Doubly High-Dimensional Contextual Bandits: An Interpretable Model with Applications to Assortment/Pricing

Anonymous Authors[1]

Abstract

We consider contextual bandits that are doubly high-dimensional in the sense that both covariates and actions are allowed to take values in high-dimensional spaces. We propose a simple model that captures the interactions between covariates and actions via a (near) low-rank representation matrix. The resulting class of models is reasonably expressive while remaining interpretable, and includes various structured linear bandit models as particular cases. We propose a computationally tractable procedure that combines an exploration/exploitation protocol with an efficient low-rank matrix estimator, and we prove bounds on its regret. Simulation results show that this method has lower regret than state-of-the-art methods applied to various standard bandit models. We also apply our method to a real-world online retail data set involving assortment and pricing; in contrast to most existing methods, our method allows the assortment-pricing problem to be solved simultaneously. We demonstrate the effectiveness of this joint approach for revenue maximization.

1. Introduction

The bandit problem, dating back to the seminal work of Robbins (1952), is a canonical problem in sequential decision-making. At each round, a decision-maker chooses an action (arm) at each round, and then observes a reward. The goal is to act strategically so as to determine a near-optimal policy without incurring large regret. There is now a very well-developed literature on the bandit problem, and its extension to the contextual bandits; see the book by Lattimore & Szepesvári (2020) and references therein for more background. Contextual bandit models and algorithms play a central role in online decision-making, due to with e-commerce and health care being two fruitful domains.

The classical bandit formulation involves a finite action space, but many modern applications lead to settings where actions take values in continuous space, also known as the *continuum-armed bandit* (e.g., Agrawal (1995); Kleinberg (2004)). For instance, in e-commerce, an online retailer seeks to decide upon product assortment and pricing so as to maximize long-term profits (Caro & Gallien, 2007; Sauré & Zeevi, 2013; den Boer & Zwart, 2014; Keskin & Zeevi, 2014). In mobile health, the personal device provides exercise and dietary suggestions to improve physical and mental health (see Debon et al. (2019) and references therein). In both of these cases, the actions take values in some subset of $\mathbb{R}^{d_a}$, where the action dimension d_a can be quite large, which we refer to as the high-dimensional setting. In the contextual bandit problem, decision-makers observe an additional vector $\boldsymbol{x} \in \mathbb{R}^{d_x}$ of features or covariates, also known as the context. In applications, the covariate dimension d_x may also be high-dimensional. The reward mean is modeled as some unknown function of the covariate-action pair $(\boldsymbol{x}, \boldsymbol{a})$.

As the action-covariate dimensionalities d_a and d_x grow, traditional bandit algorithms suffer from the curse of dimensionality; indeed, without some kind of structure, there are "no-free-lunch" theorems showing that it is prohibitively costly, both in terms of samples and computation, to learn an optimal policy (Lattimore & Szepesvári, 2020). This fact motivates various models that encode some form of low-dimensional structure in the reward function. To date, researchers have pursued structure in the covariates and actions in isolation, including sparsity for high-dimensional contextual bandit problems (e.g., Bastani & Bayati (2020)), or subspace structure for continuum-action bandits (e.g., Tyagi et al. (2016)). This line of work leaves open the following question:

Are there useful models and efficient learning procedures for contextual bandits that are high-dimensional in both actions and covariates?

In this paper, we tackle this challenge by proposing a new model that captures interactions between actions and covariates via an (approximately) low-rank matrix representation. Within this model class, we also propose a new algorithm (`Hi-CCAB`) that combines low-rank estimation with exploration, and prove some non-asymptotic bounds on its expected regret.

Our reward model takes the following form: given a covariate vector $\boldsymbol{x} \in \mathbb{R}^{d_x}$ and an action vector $\boldsymbol{a} \in \mathbb{R}^{d_a}$, we observe a noisy reward Y with conditional mean

$$\mathbb{E}[Y \mid \boldsymbol{x}, \boldsymbol{a}] = \boldsymbol{a}^T \boldsymbol{\Theta} \boldsymbol{x},$$

where $\boldsymbol{\Theta} \in \mathbb{R}^{d_a \times d_x}$ is an unknown representation matrix. As we discuss in the sequel, in many applications, it is natural to assume that this representation matrix is relatively low-rank—say with rank $r \ll \min\{d_a, d_x\}$. Given this structure, our proposed `Hi-CCAB` algorithm interleaves estimation steps, in which the low-rank representation matrix is estimated based on data observed thus far, and exploration/exploitation steps. In the proposal given here, we analyze a standard estimator based on the nuclear norm relaxation of rank (cf. Chapter 10 in Wainwright (2019) for details). While the estimator itself is not novel, our analysis of it does require new ingredients since we apply it to data adaptively collected under a bandit protocol. We further demonstrate the benefits of our methodologies in e-commerce with real sales data where the online retailer needs to decide on the product assortment and pricing jointly. The generality of our model makes it possible to learn policy on product assortment and pricing at the same time, while previous literature mostly studies the assortment and pricing problem separately.

Contributions. Let us summarize some of our main contributions:

1. We propose a new model for high-dimensional contextual bandits, in which both the covariates and actions can be high-dimensional and continuous. The crux of our model is a low-rank representation matrix that represents the interaction between action-covariate pairs via its left and right singular vectors. This model, while quite simple, unifies a number of structured bandit models analyzed in past work.
2. As we argue, an advantage of this low-rank model is its combination of prediction power with a high degree of interpretability. Performing a singular value decomposition (SVD) on the representation matrix yields the latent structure, with the left (respectively right) singualr vectors corresponding to the action (respectively covariate) space structure. In this way, our model implicitly performs a form of dimension reduction in how the actions and covariates interact to determine the reward function. On the other hand, given the covariate, our model is able to predict the reward of an unseen arm. Both interpretability and predictive power can be tremendously useful for decision-makers.
3. We propose an efficient algorithm for on-line learning in the active setting, referred to as the **Hi**gh-dimensional **C**ontextual and **Hi**gh-dimensional **C**ontinumm **A**rmed **B**andit (`Hi-CCAB`) by adopting the low-rank matrix estimator. We further provide a non-asymptotic upper bound on the expected regret of `Hi-CCAB`.
4. The generality of our model allows for a wide range of applications. Specifically, we apply `Hi-CCAB` to the joint assortment and pricing problem. We show that our model reveals insights for product designs, assortment, and pricing and that the assortment-pricing policy based on `Hi-CCAB` yields sales four times as high as the original strategy.

Connections to past work. Literature on high-dimensional bandit problems has been expanding recently, especially after statistical tools for high-dimensional problems become mature (e.g., see the book (Wainwright, 2019) and references therein). Lots of high-dimensional bandit literature focuses on contextual bandits with high-dimensional covariates, such as the LASSO bandit problem (Abbasi-Yadkori et al., 2012; Kim & Paik, 2019; Bastani & Bayati, 2020; Hao et al., 2020; Papini et al., 2021) where they assume the mean reward is a linear function of a sparse unknown parameter vector; and low-rank matrix bandits where both the covariate and unknown parameters are of matrix form (Kveton et al., 2017; Lu et al., 2021). These high-dimensional bandit models are special cases of our model. Other work uses non-parametric methods, such as boosting, random forests or neural networks, for estimating the reward function (Féraud et al., 2016; Zhou et al., 2020; Ban et al., 2022; Chen et al., 2022; Xu et al., 2022). We compare to one such method in our experimental results.

There are various other models and problems that have connections to but differ from the set-up in this paper. For example, one line of research focuses on representation learning in linear bandits, specifically for multi-task learning where several bandits are played concurrently. The arms for each task are embedded in the same space and share a common low-dimensional representation (Lale et al., 2019; Yang et al., 2020; Hu et al., 2021; Xu & Bastani, 2021). Our problem differs from this set-up of multi-task learning, since each time involves only a single bandit (and single reward), whereas observations in the the multi-task bandit problem consist of multiple rewards at each round.

Our reward model has a bilinear structure in both the covariate and action spaces. It is connected to but different from papers that propose bilinear-type reward models (e.g., Jun et al. (2019); Kim & Vojnovic (2021); Rizk et al. (2021)) in which *both* arguments of the bilinear function are part of the action. Such models can be understood as a structured linear bandit of a particular type, and unlike our models, do not capture the interaction between the covariate and action at each time step.

