,

$$f(\boldsymbol{a}_t, \boldsymbol{x}_t^\top \boldsymbol{\beta}) = f(\boldsymbol{x}_t^\top \boldsymbol{\beta} - a_t),$$

where f is a link function of GLMs. Hence, (6) holds by Negahban et al. (2010).

Now we verify that Condition 4.3. First,

$$r(a_t, \boldsymbol{x}_t^\top \boldsymbol{\beta}) = a_t \times f(\boldsymbol{x}_t^\top \boldsymbol{\beta} - a_t).$$

As $\mathcal{A}$ and f are uniformly bounded, $r(\cdot)$ is uniformly bounded.

Next,

$$|r(a_t, \boldsymbol{x}_t^\top \boldsymbol{\beta}) - r(a_t^*, \boldsymbol{x}_t^\top \boldsymbol{\beta})| \leq \frac{\partial^2 r(a_t, \boldsymbol{x}_t^\top \boldsymbol{\beta})}{\partial^2 a_t} |a_t^* - a_t|^2 \tag{15}$$

$$|\frac{\partial^2 r(a_t, \boldsymbol{x}_t^\top \boldsymbol{\beta})}{\partial^2 a_t}| = -2\dot{f}(\boldsymbol{x}_t^\top \boldsymbol{\beta} - a_t) + a_t \ddot{f}(\boldsymbol{x}_t^\top \boldsymbol{\beta} - a_t).$$

a_t^* is a solution to

$$\begin{aligned} & f(\boldsymbol{x}_t^\top \boldsymbol{\beta} - a_t^*) - a_t^* \dot{f}(\boldsymbol{x}_t \boldsymbol{\beta} - a_t^*) = 0 \\ \Longrightarrow\ & a_t^* = \frac{f(\boldsymbol{x}_t^\top \boldsymbol{\beta} - a_t^*)}{\dot{f}(\boldsymbol{x}_t^\top \boldsymbol{\beta} - a_t^*)}. \end{aligned}$$

Define

$$\psi(v) = v - \frac{f(v)}{\dot{f}(v)}.$$

It holds that

$$\begin{aligned} & \psi(\boldsymbol{x}_t^\top \boldsymbol{\beta} - a_t^*) - \boldsymbol{x}_t^\top \boldsymbol{\beta} = 0 \\ \Longrightarrow\ & a_t^* + \boldsymbol{x}_t^\top \boldsymbol{\beta} = \psi^{-1}(\boldsymbol{x}_t^\top \boldsymbol{\beta}) \\ \Longrightarrow\ & a_t^* = \psi^{-1}(-\boldsymbol{x}_t^\top \boldsymbol{\beta}) + x_t^\top \boldsymbol{\beta} := h(\boldsymbol{x}_t^\top \boldsymbol{\beta}). \end{aligned} \tag{16}$$

Assume that f is strictly increasing. Then $\psi(v)$ is injective and hence $h(\cdot)$ is injective.

$$\left|\dot{h}\right| = \left|1 - \frac{1}{\dot{\psi}(\psi^{-1}(-\boldsymbol{x}_t^\top \boldsymbol{\beta}))}\right|.$$

Hence, a_t^* is Lipchitz in $\boldsymbol{x}_t^\top \boldsymbol{\beta}$. Together with (15), we have verified Condition 4.3. □

8.2 Proof of Corollary 4.2

We only need to verify Condition 4.2 and Condition 4.3 with $\alpha = 2$. The verification is done in the following lemma.

Lemma 8.3. *Assume Condition 4.1. Then Condition 4.2 holds for the contextual bandit problem.*

Proof of Lemma 8.3. For K-arm contextual bandits,

$$\frac{1}{n}\sum_{t=1}^{n}(\boldsymbol{x}_t^\top \boldsymbol{u})\{f(\boldsymbol{a}_t, \boldsymbol{x}_t^\top(\boldsymbol{\beta} + \boldsymbol{u})) - f(\boldsymbol{a}_t, \boldsymbol{x}_t^\top \boldsymbol{\beta})\} = \frac{1}{n}\sum_{t=1}^{n}(\boldsymbol{a}_t^\top \boldsymbol{X}_t \boldsymbol{u})^2.$$

Hence, (6) holds as long as

$$\inf_{u \in 2S(\boldsymbol{\beta}), \|u\|_2 = 1} \frac{1}{n}\sum_{t=1}^{n} \min_{a_t \in \mathcal{A}} |\boldsymbol{a}_t^\top \boldsymbol{X}_t^\top u|^2 \geq c_n \geq c \tag{17}$$

with probability at least $1 - \exp(-n)$. □

Proof of Corollary 4.1. We are left to verify Condition 4.3 assuming (8).