For continuum-action bandits, there exists a thread of literature that assumes the mean reward function is smooth and continuous on the action space in some sense, e.g., the function lies in the Lipschitz or Hölder space (Agrawal, 1995;

Kleinberg, 2004; Kleinberg et al., 2019). Approaches taken in this work include discretizing the action space, or using non-parametric regression to estimate the reward function, which is quite different from our model. Other work on contextual bandits with continuous actions and covariates assumes that the reward function is continuous with Lipchitz type conditions over the action-covariate space and it is restricted to relatively low-dimension (Lu et al., 2010; Slivkins, 2011; Krishnamurthy et al., 2020). There are a few recent papers on high-dimensional models for contextual bandits (Turğay et al., 2020), but using rather different techniques with relatively strict assumptions and lacking the interpretability of our model.

Methods for low-rank matrix prediction and estimation have been studied extensively in both statistics and machine learning (e.g., (Srebro et al., 2005; Recht et al., 2010; Candes & Plan, 2010; Negahban & Wainwright, 2011; Cai & Zhang, 2018)). Our `Hi-CCAB` algorithm uses least-squares with nuclear norm regularization, which is a well-known approach, but our analysis of it in the bandit setting requires some novel results so as to deal with the adaptive nature of bandit data collection.

Finally, in the field of operations research, assortment and pricing are key decisions to be made by any firm; accordingly, there is a substantial body of past work on dynamic assortment and dynamic pricing. Much of the work on assortment is based on the multinomial logit (MNL) choice model (Caro & Gallien, 2007; Kök et al., 2008; Sauré & Zeevi, 2013); more recent work adopts multi-arm bandit techniques to the MNL model (Chen & Wang, 2017; Agrawal et al., 2019; Kallus & Udell, 2020; Chen et al., 2021). For dynamic pricing, the problem usually comes with demand learning. In presence of covariates, the demand can be modeled as a parametric function (Qiang & Bayati, 2016; Ban & Keskin, 2021) or a nonparametric function (Chen & Gallego, 2021) which adopt the continuum-action bandit techniques in Slivkins (2011). However, there are relatively few papers on the joint assortment-pricing problem. Recently, Miao & Chao (2021) provides a solution using the MNL choice model with a finite number of actions, while our model involves infinitely many actions. In addition, their model assumes the products are independent of each other and can only handle a small number of products. Their model can neither incorporates contextual information nor predicts new products.

Roadmap. The rest of the paper is organized as follows. Section 2 describes the problem formulation and introduces our model with two concrete examples in assortment-pricing and health care. Section 3 presents our `Hi-CCAB` algorithm and its convergence result. Finally, Section 4 shows the empirical results on simulated data and a case study on real sales data from one of the largest online retailers. The proof of our theorem and additional empirical results are provided in the Appendix.

Notation. We use bold lowercase for vectors and bold uppercase for matrices. For any vector $\boldsymbol{a}$, we use $\|\boldsymbol{a}\|$ to denote its ℓ_2 norm. For any matrix $\boldsymbol{A}$, we use $\|\boldsymbol{A}\|_F := \sum_{ij} a_{ij}^2$ to denote its Frobenius norm, $\|\boldsymbol{A}\|_2$ to denote its ℓ_2 spectrum norm, i.e., $\|\boldsymbol{A}\|_2 := \sup_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{A}\boldsymbol{x}\|_2$, and $\|\boldsymbol{A}\|_* := \sum_{k=1}^d s_k$ to denote its nuclear norm where d is the rank and s_k's are the singular values of $\boldsymbol{A}$. We use $\langle \boldsymbol{a}, \boldsymbol{b}\rangle := \boldsymbol{a}^\top \boldsymbol{b}$ to denote the inner product between two vectors and $\langle \boldsymbol{A}, \boldsymbol{B}\rangle := \mathrm{trace}(\boldsymbol{A}^\top \boldsymbol{B})$ between two matrices.

2. Problem formulation

In this section, we first introduce our high-dimensional contextual bandit model. We compare and contrast it with traditional bandit models, and provide intuition as to why it is suitable for two different applications (assortment-pricing in online retail, and mobile healthcare apps). Finally, we discuss how various structured bandit models can be seen as special cases of our proposal.

Problem setup. Our goal is to learn the action with the highest expected reward based on T samples in a sequential model for data collection. At each time t, we are allowed to choose an action $\boldsymbol{a}_t$ based on the data seen to date. This chosen action applies to a batch of objects of size $L \geq 1$. At each round, we observe a collection of contexts of covariate vectors $\{\boldsymbol{x}_{t,\ell}\}_{\ell=1}^L$, one associated with each of the L objects, and each lying in $\mathbb{R}^{d_x}$. At time t, we are allowed to make a decision based on all observations prior to time t along with the covariates at time t, we decide on an action $\boldsymbol{a}_t$ that takes values in a constraint set $\mathcal{A} \subseteq \mathbb{R}^{d_a}$. After taking action $\boldsymbol{a}_t$ at time t, we observe a batch of rewards

$$y_{t,j} = \boldsymbol{a}_t^\top \boldsymbol{\Theta} \boldsymbol{x}_{t,j} + \varepsilon_{t,j}, \quad \text{for } j = 1, 2, \cdots, L, \tag{1}$$

where $\boldsymbol{\Theta} \in \mathbb{R}^{d_a \times d_x}$ is an unknown low-rank matrix and $\varepsilon_{t,j}$ is independent noise with $\mathbb{E}[\varepsilon_{t,j}] = 0$ and $\mathrm{Var}[\varepsilon_{t,j}] \leq \sigma^2$.

Suppose that we run T rounds of this scheme, using some protocol π for choosing actions; doing so yields a sequence $\{\boldsymbol{a}_{t,\pi}\}_{t=1}^T$ of T actions. We measure the quality of this sequence—and hence the protocol π—via its *expected average regret*[1]

$$\mathcal{R}^\pi(T) = \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^T \sup_{\boldsymbol{a}\in\mathcal{A}_t}\left(\sum_{j=1}^L \boldsymbol{a}^\top \boldsymbol{\Theta}\boldsymbol{x}_{t,j} - \boldsymbol{a}_{t,\pi}^\top \boldsymbol{\Theta}\boldsymbol{x}_{t,j}\right)\right]. \tag{2}$$

In this definition, the expectation is taken with respect to $(\boldsymbol{x}_{t,j}, \varepsilon_{t,j})$ since $\boldsymbol{a}_{t,\pi}$ depends on both. Our goal is to design protocols π with low regret.

[1] Here we have defined the average regret (via our rescaling by T), but bounds on this quantity can immediately be translated to the cumulative regret as needed.

At the core is the model $\mathbb{E}[Y \mid \boldsymbol{x}, \boldsymbol{a}] = \boldsymbol{a}^T \boldsymbol{\Theta} \boldsymbol{x}$ of the mean reward function. It is worth noting that this form of reward function generalizes various known models, including classical K-arm bandits, K-arm bandits with context, and continuum-armed bandits (Robbins, 1952; Bastani & Bayati, 2020; Kleinberg et al., 2019). We discuss these connections in more detail at the end of this section.

The other key ingredient in our model is the low-rank representation matrix $\boldsymbol{\Theta}$. It encapsulates the effect of both the arm and covariates on the reward and exploits the low-dimensional structure in the high-dimensional actions and covariates. To understand this condition, consider a matrix $\boldsymbol{\Theta}$ that is of rank $r \ll \min\{d_a, d_x\}$. It then has a singular value decomposition of the form $\boldsymbol{\Theta} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$, where $\boldsymbol{S} = \mathrm{diag}\{s_1, \ldots, s_r\}$ is a diagonal matrix with the ordered singular values $s_1 \geq s_2 \geq \cdots \geq s_r > 0$, and both $\boldsymbol{U} \in \mathbb{R}^{d_a \times r}$ and $\boldsymbol{V} \in \mathbb{R}^{d_x \times r}$ are matrices with orthonormal columns, corresponding to the left $\{\boldsymbol{u}_j\}_{j=1}^r$ and right singular vectors $\{\boldsymbol{v}_j\}_{j=1}^r$, respectively. With this notation, the reward function (1) can be written as

$$\mathbb{E}[Y \mid \boldsymbol{a}, \boldsymbol{x}] = \boldsymbol{a}^\top \boldsymbol{\Theta} \boldsymbol{x} = \sum_{i=1}^{r} s_i \langle \boldsymbol{a}, \boldsymbol{u}_i \rangle \cdot \langle \boldsymbol{v}_i, \boldsymbol{x} \rangle. \quad (3)$$

In other words, the mean reward is the summation of inner products between the action projected on the left singular vector and the covariates projected on the right singular vector, weighted by the singular values. By assuming $\boldsymbol{\Theta}$ to be low-rank, the mean reward is assumed to be governed by only a few linear combinations of the arm attributes and covariates. Hence our model automatically explores the low-dimensional structure of the arm vector and the contextual vector in terms of its effect on the reward, from which we can draw interpretation and insights from the effective subspaces of both the arm and covariates.