Let $\Delta_{t,k} = \{(\boldsymbol{X}_t)_{k^*} - (\boldsymbol{X}_t)_k\}^\top \boldsymbol{\beta}$.

$$\begin{aligned}
reg_t(\boldsymbol{\beta}') &= (a_t^* - a_t(\boldsymbol{\beta}'))^\top \boldsymbol{X}_t \boldsymbol{\beta} \\
&= \sum_{k \neq k^*} ((\boldsymbol{X}_t)_{k^*} - (\boldsymbol{X}_t)_k)^\top \boldsymbol{\beta} \mathbb{1}(\boldsymbol{a}_t(\boldsymbol{\beta}') = e_k) \\
&\leq \sum_{k \neq k^*} (\boldsymbol{X}_t)_{k^*} - (\boldsymbol{X}_t)_k)^\top \boldsymbol{\beta} \mathbb{1}(\boldsymbol{a}_t(\boldsymbol{\beta}') = e_k, (\boldsymbol{X}_t)_k^\top \boldsymbol{\beta}_t' \geq (\boldsymbol{X}_t)_{k^*}^\top \boldsymbol{\beta}) \\
&\leq \max_{k \leq K} (\boldsymbol{X}_t)_{k^*} - (\boldsymbol{X}_t)_k)^\top \boldsymbol{\beta} \mathbb{1}(\boldsymbol{a}_t(\boldsymbol{\beta}') = e_k, (\boldsymbol{X}_t)_k^\top \boldsymbol{\beta}_t' \geq (\boldsymbol{X}_t)_{k^*}^\top \boldsymbol{\beta}) \\
&\leq \max_{k \leq K} \Delta_{t,k} \mathbb{1}(\{(\boldsymbol{X}_t)_k - (\boldsymbol{X}_t)_{k^*}\}^\top (\boldsymbol{\beta}' - \boldsymbol{\beta}) \geq \Delta_{t,k}),
\end{aligned}$$

where the first inequality is due to the optimality of $a_t(\boldsymbol{\beta}')$, i.e., $r(a_t(\boldsymbol{\beta}'), \boldsymbol{x}_t^T \boldsymbol{\beta}') \geq r(a_t^*, \boldsymbol{x}_t^T \boldsymbol{\beta}')$.

We decompose $reg_t(\boldsymbol{\beta}')$ into two terms. Let $v_t = 2 \max_{j,k} |(\boldsymbol{X}_t)_{j,k}| \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_1$. Notice that $\Delta_{t,k} \leq v_t$.

$$\begin{aligned}
reg_t(\boldsymbol{\beta}') \leq & \underbrace{\max_{k \leq K} \Delta_{t,k} \mathbb{1}(\{(\boldsymbol{X}_t)_k - (\boldsymbol{X}_t)_{k^*}\}^\top (\boldsymbol{\beta}' - \boldsymbol{\beta}) \geq \Delta_{t,k}, \Delta_{t,k} \geq c_0)}_{U_{1,t}} \\
& + \underbrace{\max_{k \leq K} \Delta_{t,k} \mathbb{1}(\{(\boldsymbol{X}_t)_k - (\boldsymbol{X}_t)_{k^*}\}^\top (\boldsymbol{\beta}' - \boldsymbol{\beta}) \geq \Delta_{t,k}, \Delta_{t,k} \leq v_k \wedge c_0)}_{U_{2,t}}.
\end{aligned}$$

$$\begin{aligned}
U_{1,t} &\leq \max_{k \leq K} \{(\boldsymbol{X}_t)_k - (\boldsymbol{X}_t)_{k^*}\}^\top (\boldsymbol{\beta}' - \boldsymbol{\beta}) \mathbb{1}(\max_{k \leq K} \{(\boldsymbol{X}_t)_k - (\boldsymbol{X}_t)_{k^*}\}^\top (\boldsymbol{\beta}' - \boldsymbol{\beta}) \geq c_0) \\
&\leq \frac{1}{c_0} |\{\max_{k \leq K} (\boldsymbol{X}_t)_k - (\boldsymbol{X}_t)_{k^*}\}^\top (\boldsymbol{\beta}' - \boldsymbol{\beta})|^2 \\
&\leq \frac{2}{c_0} \{\max_{k \leq K} (\boldsymbol{X}_t)_k^\top (\boldsymbol{\beta}' - \boldsymbol{\beta})\}^2 \leq \frac{2}{c_0} v_t^2.
\end{aligned}$$