So as to explain why the low-rank assumption is well-motivated in practice, let us discuss some concrete use cases.

Example 1 (Assortment and Pricing). The assortment problem, which arises in retail and e-commerce, is to decide which combination of products to present at each given time while satisfying capacity constraints (Kök et al., 2008). The closely related pricing problem is to decide the prices of the products. A given firm wants to solve these problems so as to maximize a certain objective (e.g., revenue or profit).

In these problems, the action vector consists of both prices and attributes associated with products. Attributes differ in a case-by-case manner: for clothes, the pattern, color and size are standard attributes, whereas for electronics, the technical specifications provide attributes. In our case study, as described in more detail in Section 4, we focus on boxes of instant noodles. Box j has a price p_j, and can contain packages that correspond of one of m total flavors; this can be encoded by an attribute vector $\boldsymbol{f}_j = (f_{j,1}, \ldots, f_{j,m})$ where $f_{j,\ell}$ is the number of packages of flavor ℓ in box j. Given a total of $K \geq 1$ slots in which to present products, a given store needs to decide which products to present, along with their corresponding prices. To formalize this set-up, the action vector takes the form

$$\boldsymbol{a} = (\boldsymbol{f}_1, p_1, \boldsymbol{f}_2, p_2, \cdots, \boldsymbol{f}_K, p_K, 1),$$

and so is high-dimensional. At the same time, we observe covariates associated with the observations; they represent information such as geographic location, seasonal information at the aggregated level, or demographic information at the user level. observe the covariate $\boldsymbol{x}$ for each period of time, such as the location and season at the aggregated level or demographics information at the user level.

The demand and sales of products with similar attributes react similarly to the same market conditions. It is often the case that there exist latent factors of the products that govern the demand and sales. Therefore, it is reasonable to parameterize the reward function (1) rather than ignoring the similarity between products as in the literature (Miao & Chao, 2021; Kallus & Udell, 2020; Chen et al., 2021). Our model can further suggest new products rather than only the products that have already been provided. ♣

Example 2 (Health-care). Bandits are used for health monitoring, which monitors health conditions and give suggestions on actions to take for users, the arm ($\boldsymbol{a}$) is high-dimensional and continuous (e.g., sleeping time, length and kind of exercise, usage of social media, diet choices including energy, water, protein, minerals, and nutrition intakes), and the health outcome depends on not only our suggestions, but also the user's characteristics (e.g., age, gender, weight, height, basic health status, and the tendency of following suggestions) as contextual variables ($\boldsymbol{x}$). Clearly, both the arm and the contextual variable vectors are possibly high-dimensional and the arm can take continuous values. The classical bandit models do not fit the situation. The actions usually share similar effects on health and the user's characteristics can be usually captured by a few latent factors. Therefore, it is reasonable to assume $\boldsymbol{\Theta}$ to be low-rank. ♣

To close this section, let us summarize formally some classical bandit models that are special cases of our reward model (1).

1. Multi-arm bandit: For i-th arm, $\boldsymbol{a} = (0, 0, \cdots, 1, \cdots, 0)$, where 1 is in i-th element. Suppose $\boldsymbol{x}$ has its first element being constant. Then $\boldsymbol{\Theta}_{i,1} = \mu_i$, where μ_i is the mean reward of the i-th arm, and $\boldsymbol{\Theta}_{i,j} = 0$ if $j \neq 1$. Clearly, $\boldsymbol{\Theta}$ has rank 1.
2. Multi-arm high-dimensional contextual bandit: for i-th arm, $\boldsymbol{a} = (0, 0, \cdots, 1, \cdots, 0)$, where 1 is in i-th element. $\boldsymbol{x}$ is the contextual vector. Then $\boldsymbol{\Theta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_m)^\top$, where $\boldsymbol{\beta}_i$ is the parameter vector corresponding to i-th arm (Bastani & Bayati, 2020).

3. Continuum-action bandits (without context): Suppose the arm in the original continuum arm bandit is denoted by a, and the mean reward function is $\mu(a)$. Since all continuous functions on a bounded interval can be approximated by polynomial functions to arbitrary precision, it is reasonable to assume $\mu(a)$ to be polynomial of order n, which is not known precisely and only an upper bound N is known. Let $\boldsymbol{a} = (1, a, a^2, a^3, \cdots, a^n, \cdots, a^N)$, and suppose the first element of $\boldsymbol{x}$ is constant 1, then $\boldsymbol{\Theta}_{i,j} = \frac{1}{i!}f^{(i)}(a)$ for $j = 1$ and $\boldsymbol{\Theta}_{i,j} = 0$ for $j \neq 1$. Clearly, $\boldsymbol{\Theta}$ is rank n.

3. Hi-CCAB algorithm and theoretical results

In this section, we present our learning algorithm with a regret upper bound. Specifically, we detail the Hi-CCAB algorithm in Section 3.1 and establish an upper bound for its convergence rate of the expected regret in Section 3.2.

3.1. Description of the learning algorithm

Our policy consists of two phases for each period $t \in [T]$: the first phase learns a low-rank representation and the second phase determines the assortment and the selling prices. In the first phase, our policy estimates $\widehat{\boldsymbol{\Theta}}_t$ by a penalized least-square estimator using $(\boldsymbol{a}_i, \boldsymbol{x}_{i,j}, y_{i,j})$ for $i = 1, \ldots, t$ and $l = 1, \ldots, L$. Based on $\widehat{\boldsymbol{\Theta}}_t$, we look for the optimal assortment and pricing within the action space $\mathcal{A}_t$. Algorithm 1 describes the detailed procedure of our policy.

Low-rank representation learning. As mentioned in Section 2, both the arm and the contextual vectors $\boldsymbol{a} \in \mathbb{R}^{d_a}$ and $\boldsymbol{x} \in \mathbb{R}^{d_x}$ are high-dimensional, and thus $\boldsymbol{\Theta} \in \mathbb{R}^{d_a \times d_x}$ is also high-dimensional. Fortunately, there often exists structure in both the arm and covariate space as explained in Section 1. To leverage the underlying structure, we impose a low-rank assumption on $\boldsymbol{\Theta}$, which automatically explores the effect of the low-rank structure and the relationships between the action and the contextual arms.

In order to estimate the low-rank representation of $\boldsymbol{\Theta}$ at time t, one could in principle solve the rank-regularized least-squares problem

$$\arg\min_{\boldsymbol{\Theta}} \left\{ \sum_{i=1}^{t} \sum_{j=1}^{L} \left(\boldsymbol{a}_i^\top \boldsymbol{\Theta} \boldsymbol{x}_{i,j} - y_{i,j} \right)^2 + \lambda_t \cdot \text{rank}(\boldsymbol{\Theta}) \right\},$$

where rank$(\boldsymbol{\Theta})$ is the rank function, and $\lambda_t > 0$ is a regularization parameter. Rank penalization leads to a non-convex problem with associated computational challenges, so that it is standard to replace it with the nuclear norm so as to obtain a convex problem. Doing so in our context yields the nuclear-norm regularized estimator

$$\widehat{\boldsymbol{\Theta}}_t := \arg\min_{\boldsymbol{\Theta}} \left\{ \sum_{i=1}^{t} \sum_{j=1}^{L} \left(\boldsymbol{a}_i^\top \boldsymbol{\Theta} \boldsymbol{x}_{i,j} - y_{i,j} \right)^2 + \lambda_t \cdot \|\boldsymbol{\Theta}\|_* \right\}. \tag{4}$$