For $U_{2,t}$,

$$\begin{aligned}
U_{2,t} &\leq \max_{k \leq K} \Delta_{t,k} \mathbb{1}(\max_{t,k} \Delta_{t,k} \leq v_t \wedge c_0) \\
&\leq v_t \wedge c_0 \mathbb{1}(\max_{k \leq K} \Delta_{t,k} \leq v_t \wedge c_0).
\end{aligned}$$

Notice the upper bound on $U_{2,t}$ is independent of $\boldsymbol{\beta}'$. Hence,

$$\mathbb{E}[M_e(S, \kappa)] \leq \mathbb{E}[\sup_{\|\boldsymbol{u}\|_2 \leq \kappa, \boldsymbol{u} \in \mathcal{C}(3,S)} \frac{1}{\tau_e} \sum_{t=\tau_{e-1}+1}^{\tau_e} U_{1,t}] + \mathbb{E}[\sup_{\|\boldsymbol{u}\|_2 \leq \kappa, \boldsymbol{u} \in \mathcal{C}(3,S)} \frac{1}{\tau_e} \sum_{t=\tau_{e-1}+1}^{\tau_e} U_{2,t}].$$

$$\mathbb{E}[\sup_{\|\boldsymbol{u}\|_2\le\kappa,\boldsymbol{u}\in\mathcal{C}(3,S)}\frac{1}{\tau_e}\sum_{t=\tau_{e-1}+1}^{\tau_e}U_{1,t}]\le\frac{2}{c_0}\mathbb{E}[\sup_{\|\boldsymbol{u}\|_2\le\kappa,\boldsymbol{u}\in\mathcal{C}(3,S)}\frac{1}{\tau_e}\sum_{t=\tau_{e-1}+1}^{\tau_e}v_t^2]$$
$$\le s\kappa^2\mathbb{E}[\frac{1}{\tau_e}\sum_{t=\tau_{e-1}+1}^{\tau_e}\max_{j,k}(\boldsymbol{X}_t)_{j,k}^2]$$
$$\le s\kappa^2(\log d+\log K).$$

Notice that $U_{2,t}$ is an increasing function of v_t and $\sup_{\|\boldsymbol{u}\|_2\le\kappa,\boldsymbol{u}\in\mathcal{C}(3,S)}v_t\le\max_{j,k}|\boldsymbol{X}_t|_{j,k}\sqrt{s}\kappa$. Moreover, $\Delta_{t,k}$ is independent of $\boldsymbol{u}$. Hence

$$\mathbb{E}[\sup_{\|\boldsymbol{u}\|_2\le\kappa,\boldsymbol{u}\in\mathcal{C}(3,S)}\frac{1}{\tau_e}\sum_{t=\tau_{e-1}+1}^{\tau_e}U_{2,t}]\le\mathbb{E}[\frac{1}{\tau_e}\sum_{t=\tau_{e-1}+1}^{\tau_e}(\max_{j,k}|\boldsymbol{X}_t|_{j,k}\sqrt{s}\kappa)\wedge c_0\mathbb{1}(\max_{k\le K}\Delta_{t,k}\le\max_{j,k}|\boldsymbol{X}_t|_{j,k}\sqrt{s}\kappa\wedge c_0)]$$
$$=\underbrace{\mathbb{E}[(\max_{j,k}|\boldsymbol{X}_t|_{j,k}\sqrt{s}\kappa)\wedge c_0\mathbb{1}(\max_{k\le K}\Delta_{t,k}\le\max_{j,k}|\boldsymbol{X}_t|_{j,k}\sqrt{s}\kappa\wedge c_0)]}_{U_{2,e}}$$
$$\le\mathbb{E}[U_{2,e}\mathbb{1}(\max_{j,k}|(\boldsymbol{X}_t)_{j,k}|\le C\sqrt{\log d+\log K})]+\mathbb{E}[U_{2,e}\mathbb{1}(\max_{j,k}|(\boldsymbol{X}_t)_{j,k}|\ge C\sqrt{\log d+\log K})].$$