Algorithm 1 The Hi-CCAB Algorithm.

```
Result: Actions $\boldsymbol{a}_{t_1+1}, \ldots, \boldsymbol{a}_T$.
Input: The number of steps for initialization $t_1$, set of possible actions $\mathcal{A}_{t_1}$, action vectors based on domain knowledge $\{\boldsymbol{a}_i\}_{i=1}^{t_1}$, covariate vectors $\{\boldsymbol{x}_{i,j}\}_{i=1}^{t_1}$, rewards $y_{i,j}$ for $j = 1, \ldots, L$, and exploration parameter $h$.
Initialization: $\lambda_0 \leftarrow \|\frac{1}{2t_1 L}\sum_{i=1}^{t_1}\sum_{j=1}^{L}|\mathbf{a}_i^\top\widehat{\boldsymbol{\Theta}}_{t_1}\mathbf{x}_{i,j} - y_{i,j}|\mathbf{x}_{i,j}\mathbf{a}_i^\top\|_2$, $t \leftarrow t_1$.
while $t < T$ do
    $\lambda_t \leftarrow \lambda_0/\sqrt{t}$;
    Low-rank representation learning:
    $\widehat{\boldsymbol{\Theta}}_t \leftarrow \arg\min_{\boldsymbol{\Theta}} \frac{1}{tL}\sum_{i=1}^{t}\sum_{j=1}^{L}(\boldsymbol{a}_i^\top\boldsymbol{\Theta}\boldsymbol{x}_{i,j} - y_{i,j})^2 + \lambda_t\|\boldsymbol{\Theta}\|_*$;
    Policy learning:
        $\hat{\boldsymbol{a}}_{t+1} \leftarrow \arg\max_{\boldsymbol{a}\in\mathcal{A}_t}\sum_{j=1}^{L}\boldsymbol{a}^\top\widehat{\boldsymbol{\Theta}}_t\boldsymbol{x}_{t+1,j}$;
    if $t \notin \{\lfloor w^{\frac{3}{2}} \rfloor : w \in \mathbb{Z}_+\}$ then
        Exploitation: $\boldsymbol{a}_{t+1} \leftarrow \hat{\boldsymbol{a}}_{t+1}$;
    else
        Exploration: $\boldsymbol{a}_{t+1} \leftarrow \hat{\boldsymbol{a}}_{t+1} + \boldsymbol{\delta}_{t+1}$ where $\boldsymbol{\delta}_{t+1} \sim N(\mathbf{0}_{d_a}, h\boldsymbol{I}_{d_a})$, update action space $\mathcal{A}_{t+1}$;
    end if
    Apply action $\boldsymbol{a}_{t+1}$ and observe reward $y_{t+1,j}$ for $j = 1, \ldots, L$;
    $t \leftarrow t + 1$;
end while
```

The penalization parameter λ_t is updated in each iteration according to the schedule $\lambda_t = \lambda_0/\sqrt{t}$, where λ_0 is the initialized penalization parameter, which can be chosen by cross-validation or guided by $\|\frac{1}{2t_1 L}\sum_{i=1}^{t_1}\sum_{j=1}^{L}|\boldsymbol{a}_i^\top\widehat{\boldsymbol{\Theta}}_{t_1}\boldsymbol{x}_{i,j} - y_{i,j}|\boldsymbol{x}_{i,j}\boldsymbol{a}_i^\top\|_2$.

Policy learning. Once we estimated the low-rank representation of $\boldsymbol{\Theta}$, we can proceed to the action step. The goal of the action step is to *exploit* the knowledge we have learned, i.e., $\widehat{\boldsymbol{\Theta}}_t$, so as to decide on the next action $\boldsymbol{a}_{t+1}$ that maximizes the reward, and at the same time to *explore* actions that better inform the true $\boldsymbol{\Theta}$, which in turn will help make better decisions to achieve higher long-term rewards. Specifically, given $\widehat{\boldsymbol{\Theta}}_t$ and the covariate $\boldsymbol{x}_{t+1,j}$ for $j = 1, \ldots, L$, we look for an action $\hat{\boldsymbol{a}}_{t+1}$ in the action space $\mathcal{A}_t$ that maximizes the total rewards across L objects:

$$\hat{\boldsymbol{a}}_{t+1} := \arg\max_{\boldsymbol{a}\in\mathcal{A}_t} \left\{ \sum_{j=1}^{L} \boldsymbol{a}^\top \widehat{\boldsymbol{\Theta}}_t \boldsymbol{x}_{t+1,j} \right\}. \tag{5}$$

At a subset of times, we further perturb $\hat{\boldsymbol{a}}_{t+1}$ for the purpose of exploration by adding random noise to each coordinate as follows: $\boldsymbol{a}_{t+1} = \hat{\boldsymbol{a}}_{t+1} + \boldsymbol{\delta}_{t+1}$ where $\boldsymbol{\delta}_{t+1} \sim N(\mathbf{0}_{d_a}, h\boldsymbol{I}_{d_a})$ and h is a tuning parameter. In our current algorithm, we perform this perturbation at times $t \in \{\lfloor w^{\frac{3}{2}} \rfloor : w \in \mathbb{Z}_+\}$.

The intuition for this particular choice ($\lfloor w^{\frac{3}{2}} \rfloor$) is to explore more in the initial stage and exploit more in the later stage of the algorithm. To be specific, there are approximately $T^{\frac{2}{3}}$ steps for exploration before time T. The density of exploration at a small time frame around T is $T^{-\frac{1}{3}}$, which goes to zero as $T \to \infty$. Note that the exponent need not be $\frac{3}{2}$, but can be any number strictly larger than 1; this choice will affect the decay rate of the regret, as will be discussed later in Remark 3.3.

The polynomial form can be changed as well. For each exploration step, one can also let $\boldsymbol{\delta}_{t+1} \sim N(\mathbf{0}_{d_a}, diag(\hat{\boldsymbol{\tau}}_t))$ where each element of $\hat{\boldsymbol{\tau}}_t$ is the coordinate-wise standard error of the previous actions $\{\boldsymbol{a}_i\}_{i=1}^{t}$. The intuition is to avoid tuning parameter h while taking the right scale. Finally, we update the action space $\mathcal{A}_{t+1}$ according to $\boldsymbol{a}_{t+1}$. For example, if the action space $\mathcal{A}_t \in \mathbb{R}^{d_a}$ can be defined by an upper limit $\bar{\boldsymbol{a}}_t$ and a lower limit $\underline{\boldsymbol{a}}_t$, then we simply expand the action space by pushing the boundary of each coordinate to $\boldsymbol{a}_{t+1,j}$ if $\boldsymbol{a}_{t+1,j} \notin [\underline{\boldsymbol{a}}_{t,j}, \bar{\boldsymbol{a}}_{t,j}]$ for $j = 1, \ldots, d_a$.

Remark 3.1. To take advantage of the interpretability of our model, we can further explore the structure of the $\widehat{\boldsymbol{\Theta}}_t$. Specifically, we can apply singular value decomposition (SVD) on $\widehat{\boldsymbol{\Theta}}_t$ to explore the underlying latent structure of the covariates from the right singular vectors; and apply SVD on $(\widehat{\boldsymbol{\Theta}}_t \sum_j^L \boldsymbol{x}_{t,j})$ to explore the latent structure of the arms from the left singular vectors. One can further rotate the singular vectors so as to reveal the underlying factors using techniques in factor analysis such as Varimax (Kaiser, 1958; Rohe & Zeng, 2020) or to perform clustering analysis by performing K-means on the singular vectors.

3.2. Theoretical Results

In this section, we now state a theorem that provides a bound on the expected regret associated with Algorithm 1. It shows that in the worst-case and for any dimensions, the expected regret decays to zero as $T^{-2/15}$.

Our analysis applies to an instantiation of Algorithm 1 in which the exploratory actions are chosen as

$$\boldsymbol{a}_t = \hat{\boldsymbol{a}}_t + \boldsymbol{\delta}_t \qquad \text{for each } t \in \{\lfloor w^{\frac{3}{2}} \rfloor : w \in \mathbb{Z}_+\},$$

where $\boldsymbol{\delta}_t \sim N(\mathbf{0}_{d_a}, h\boldsymbol{I}_{d_a})$, $\mathcal{A}_t = \{\boldsymbol{a} \in \mathbb{R}^{d_a} : \|\boldsymbol{a}\| \le 1\}$. Finally, our statement involves a burn-in period $B_{\text{init}} = C_{h,L,\lambda_0}(d_x + d_a)^6 \left(\log(d_x + d_a)\right)^3$.

Theorem 3.2. *Suppose that $\boldsymbol{\Theta}$ has rank r, we observe covariates $\boldsymbol{x}_{t,l} \overset{i.i.d}{\sim} N(\mathbf{0}_{d_x}, \boldsymbol{I}_{d_x})$, and the reward errors $\varepsilon_{t,j} \overset{i.i.d}{\sim} N(0, \sigma^2)$ in equation (1). Then there are universal constants $\{c_j\}_{j=1}^3$ such that for all $T \ge B_{init}$, the expected regret is bounded as*

$$\mathcal{R}^\pi(T) \le \frac{c_1}{T}\sqrt{Ld_x}\|\boldsymbol{\Theta}\|_2 B_{init} + \frac{c_2}{T^{1/6}}\lambda_0 \frac{\sqrt{2rd_xL}}{h^2} + \frac{c_3}{T^{2/15}}\Big\{\sqrt{Ld_x}\|\boldsymbol{\Theta}\|_2 + \frac{\sigma(d_x+1)}{h^2}\Big\}. \tag{6}$$

Remark 3.3 (Convergence rate). An intuitive understanding of Theorem 3.2 is that the expected regret converges to zero at least as quickly as $T^{-\frac{2}{15}}$ as T tends to infinity. The convergence rate depends on the frequency of the exploration which depends on the exponent $\frac{3}{2}$ in the exploration set, $\{\lfloor w^{\frac{3}{2}} \rfloor : w \in \mathbb{Z}_+\}$. Recall that the exponent can be changed with any number larger than 1, which can be considered as a tuning parameter.