For the first term, we can use the margin condition (8) such that

$$\mathbb{E}[U_{2,e}\mathbb{1}(\max_{j,k}|(\boldsymbol{X}_t)_{j,k}|\le C\sqrt{\log d+\log K})]$$
$$\le(\sqrt{(\log d+\log K)s}\kappa)\wedge c_0\mathbb{P}(\max_{j,k}|(\boldsymbol{X}_t)_{j,k}|\le\sqrt{(\log d+\log K)s}\kappa\wedge c_0)$$
$$\le(\log d+\log K)s\kappa^2\wedge c_0.$$

For the second term,

$$\mathbb{E}[U_{2,e}\mathbb{1}(\max_{j,k}|(\boldsymbol{X}_t)_{j,k}|\ge C\sqrt{\log d+\log K})]\le c_0\mathbb{P}(\max_{j,k}|(\boldsymbol{X}_t)_{j,k}|\ge C\sqrt{\log d+\log K})=o(1).$$

To summarize,

$$\mathbb{E}[M_e(S,\kappa)]\le(\log d+\log K)s\kappa^2+o(1).$$

□

Proof of Corollary 4.3. We only need to verify Condition 4.3 with $\alpha=1$. For any $\boldsymbol{\beta}'$ such

that $\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2 \leq \kappa$, it holds that

$$\begin{aligned} r(a^*, \boldsymbol{x}_t^T\boldsymbol{\beta}) - r(a_t(\boldsymbol{\beta}'), \boldsymbol{x}_t^T\boldsymbol{\beta}) &= (a_t^* - a_t(\boldsymbol{\beta}'))\boldsymbol{X}_t\boldsymbol{\beta} \\ &\leq (a_t^* - a_t(\boldsymbol{\beta}'))\boldsymbol{X}_t(\boldsymbol{\beta} - \boldsymbol{\beta}_t'), \end{aligned}$$

where the last step is due to $r(a_t(\boldsymbol{\beta}'), \boldsymbol{x}_t^T\boldsymbol{\beta}') \geq r(a_t, \boldsymbol{x}_t^T\boldsymbol{\beta}')$. Hence,

$$\begin{aligned} \mathbb{E}[M_e(S,\kappa)] &\leq 2\mathbb{E}[\sup_{\|u\|_2=\kappa, u\in\mathcal{C}(3,S)} \frac{1}{\tau_e}\sum_{t=\tau_{e-1}+1}^{\tau_e} \max_{k,j} |(\boldsymbol{X}_t)_{k,j}|\|u\|_1] \\ &\leq 6\sqrt{s}\kappa\mathbb{E}[\max_{k,j}|(\boldsymbol{X}_t)_{k,j}|]. \end{aligned} \tag{18}$$

Using the sub-Gaussian property of $(\boldsymbol{X}_t)_k$,

$$\mathbb{E}[|\max_{k,j}|(\boldsymbol{X}_t)_{k,j}|] \leq C_{\max}\sqrt{\log K + \log d}.$$

Therefore, (18) implies that Condition 4.3 holds for $L_d = \sqrt{s(\log K + \log d)}$.

□

8.3 Proof of Theorem 4.2 (lower bound)

Proof of Theorem 4.2. Consider a proposed policy $\pi_t = \boldsymbol{a}(\boldsymbol{x}_t^\top\widehat{\boldsymbol{\beta}}_t)$, where $\widehat{\beta}_t$ is $\mathcal{H}_{t-1}$-measureable. The regret under π is

$$\begin{aligned} \sup_\beta reg_t &= \sup_\beta \mathbb{E}[r(\boldsymbol{a}_t^*, \boldsymbol{x}_t^\top\boldsymbol{\beta}) - r(\boldsymbol{a}(\widehat{\boldsymbol{\beta}}_t), \boldsymbol{x}_t^\top\boldsymbol{\beta})] \\ &= \sup_\beta \mathbb{E}[\mathbb{E}[r(\boldsymbol{a}_t^*, \boldsymbol{x}_t^\top\boldsymbol{\beta}) - r(\boldsymbol{a}(\widehat{\boldsymbol{\beta}}_t), \boldsymbol{x}_t^\top\boldsymbol{\beta})|\widetilde{\mathcal{H}}_{t-1}]] \\ &\geq \sup_\beta L_d\mathbb{E}[\|\widehat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}\|_2^2]. \end{aligned}$$