Remark 3.4 ("Burn-in" term). The first term in the bound (6) is a "burn-in" term, where the algorithm is gaining knowledge of $\boldsymbol{\Theta}$ from scratch. We do not impose any assumptions on these starting steps so that we have a relatively conservative "burn-in" term. However, in practice, we usually have historical data to start with so that we can start from a reasonable estimation of $\boldsymbol{\Theta}$ and a much smaller "burn-in" term. Recall that the exponent of the exploration set can be any number larger than 1. The order of the "burn-in" term depends on the exponent of the w in the exploration set — the more exploration there is, the smaller the "burn-in" term. The exponent can be chosen depending on the situation — how ample the historical data is.

Remark 3.5 (Constant C_{h,L,λ_0} of B_{init}). While constant C_{h,L,λ_0} depends on h, L, λ_0, the primary dependency is actually on h and L. The order of λ_0 in terms of dimensions and noise level is $\sigma\sqrt{d_x}$. We do not assume the order of λ_0 or bound it with a high probability bound in order to show its role in time-averaged expected cumulative regret. If we utilize the order $\sigma\sqrt{d_x}$, then C_{h,L,λ_0} can be replaced by a constant depending on h and L only.

Remark 3.6 (Dependence on dimensions d_a, d_x and rank r). When T is small, the "burn-in" term (the first term) dominates. It depends on T and the dimensions but not the rank as $(d_a + d_x)^6(\log(d_a + d_x))^3 T^{-1}$, whose order depends on the exponent defining the exploration set (i.e., how frequently we explore). As T grows, the second term dominates. Recall Remark 3.5, λ_0 is of order $\sigma\sqrt{d_x}$, so the second terms depends on T, d_x and r but not d_a at the order of $\Omega(d_x\sqrt{r}T^{-\frac{1}{6}})$. Without the low-rank assumption, the order would be $\Omega(d_x^{\frac{3}{2}}T^{-\frac{1}{6}})$ instead. When T becomes even larger, the last two terms dominate, at the order $\Omega(d_xT^{-\frac{2}{15}})$. However, the last case rarely happens, as it requires the order of T equal to or larger than d_x^{15}. Therefore, taking dimensions and rank into consideration, the expected regret is mostly at the order of $\Omega(d_x\sqrt{r}T^{-\frac{1}{6}})$.

Proof sketch Due to space constraints, we limit ourselves to an outline of the proof of Theorem 3.2. There are two major steps: (1) bounding the estimation error for the low-rank representation matrix estimator; (2) bounding the expected regret. See Appendix A for the full proof.

(1) *High-probability bound on the estimation error of $\widehat{\boldsymbol{\Theta}}_t$:* Introduce the shorthand $\boldsymbol{\Delta}_t = \widehat{\boldsymbol{\Theta}}_t - \boldsymbol{\Theta}$. We show that

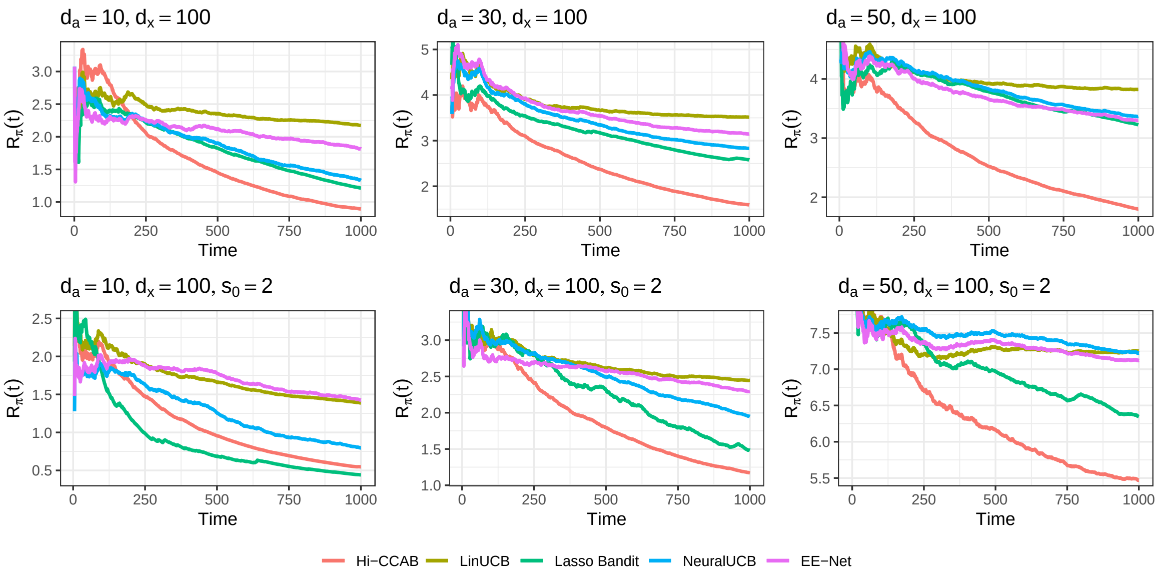


Figure 1: The expected average regret under the non-sparse (first row) and sparse (second row) settings.

for a large t,

$$\mathbb{P}\left(\|\boldsymbol{\Delta}_t\|_F \leq \frac{3}{T^{\frac{2}{15}}}\frac{\sigma\sqrt{d_x+1}}{\sqrt{L}h^2} + 6\lambda_0\frac{\sqrt{2d}}{h^2T^{\frac{1}{6}}}\right) \leq 1-\left(\frac{3}{t}+\frac{2}{t^2}+\frac{2}{Lt}+\frac{2}{L^3t^3}+\frac{1}{t^{\frac{2}{15}}}\right). \tag{7}$$

Note that the action taken is based on previous estimators and affects the accuracy of future estimators, leading to lots of dependencies. The classical matrix completion results can no longer apply. Through careful use of conditional expectations, martingales, and empirical process we separate out different sources of randomness (i.e., $\delta_1, \cdots, \delta_t, \boldsymbol{x}_{1,\cdot}, \cdots, \boldsymbol{x}_{t,\cdot}$) to derive the bounds. Lemma A.1 establishes a restricted-strong-convexity-type result of the sum of squares in the objective function. Lemma A.2 establishes a Lipschitz-type result of the sum of squares in the objective function. Further analysis of the nuclear-norm-penalized sum of squares with the two lemmas and low-rank properties gives the tail bound of the estimation error.

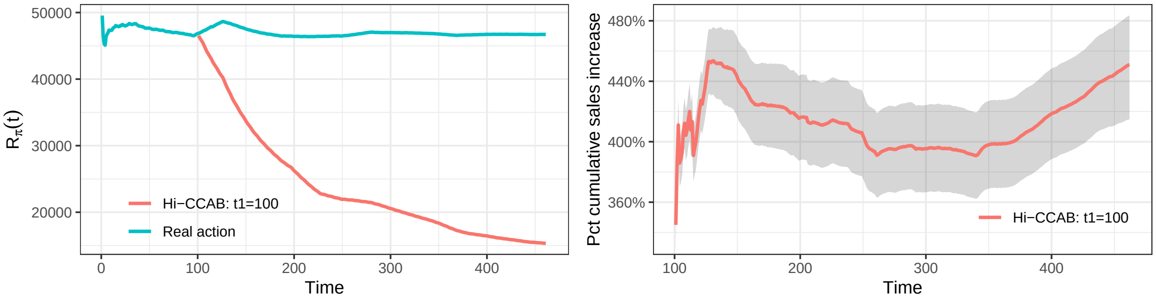
(2) *Bounding the expected regret:* At each round t, we define the event $\mathcal{E}_t = \{\|\boldsymbol{\Delta}_t\|_F \leq \frac{3}{T^{\frac{2}{15}}}\frac{\sigma\sqrt{d_x+1}}{\sqrt{L}h^2} + 6\lambda_0\frac{\sqrt{2r}}{h^2T^{\frac{1}{6}}}\}$. From the first step, we know for large t, $\mathbb{P}(\mathcal{E}_t^c) \leq \frac{3}{t}+\frac{2}{t^2}+\frac{2}{Lt}+\frac{2}{L^3t^3}+\frac{1}{t^{\frac{2}{15}}}$. Consider the expectation of the regret on $\mathcal{E}_t$ and $\mathcal{E}_t^c$ separately and both terms vanish with t at the polynomial rate.

4. Experimental evaluations

In this section, we report some experimental evaluations of both synthetic and real-world datasets. First, we conduct simulation studies to compare the proposed `Hi-CCAB` with LinUCB (Li et al., 2010), Lasso Bandit (Bastani & Bayati, 2020), NeuralUCB (Zhou et al., 2020) and EE-Net (Ban et al., 2022); we then study the joint assortment-pricing problem on the e-commerce platform for one of the largest instant noodles producers in China. Details on the tuning parameters of each algorithm and additional results of the case study are provided in Appendices B–C.

Simulation study We consider the multi-armed linear bandit setup, a special case of our model with $\boldsymbol{\Theta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_m)^\top$ so that each row of $\boldsymbol{\Theta}$ is the parameter of each arm for the multi-arm contextual bandit. Specifically, we set the number of arms $d_a = \{10, 30, 50\}$ and the dimension of covariates $d_x = 100$. For $\boldsymbol{\Theta}$, we consider a non-sparse and sparse case. For the non-sparse case, we generate $\boldsymbol{\Theta} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ where $\boldsymbol{U} \in \mathbb{R}^{d_a\times r}, \boldsymbol{V} \in \mathbb{R}^{d_x\times r}$ ($r = 5$), and $\boldsymbol{D}$ is a diagonal matrix with $(1, .9, .9, .8, .5)$ as the diagonal entries. All entries of $\boldsymbol{U}$ and $\boldsymbol{V}$ are first generated from i.i.d. $N(0, 1)$, and then applied Gram–Schmidt to make each column orthogonal. The matrix $\boldsymbol{U}$ is scaled to have length $\sqrt{d_a}$ so that the rewards are comparable across different d_a's. For the sparse case, each row of $\boldsymbol{\Theta}$ is set as zero except for $s_0 = 2$ randomly selected elements that are drawn from $N(0, 1)$. We generate the covariate $\boldsymbol{x} \overset{i.i.d}{\sim} N(0, \boldsymbol{I}_{d_x})$ and the rewards from (1) with $\sigma = 0.1$.

Figure 1 shows the regret (averaged over 50 simulations). For the non-sparse case, `Hi-CCAB` converges faster than all other methods. The advantage of `Hi-CCAB` is more pronounced when the dimension of arms becomes larger. For the sparse case, which is not to the advantage of `Hi-CCAB`, when the dimension of arms is relatively small ($d_a = 10$), Lasso Bandit converges faster but the gap between `Hi-CCAB` and Lasso Bandit is small. As the number of arms increases, `Hi-CCAB` outperforms all other methods.



(a) Expected average regret.

(b) Percentage cumulative sales gain.

Figure 2. Performance of Hi−CCAB compared with real actions. The boundaries of the shadow are the 5-th and 95-th quantiles.

Assortment-pricing case study. The original data contains daily sales of 176 products across 369 cities from March 1st, 2021 to May 31st, 2022 ($T = 456$ days). We aggregate the sales by 31 provinces. Each product is of either single or assorted flavors (13 possible flavors) with different counts. The assortment and price of each product changed daily. We also know the dates for promotion. The assortment, prices, and promotions were the same across locations. The maximum number of products to be shown on the homepage is $K = 30$. The total possible combinations are $\binom{176}{30}$ and therefore if we consider one combination as one arm, the arm space is extremely high-dimensional, for which most multi-arm bandit algorithms are not applicable.

To apply Hi−CCAB, we specify the arms $\boldsymbol{a}_t$ and the covariate vectors $\{\boldsymbol{x}_{t,j}\}_{j=1}^{L=31}$ at given time t following the setup in Example 1. The arm is represented as $\boldsymbol{a} = (\boldsymbol{f}_1, \boldsymbol{f}_1^2, p_1, p_1^2, promo_1, promo_1^2, \cdots, \boldsymbol{f}_K, \boldsymbol{f}_K^2, p_K, p_K^2, promo_K, promo_K^2, 1) \in \mathbb{R}^{2(m+2)K+1=901}$ where $\boldsymbol{f}_k = (f_{k,1}, \cdots, f_{k,m})$ is a vector of non-negative integers to denote the counts of $m = 13$ flavors, p_k is the price, $promo_k$ is the indicator of promotion of product k, and $\boldsymbol{f}_k^2$ is the element-wise quadratics. The covariate $\boldsymbol{x}_{t,j} \in \mathbb{R}^{50}$ for location j includes dummy variables of 31 provinces, the year 2021/2022, 12 months, weekdays, and an indicator of the annual sales event on Jun 18 and Nov 11. More details are deferred to Appendix C.

To run simulations using the dataset, we first create a pseudo-truth model. To be specific, we estimate $\boldsymbol{\Theta}$ and σ using all data of 456 days and consider them as the pseudo-ground truth. We perform a sanity check on our model assumption (1), the pseudo ground truth against our data before preceding the formal analysis and further examine the structure of the representation matrix $\boldsymbol{\Theta}$ in Appendix C. We evaluate the performance of Hi−CCAB in terms of the cumulative regret (2) and the percentage gain of the cumulative sales by comparing with the original actions, since no existing bandit algorithm is applicable to this problem.

Figure 2a shows the time-averaged cumulative regret (averaged over 100 simulations) and Figure 2b shows the percentage gain in cumulative sales compared to the real sales The expected average regret of Hi−CCAB converges to zero while that of original actions remains flat. In terms of percentage gain in cumulative sales, Hi−CCAB boosts cumulative sales by more than 4 times. On a separate note, Hi−CCAB with exploration performs better in terms of both regret and percentage sales gain than Hi−CCAB without exploration.

5. Discussion

With an increasing demand for online decision-making, the bandit problem is receiving increasingly more attention from both theoreticians and practitioners. Despite the richness of the bandit literature, to date, there has been relatively little work on contextual bandits in which both the covariate and action spaces are high-dimensional. In this paper, we have argued that many applications of bandit have this “doubly” high-dimensional nature, and we have provided a structured matrix model for capturing interactions between covariates and actions. This model is reasonably general, including a number of structured bandit models as special cases, but also interpretable. We propose an efficient algorithm Hi−CCAB that interleaves steps of low-rank matrix estimation with exploration/exploitation, and we proved a non-asymptotic upper bound on its expected regret. The generality and flexibility of our model allow for its application in the joint assortment-pricing problem, where the operations research community has studied the assortment and pricing optimization problems extensively, but not as a joint optimization problem. By applying our model and algorithm to a real case study on the joint assortment-pricing problem for one of the largest instant noodles producers in China, our method can boost sales by a factor of four times, while also provide insights into the underlying structure of the effect on the reward of the arms and covariates such as purchasing behaviors. As for future directions, since our model is new to the bandit literature, there remains space for improvement in our regret analysis – we may further tighten the regret bound and establish a possibly matching lower bound.

References

Abbasi-Yadkori, Y., Pal, D., and Szepesvari, C. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Artificial Intelligence and Statistics*, pp. 1–9. PMLR, 2012.

Agrawal, R. The continuum-armed bandit problem. *SIAM journal on control and optimization*, 33(6):1926–1951, 1995.

Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. Mnl-bandit: A dynamic learning approach to assortment selection. *Operations Research*, 67(5):1453–1485, 2019.

Ban, G.-Y. and Keskin, N. B. Personalized dynamic pricing with machine learning: High-dimensional features and heterogeneous elasticity. *Management Science*, 67(9): 5549–5568, 2021.

Ban, Y., Yan, Y., Banerjee, A., and He, J. Ee-net: Exploitation-exploration neural networks in contextual bandits. 2022.

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

Cai, T. T. and Zhang, A. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1): 60–89, 2018.

Candes, E. J. and Plan, Y. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

Caro, F. and Gallien, J. Dynamic assortment with demand learning for seasonal consumer goods. *Management science*, 53(2):276–292, 2007.

Chen, N. and Gallego, G. Nonparametric pricing analytics with customer covariates. *Operations Research*, 69(3): 974–984, 2021.

Chen, X. and Wang, Y. A note on a tight lower bound for mnl-bandit assortment selection models. *arXiv preprint arXiv:1709.06109*, 2017.

Chen, X., Shi, C., Wang, Y., and Zhou, Y. Dynamic assortment planning under nested logit models. *Production and Operations Management*, 30(1):85–102, 2021.

Chen, Y., Xie, M., Liu, J., and Zhao, K. Interconnected neural linear contextual bandits with ucb exploration. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 169–181. Springer, 2022.

Debon, R., Coleone, J. D., Bellei, E. A., and De Marchi, A. C. B. Mobile health applications for chronic diseases: A systematic review of features for lifestyle improvement. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 13(4):2507–2512, 2019.