Therefore,

$$\inf_\pi \sup_\beta Reg(T) \geq L_d \sum_{t=1}^{\top} \inf_{\widehat{\boldsymbol{\beta}}_t} \sup_{\boldsymbol{\beta}} \mathbb{E}[\|\widehat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}\|_2^2]$$

We state Fano's inequalty here (Lemma 7 in Zhao, Cai, and Zhou)

Lemma 8.4. *Let $\Theta = \{\theta_i : i = 0, \ldots, m_*\}$ be a parameter set and ρ be a distance over Θ.*

Let $\{\mathbb{P}_\theta : \theta \in \Theta\}$ *be a collection of probability distribution satisfying*

$$\frac{1}{m^*} \sum_{1 \le i \le m_*} KL(\mathbb{P}_{\theta_i}, \mathbb{P}_{\theta_0}) \le c \log m_*$$

with $0 < c < 1/8$. *Let* $\widehat{\theta}$ *be any estimator based on an observation from a distribution in* $\{\mathbb{P}_\theta, \theta \in \Theta\}$. *Then*

$$\sup_{\theta \in \Theta} \mathbb{E}[d^2(\widehat{\theta}, \theta)] \ge \min_{i \ne j} \frac{d^2(\theta_i, \theta_j)}{4}.$$

By Lemma 4 of Raskutti et al. (2009), there exist $\{\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^M\} \in \mathcal{B}_0(s)$ such that $\|\boldsymbol{\beta}^j - \boldsymbol{\beta}^k\|_2^2 \ge s \log d/t$ for any $j \ne k$ with $M \ge s \log d$. In fact, each element of $\boldsymbol{\beta}^i$ is in $\{-\sqrt{\log d/t}, 0, \sqrt{\log d/t}\}$. Next, we upper bound the KL distance between $\mathbb{P}^\top_{\boldsymbol{\beta}^j}$ and $\mathbb{P}^\top_{\boldsymbol{\beta}^0}$. It is easy to verify that

$$KL(\mathbb{P}^\top_{\theta_i}, \mathbb{P}^\top_{\theta_0}) \le cs \log d.$$

Using the above lemma,

$$\sup_{\theta \in \Theta} \mathbb{E}[d^2(\widehat{\theta}, \theta)] \ge c_2 \frac{s \log p}{t}.$$

Therefore,

$$\mathrm{Reg}(T) = \sum_{t=1}^{\top} \mathbb{E}[\mathrm{Reg}_t] \ge \sum_{t=1}^{\top} c_2 \frac{s \log d}{t} = c_2 s \log d \log T.$$

□

8.4 Proofs for Section 5

Lemma 8.5 (A lower bound of V_t)**.**

$$\widehat{\boldsymbol{w}}_t^T \widehat{\Sigma}_{\boldsymbol{\beta}, e-1} \widehat{\boldsymbol{w}}_t \ge c_1 \|x_t\|_2^2 (1 - o_P(1)).$$

Proof of Lemma 8.5. By (9), for any feasible solution $\boldsymbol{w}$,

$$\boldsymbol{x}_t^T (\boldsymbol{x}_t - \widehat{\Sigma}_{\widehat{\boldsymbol{\beta}}, e-1} \boldsymbol{w}) \le \lambda_e \|\boldsymbol{x}_t\|_2^2$$

For any $c \geq 0$,

$$\begin{aligned}\boldsymbol{w}^T\widehat{\Sigma}_{\widehat{\boldsymbol{\beta}},e-1}\boldsymbol{w} &\geq \boldsymbol{w}^T\widehat{\Sigma}_{\widehat{\boldsymbol{\beta}},e-1}\boldsymbol{w} + c(\|\boldsymbol{x}_t\|_2^2 - \lambda_e\|\boldsymbol{x}_t\|_2^2 - \boldsymbol{x}_t^T\widehat{\Sigma}_{\widehat{\boldsymbol{\beta}},e-1}\boldsymbol{w}) \\ &\leq \min_{\boldsymbol{w}}\{\boldsymbol{w}^T\widehat{\Sigma}_{\widehat{\beta},e-1}\boldsymbol{w} + c(\|\boldsymbol{x}_t\|_2^2 - \lambda_e\|\boldsymbol{x}_t\|_2^2 - \boldsymbol{x}_t^T\widehat{\Sigma}_{\widehat{\boldsymbol{\beta}},e-1}\boldsymbol{w})\} \\ &= -\frac{c^2}{2}\boldsymbol{x}_t^T\widehat{\Sigma}_{\widehat{\boldsymbol{\beta}},e-1}\boldsymbol{x}_t + c(\|\boldsymbol{x}_t\|_2^2 - \lambda_e\|\boldsymbol{x}_t\|_2^2).\end{aligned}$$

Taking $c = \|\boldsymbol{x}_t\|_2^2\frac{1-\lambda_e}{\boldsymbol{x}_t^T\widehat{\Sigma}_{\widehat{\boldsymbol{\beta}},e-1}\boldsymbol{x}_t}$, we have

$$\widehat{\boldsymbol{w}}_t^T\widehat{\Sigma}_{\widehat{\boldsymbol{\beta}},e-1}\widehat{\boldsymbol{w}}_t \geq \|\boldsymbol{x}_t\|_2^4\frac{(1-\lambda_e)^2}{2\boldsymbol{x}_t^T\widehat{\Sigma}_{\widehat{\boldsymbol{\beta}},e-1}\boldsymbol{x}_t}.$$

Conditioning on $\boldsymbol{x}_t$ and $\widehat{\boldsymbol{\beta}}_{e-1}$,

$$\boldsymbol{x}_t^T\widehat{\Sigma}_{\widehat{\boldsymbol{\beta}},e-1}\boldsymbol{x}_t = \boldsymbol{x}_t^T\mathbb{E}[\widehat{\Sigma}_{\widehat{\boldsymbol{\beta}},e-1}|\widehat{\boldsymbol{\beta}}_{e-1}]\boldsymbol{x}_t + O_P(\frac{\|\boldsymbol{x}_t\|_2^2}{\sqrt{\tau_{e-1}}}) \leq (C_1 - o_P(1))\|\boldsymbol{x}_t\|_2^2.$$

As a result,

$$\widehat{\boldsymbol{w}}_t^T\widehat{\Sigma}_{\widehat{\boldsymbol{\beta}},e-1}\widehat{\boldsymbol{w}}_t \geq c_1\|\boldsymbol{x}_t\|_2^2(1-\lambda_e)^2(1-o_P(1)).$$

$$\widehat{\boldsymbol{w}}_t^T\widehat{\Sigma}_{\boldsymbol{\beta},e-1}\widehat{\boldsymbol{w}}_t \geq \widehat{\boldsymbol{w}}_t^T\widehat{\Sigma}_{\widehat{\boldsymbol{\beta}},e-1}\widehat{\boldsymbol{w}}_t(1 - \max_{\tau_k^-+1\leq j\leq\tau_k}|\dot{f}(\boldsymbol{x}_j^T\boldsymbol{\beta})/\dot{f}(\boldsymbol{x}_j^T\widehat{\boldsymbol{\beta}}_{e-1}) - 1|) = \widehat{\boldsymbol{w}}_t^T\widehat{\Sigma}_{\widehat{\boldsymbol{\beta}},e-1}\widehat{\boldsymbol{w}}_t(1-o_P(1)).$$

□

Proof of Theorem 5.1. By (10),

$$\begin{aligned}&\widehat{\mu}_t - \boldsymbol{x}_t^T\boldsymbol{\beta}^* = \boldsymbol{x}_t^T\widehat{\boldsymbol{\beta}}_e - \boldsymbol{x}_t^T\boldsymbol{\beta}^* + \frac{\sum_{j=\tau_{e-1}+1}^{\tau_e}\langle\boldsymbol{x}_j^T\widehat{\boldsymbol{w}}_t, y_j - f(p_j - \boldsymbol{x}_j^T\widehat{\boldsymbol{\beta}}_e)\rangle}{\tau_{e-1}} \\ &= x_t^T(\widehat{\boldsymbol{\beta}}_e - \boldsymbol{\beta}) - \frac{\sum_{j=\tau_{e-1}+1}^{\tau_e}\langle\boldsymbol{x}_j^T\widehat{\boldsymbol{w}}_t, f(p_j - \boldsymbol{x}_j^T\widehat{\boldsymbol{\beta}}_e) - f(p_j - \boldsymbol{x}_j^T\boldsymbol{\beta})\rangle}{\tau_{e-1}} + \underbrace{\frac{\sum_{j=\tau_{e-1}+1}^{\tau_e}\langle\boldsymbol{x}_j^T\widehat{\boldsymbol{w}}_t, y_j - f(p_j - \boldsymbol{x}_j^T\boldsymbol{\beta})\rangle}{\tau_{e-1}}}_{R_{1,t}} \\ &= \underbrace{(\boldsymbol{x}_t - \frac{\sum_{j=\tau_{e-1}+1}^{\tau_e}\boldsymbol{x}_j^T\widehat{\boldsymbol{w}}_t\{f(p_j - \boldsymbol{x}_j^T\widehat{\boldsymbol{\beta}}_e) - f(p_j - \boldsymbol{x}_j^T\boldsymbol{\beta})\}}{\tau_{e-1}}}_{R_{2,t}} + R_{1,t}.\end{aligned}$$