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

Féraud, R., Allesiardo, R., Urvoy, T., and Clérot, F. Random forest for the contextual bandit problem. In *Artificial intelligence and statistics*, pp. 93–101. PMLR, 2016.

Hao, B., Lattimore, T., and Wang, M. High-dimensional sparse linear bandits. *Advances in Neural Information Processing Systems*, 33:10753–10763, 2020.

Hu, J., Chen, X., Jin, C., Li, L., and Wang, L. Near-optimal representation learning for linear bandits and linear rl. In *International Conference on Machine Learning*, pp. 4349–4358. PMLR, 2021.

Jun, K.-S., Willett, R., Wright, S., and Nowak, R. Bilinear bandits with low-rank structure. In *International Conference on Machine Learning*, pp. 3163–3172. PMLR, 2019.

Kaiser, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, 1958.

Kallus, N. and Udell, M. Dynamic assortment personalization in high dimensions. *Operations Research*, 68(4): 1020–1037, 2020.

Keskin, N. B. and Zeevi, A. Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations research*, 62(5):1142–1167, 2014.

Kim, G.-S. and Paik, M. C. Doubly-robust lasso bandit. *Advances in Neural Information Processing Systems*, 32, 2019.

Kim, J.-h. and Vojnovic, M. Scheduling servers with stochastic bilinear rewards. *arXiv preprint arXiv:2112.06362*, 2021.

Kleinberg, R. Nearly tight bounds for the continuum-armed bandit problem. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS'04, pp. 697—704, Cambridge, MA, USA, 2004. MIT Press.

Kleinberg, R., Slivkins, A., and Upfal, E. Bandits and experts in metric spaces. *J. ACM*, 66(4), May 2019. ISSN 0004-5411. doi: 10.1145/3299873. URL https://doi.org/10.1145/3299873.

Kök, A. G., Fisher, M. L., and Vaidyanathan, R. Assortment planning: Review of literature and industry practice. *Retail supply chain management*, 122(1):99–153, 2008.

Krishnamurthy, A., Langford, J., Slivkins, A., and Zhang, C. Contextual bandits with continuous actions: Smoothing, zooming, and adapting. *The Journal of Machine Learning Research*, 21(1):5402–5446, 2020.

Kveton, B., Szepesvári, C., Rao, A., Wen, Z., Abbasi-Yadkori, Y., and Muthukrishnan, S. Stochastic low-rank bandits. *arXiv preprint arXiv:1712.04644*, 2017.

Lale, S., Azizzadenesheli, K., Anandkumar, A., and Hassibi, B. Stochastic linear bandits with hidden low rank structure. *arXiv preprint arXiv:1901.09490*, 2019.

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

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

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

Lu, Y., Meisami, A., and Tewari, A. Low-rank generalized linear bandit problems. In *International Conference on Artificial Intelligence and Statistics*, pp. 460–468. PMLR, 2021.

Miao, S. and Chao, X. Dynamic joint assortment and pricing optimization with demand learning. *Manufacturing & Service Operations Management*, 23(2):525–545, 2021.

Negahban, S. and Wainwright, M. J. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011.

Papini, M., Tirinzoni, A., Restelli, M., Lazaric, A., and Pirotta, M. Leveraging good representations in linear contextual bandits. In *International Conference on Machine Learning*, pp. 8371–8380. PMLR, 2021.

Qiang, S. and Bayati, M. Dynamic pricing with demand covariates. *arXiv preprint arXiv:1604.07463*, 2016.

Recht, B., Fazel, M., and Parrilo, P. A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

Rizk, G., Thomas, A., Colin, I., Laraki, R., and Chevaleyre, Y. Best arm identification in graphical bilinear bandits. In *International Conference on Machine Learning*, pp. 9010–9019. PMLR, 2021.

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

Rohe, K. and Zeng, M. Vintage factor analysis with varimax performs statistical inference. *arXiv preprint arXiv:2004.05387*, 2020.

Sauré, D. and Zeevi, A. Optimal dynamic assortment planning with demand learning. *Manufacturing & Service Operations Management*, 15(3):387–404, 2013.

Slivkins, A. Contextual bandits with similarity information. In *Proceedings of the 24th annual Conference On Learning Theory*, pp. 679–702. JMLR Workshop and Conference Proceedings, 2011.

Srebro, N., Alon, N., and Jaakkola, T. S. Generalization error bounds for collaborative prediction with low-rank matrices. In *Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2005.

Tropp, J. A. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12 (4):389–434, 2012.

Turğay, E., Bulucu, C., and Tekin, C. Exploiting relevance for online decision-making in high-dimensions. *IEEE Transactions on Signal Processing*, 69:1438–1451, 2020.

Tyagi, H., Stich, S. U., and Gärtner, B. On two continuum armed bandit problems in high dimensions. *Theory of Computing Systems*, 58(1):191–222, 2016.

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

Xu, K. and Bastani, H. Learning across bandits in high dimension via robust statistics. *arXiv preprint arXiv:2112.14233*, 2021.

Xu, P., Wen, Z., Zhao, H., and Gu, Q. Neural contextual bandits with deep representation and shallow exploration. 2022.

Yang, J., Hu, W., Lee, J. D., and Du, S. S. Impact of representation learning in linear bandits. *arXiv preprint arXiv:2010.06531*, 2020.

Zhou, D., Li, L., and Gu, Q. Neural contextual bandits with ucb-based exploration. In *International Conference on Machine Learning*, pp. 11492–11502. PMLR, 2020.

A. Proof of Theorem 3.2

In this proof, we denote the true parameter as $\boldsymbol{\Theta}^*$.

Let $\mathcal{L}_T(\boldsymbol{\Theta}) := \frac{1}{2LT}\sum_{t=1}^{T}\sum_{l=1}^{L}(\boldsymbol{a}_{t,l}^\top \boldsymbol{\Theta}\boldsymbol{x}_{t,l} - y_{t,l})^2$. Then we have the following lemmas that we will prove later.

Lemma A.1. *Suppose all the assumptions in Theorem 3.2 holds. Denote $\mathcal{E}_T(\Delta) = \mathcal{L}_T(\boldsymbol{\Theta}^* + \Delta) - \mathcal{L}_T(\boldsymbol{\Theta}^*) - \langle \nabla \mathcal{L}_T(\boldsymbol{\Theta}^*), \Delta\rangle$. Then with probability at least $1 - \frac{1}{LT} - \frac{2}{T} - \frac{1}{T^2}$,*

$$\mathcal{E}_T(\Delta) \geq \frac{\lfloor T^{\frac{2}{3}} \rfloor}{2T} h^2 \|\Delta\|_F^2 - 14T^{-\frac{2}{3}}\left(h + h^2\right)\left(2d_x + 2d_a + 6\log T + 6\log L\right)^2 \log T \|\Delta\|_2^2. \tag{8}$$

Lemma A.2. *Suppose all the assumptions in Theorem 3.2 holds. With probability at least $1 - \frac{1}{T^{\frac{2}{15}}} - \frac{2}{L^3T^3} - \frac{1}{LT} - \frac{1}{T} - \frac{1}{T^2}$, the following holds for all Δ*

$$\begin{aligned}|\langle \nabla \mathcal{L}_T(\boldsymbol{\Theta}^*), \Delta\rangle| \leq & \|\Delta\|_F \frac{\sigma\sqrt{d_x+1}}{\sqrt{LT}} T^{\frac{1}{30}} + \\ & \Big(2h\sigma T^{-2/3}\log T \sqrt{\frac{\max\{d_a, d_x\}\log\left(d_a + d_x\right)}{L}} + \\ & \frac{8h\sigma}{T}\sqrt{\log\left(TL\right)}\sqrt{(d_x + 3\log(LT))(d_a + 3\log T)}(\log(d_x+d_a) + 2\log T)\Big)\|\Delta\|_*.\end{aligned} \tag{9}$$

Recall the definition of $\hat{\boldsymbol{\Theta}}_t$, we know that

$$\mathcal{L}_T(\hat{\boldsymbol{\Theta}}_T) + \lambda_T \|\hat{\boldsymbol{\Theta}}_T\|_* \leq \mathcal{L}_T(\boldsymbol{\Theta}^*) + \lambda_T\|\boldsymbol{\Theta}^*\|_*. \tag{10}$$

Denote $\boldsymbol{\Delta}_t = \hat{\boldsymbol{\Theta}}_t - \boldsymbol{\Theta}^*$ and for notation simplicity we will drop the subscript t for $\boldsymbol{\Delta}_t$ in the following when there is no confusion. Equation (10) then implies that

$$\mathcal{E}_T(\boldsymbol{\Delta}) \leq -\langle \nabla\mathcal{L}_T(\boldsymbol{\Theta}^*), \boldsymbol{\Delta}\rangle + \lambda_T\left(\|\boldsymbol{\Theta}^*\|_* - \|\boldsymbol{\Theta}^* + \boldsymbol{\Delta}\|_*\right). \tag{11}$$

Suppose the singular value decomposition of $\boldsymbol{\Theta}^*$ is $\boldsymbol{\Theta}^* = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$, where $\boldsymbol{S}$ is an $r \times r$ diagonal matrix. Let $\boldsymbol{U}_\top$ be an $d_a \times (d_a - r)$ matrix satisfying $(\boldsymbol{U}, \boldsymbol{U}_\perp)(\boldsymbol{U}, \boldsymbol{U}_\perp)^\top = \boldsymbol{I}_{d_a}$. We define $\boldsymbol{V}_\perp$ similarly.