For $R_{1,t}$, we prove its asymptotic normality. Specifically, $\widehat{\boldsymbol{w}}_t$ is a function of $\{\boldsymbol{x}_j\}_{j=\tau_{e-1}+1}^{\tau_e}$ and $\widehat{\boldsymbol{\beta}}_{e-1}$. Conditioning on $\{\boldsymbol{x}_j\}_{j=\tau_{e-1}+1}^{\tau_e}$ and $\mathcal{H}_{\tau_{e-1}}$, $R_{1,t}$ is a sum of independent random

variables with mean zero. As $\max_{\tau_{e-1}+1\le j\le\tau_e}|\boldsymbol{x}_j^T\widehat{\boldsymbol{w}}_t|\le C\sqrt{\log n}$. Together with Lemma 8.5, one can easily use the Lyapunov central limit theorem to show that

$$\sqrt{\frac{\tau_{e-1}}{V_t}}R_{1,t}\xrightarrow{D}N(0,1).$$

For $R_{2,t}$, by Taylor expansion,

$$\begin{aligned}|R_{2,t}| &= (\boldsymbol{x}_t-\widehat{\Sigma}_{\widehat{\boldsymbol{\beta}},e-1}\widehat{\boldsymbol{w}}_t)^T(\widehat{\boldsymbol{\beta}}_e-\boldsymbol{\beta})+\frac{c_1}{\tau_{e-1}}\sum_{j=\tau_{e-1}+1}^{\tau_e}(\boldsymbol{x}_j^T\widehat{\boldsymbol{w}}_t)\{\boldsymbol{x}_j^T(\widehat{\boldsymbol{\beta}}_e-\boldsymbol{\beta})\}\{\boldsymbol{x}_j^T(\widehat{\boldsymbol{\beta}}_{e-1}-\boldsymbol{\beta})\}\\ &\le \lambda_e\|\widehat{\boldsymbol{\beta}}_e-\boldsymbol{\beta}\|_1+\|\boldsymbol{x}_t\|_2\sqrt{\sum_{j=\tau_{e-1}+1}^{\tau_e}|\{\boldsymbol{x}_j^T(\widehat{\boldsymbol{\beta}}_e-\boldsymbol{\beta})\}^2}\sqrt{\sum_{j=\tau_{e-1}+1}^{\tau_e}|\{\boldsymbol{x}_j^T(\widehat{\boldsymbol{\beta}}_{e-1}-\boldsymbol{\beta})\}^2}\\ &= O_P\left(\|\boldsymbol{x}_t\|_2\frac{\sqrt{\log\tau_{e-1}}s\log d}{\tau_{e-1}}\right)\end{aligned}$$

Hence,

$$\sqrt{\frac{\tau_{e-1}}{2V_t}}(\widehat{\mu}_t-\boldsymbol{x}_t^T\boldsymbol{\beta})=Z_t+O_P\left(\frac{\sqrt{\log\tau_{e-1}}s\log d}{\tau_{e-1}}\right).$$

□

8.5 Minimax optimality

Consider the parameter space

$$\Omega = \{\|\beta\|_0 \leq s, c_1 I \preceq \ \Sigma \preceq c_2 I\}.$$

By Lemma 16 in Javanmard and Nazerzadeh (2019b), there exists some constant δ such that

$$\mathbb{E}[\text{Reg}_t|\widetilde{\mathcal{H}}_{t-1}] = c_1 \sum_{t=1}^{T} \ddot{\text{rev}}_t(\widetilde{p})(p_t - p_t^*)^2 \geq c_2 \sum_{t=1}^{T} \min\{|x_t^T(\beta_t - \beta)|^2, \delta^2\}.$$

Therefore,

$$\mathbb{E}[\text{Reg}_t] \geq c_3 \mathbb{E}[\min\{\|\beta_t - \beta\|_2^2, \delta^2\}].$$

By Fano's inequality,

$$\inf_{\beta_t} \sup_{\beta} \mathbb{E}[\|\beta_t - \beta\|_2^2] \geq \rho_t \mathbb{P}\left(\inf_{\beta_t} \sup_{\beta} \|\beta_t - \beta\|_2^2 \geq \rho_t\right) \geq \rho_n \min_{\widetilde{\beta}} \mathbb{P}\left(\widetilde{\beta} \neq B_t\right),$$

where $B_t \in \mathbb{R}^d$ is uniformly distributed over the packing set $\{\beta^1, \ldots, \beta^M\}$ and $\widetilde{\beta}$ takes value in the packing set. We can choose $\rho_n = c_1 s \log d/t$ and show that

$$\min_{\widetilde{\beta}} \mathbb{P}\left(\widetilde{\beta} \neq B_t\right) \geq \frac{1}{4}.$$

Therefore,

$$\text{Reg}(T) = \sum_{t=1}^{T} \mathbb{E}[\text{Reg}_t] \geq \sum_{t=1}^{T} c_2 \frac{s \log d}{t} = c_2 s \log d \log T.$$

References

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

Ariu, K., K. Abe, and A. Proutière (2020). Thresholded lasso bandit. *arXiv preprint arXiv:2010.11994*.

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

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

Bird, S., S. Barocas, K. Crawford, F. Diaz, and H. Wallach (2016). Exploring or exploiting? social and ethical implications of autonomous experimentation in ai. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

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

Chen, H., W. Lu, and R. Song (2020a). Statistical inference for online decision-making: In a contextual bandit setting. *Journal of the American Statistical Association* (Forthcoming), 1–22.

Chen, H., W. Lu, and R. Song (2020b). Statistical inference for online decision making via stochastic gradient descent. *Journal of the American Statistical Association* (Forthcoming), 1–45.

Chen, X. and D. Simchi-Levi (2012). Pricing and inventory management. *The Oxford handbook of pricing management 1*, 784–824.

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

Dani, V., T. P. Hayes, and S. M. Kakade (2008). Stochastic linear optimization under bandit feedback. pp. 355–366.

Dean, S., H. Mania, N. Matni, B. Recht, and S. Tu (2018). Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, pp. 4188–4197.

den Boer, A. V. and B. Zwart (2014). Simultaneously learning and optimizing using controlled variance pricing. *Management science 60*(3), 770–783.

Ferreira, K. J., D. Simchi-Levi, and H. Wang (2018). Online network revenue management using thompson sampling. *Operations research 66*(6), 1586–1602.

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

Javanmard, A. and H. Nazerzadeh (2019a). Dynamic pricing in high-dimensions. *The Journal of Machine Learning Research 20*(1), 315–363.

Javanmard, A. and H. Nazerzadeh (2019b). Dynamic pricing in high-dimensions. *The Journal of Machine Learning Research 20*(1), 315–363.

Javanmard, A., H. Nazerzadeh, and S. Shao (2020). Multi-product dynamic pricing in high-dimensions with heterogeneous price sensitivity. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 2652–2657. IEEE.

Kim, G.-S. and M. C. Paik (2019). Doubly-robust lasso bandit. In *Advances in Neural Information Processing Systems*, pp. 5877–5887.

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

Langford, J. and T. Zhang (2007). The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in Neural Information Processing Systems 20*, 817–824.

Lattimore, T. and C. Szepesvári (2020). *Bandit algorithms*. Cambridge University Press.

Li, L., W. Chu, J. Langford, and R. E. Schapire (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670.

Mills, E. S. (1959). Uncertainty and price theory. *The Quarterly Journal of Economics 73*(1), 116–130.

Oh, M.-h., G. Iyengar, and A. Zeevi (2020). Sparsity-agnostic lasso bandit. *arXiv preprint arXiv:2007.08477*.

Petruzzi, N. C. and M. Dada (1999). Pricing and the newsvendor problem: A review with extensions. *Operations research 47*(2), 183–194.

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

Salant, Y. and J. Cherry (2020). Statistical inference in games. *Econometrica 88*(4), 1725–1752.

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

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