Denote $\boldsymbol{\Delta}_\perp = \boldsymbol{U}_\perp^\top \boldsymbol{\Delta}\boldsymbol{V}_\perp$. Then $\|\boldsymbol{\Theta}^* + \boldsymbol{\Delta}\|_* \geq \|\boldsymbol{\Theta}^* + \boldsymbol{\Delta}_\perp\|_* - \|\boldsymbol{\Delta} - \boldsymbol{\Delta}_\perp\|_* = \|\boldsymbol{\Theta}^*\|_* + \|\boldsymbol{\Delta}_\perp\|_* - \|\boldsymbol{\Delta} - \boldsymbol{\Delta}_\perp\|_* \geq \|\boldsymbol{\Theta}^*\|_* + \|\boldsymbol{\Delta}_\perp\|_* - \sqrt{2r}\|\boldsymbol{\Delta} - \boldsymbol{\Delta}_\perp\|_F$.

Going back to inequality (11), and combing with Lemma A.1 and Lemma A.2, we have, with probability at least $1 - \frac{3}{T} - \frac{2}{T^2} - \frac{2}{LT} - \frac{2}{L^3T^3} - \frac{1}{T^{\frac{1}{3}}}$, the following holds

$$\begin{aligned}&\left(\frac{\lfloor T^{\frac{2}{3}}\rfloor}{2T}h^2 - 14T^{-\frac{2}{3}}\left(h + h^2\right)\left(2d_x + 2d_a + 6\log T + 6\log L\right)^2\log T\right)\|\boldsymbol{\Delta}\|_F^2 \\ \leq & \|\boldsymbol{\Delta}\|_F\frac{\sigma\sqrt{d_x+1}}{\sqrt{LT}}T^{\frac{1}{30}} + \\ & \Big(\frac{8h\sigma}{T}\sqrt{\log\left(TL\right)}\sqrt{(d_x + 3\log(LT))(d_a + 3\log T)}(\log(d_x + d_a) + 2\log T) + \\ & 2h\sigma T^{-2/3}\log T\sqrt{\frac{\max\{d_a, d_x\}\log\left(d_a + d_x\right)}{L}}\Big)\left(\|\boldsymbol{\Delta} - \boldsymbol{\Delta}_\perp\|_* + \|\boldsymbol{\Delta}_\perp\|_*\right) \\ & + \lambda_0\frac{\sqrt{T}}{T}\sqrt{2r}\|\boldsymbol{\Delta}\|_F - \lambda_0\frac{\sqrt{T}}{T}\|\boldsymbol{\Delta}_\perp\|_*.\end{aligned} \tag{12}$$

Note that $\|\boldsymbol{\Delta} - \boldsymbol{\Delta}_\perp\|_* \leq \sqrt{2r}\|\boldsymbol{\Delta} - \boldsymbol{\Delta}_\perp\|_F$, divide both side with $\|\boldsymbol{\Delta}\|_F$ and multiply both sides with $3T^{\frac{1}{3}}/h^2$. Suppose

B_{init} satisfies

$$\begin{aligned} B_{\text{init}} \geq & 8, \\ B_{\text{init}}^{\frac{1}{3}} \geq & 12 \times 14(1+\frac{1}{h})\left(2d_x + 2d_a + 6\log B_{\text{init}} + 6\log L\right)^2 \log B_{\text{init}}, \\ \lambda_0 B_{\text{init}}^{\frac{1}{6}} \geq & \frac{8h\sigma}{B_{\text{init}}^{\frac{1}{3}}}\sqrt{\log\left(B_{\text{init}}L\right)}\sqrt{(d_x + 3\log(LT))(d_a + 3\log B_{\text{init}})}(\log(d_x+d_a) + 2logB_{\text{init}}) + \\ & 2h\sigma B_{\text{init}}^{-2/3}\log B_{\text{init}}\sqrt{\frac{\max\{d_a, d_x\}\log\left(d_a+d_x\right)}{L}} \end{aligned} \tag{13}$$

we have

$$\|\delta\mathbf{\Theta}\|_F \leq \frac{3}{T^{\frac{2}{15}}}\frac{\sigma\sqrt{d_x+1}}{\sqrt{L}h^2} + 6\lambda_0\frac{\sqrt{2r}}{h^2T^{\frac{1}{6}}}. \tag{14}$$

Note that there is a constant C_{h,L,λ_0} depending on L, h and λ_0 such that for

$$B_{\text{init}} \geq C_{h,L,\lambda_0}(d_x+d_a)^6\left(\log(d_x+d_a)\right)^3,$$

inequalities (13) holds.

Next we will proceed to bound the regret. Denote the event that equation (14) holds to be Q_t and its complement as Q_t^c. Then $\mathbb{P}(Q_t^c) \leq \frac{3}{T} + \frac{2}{T^2} + \frac{2}{LT} + \frac{2}{L^3T^3} + \frac{1}{T^{\frac{2}{15}}}$ and $\left(Q_t^c, \hat{\mathbf{\Theta}}_t\right) \perp\!\!\!\perp \boldsymbol{b}_{t+1}$. Let the oracle optimal action at time t be $\boldsymbol{a}_t^*$ and $\boldsymbol{b}_t = \sum_{l=1}^L \boldsymbol{x}_{t,l}$. Then

$$\begin{aligned} & T\mathcal{R}^\pi(T) - \mathbb{E}\left(\sum_{t=0}^{B_{\text{init}}-1}\sum_{l=1}^{L}\left(\boldsymbol{a}_{t+1}^{*\top}\mathbf{\Theta}^*\boldsymbol{x}_{t+1,l} - \boldsymbol{a}_{t+1}^{\top}\mathbf{\Theta}^*\boldsymbol{x}_{t+1,l}\right)\right) \\ & \leq \mathbb{E}\left(\sum_{t=B_{\text{init}}}^{T-1}\sum_{l=1}^{L}\boldsymbol{a}_{t+1}^{*\top}\mathbf{\Theta}^*\boldsymbol{x}_{t+1,l} - \boldsymbol{a}_{t+1}^{\top}\mathbf{\Theta}^*\boldsymbol{x}_{t+1,l}\right) \\ & \leq \mathbb{E}\left(\sum_{t=B_{\text{init}}}^{T-1}\left\langle\frac{\mathbf{\Theta}^*\boldsymbol{b}_{t+1}}{\|\mathbf{\Theta}^*\boldsymbol{b}_{t+1}\|_2} - \frac{\hat{\mathbf{\Theta}}_t\boldsymbol{b}_{t+1}}{\|\hat{\mathbf{\Theta}}_t\boldsymbol{b}_{t+1}\|_2}, \mathbf{\Theta}^*\boldsymbol{b}_{t+1}\right\rangle\right) \\ & = \sum_{t=B_{\text{init}}}^{T-1}\mathbb{E}\left(\left\langle\frac{\mathbf{\Theta}^*\boldsymbol{b}_{t+1}}{\|\mathbf{\Theta}^*\boldsymbol{b}_{t+1}\|_2} - \frac{\hat{\mathbf{\Theta}}_t\boldsymbol{b}_{t+1}}{\|\hat{\mathbf{\Theta}}_t\boldsymbol{b}_{t+1}\|_2}, \mathbf{\Theta}^*\boldsymbol{b}_{t+1}\right\rangle\mathbb{1}\{Q_t\}\right) + \mathbb{E}\left(\left\langle\frac{\mathbf{\Theta}^*\boldsymbol{b}_{t+1}}{\|\mathbf{\Theta}^*\boldsymbol{b}_{t+1}\|_2} - \frac{\hat{\mathbf{\Theta}}_t\boldsymbol{b}_{t+1}}{\|\hat{\mathbf{\Theta}}_t\boldsymbol{b}_{t+1}\|_2}, \mathbf{\Theta}^*\boldsymbol{b}_{t+1}\right\rangle\mathbb{1}\{Q_t^c\}\right) \\ & \leq \sum_{t=B_{\text{init}}}^{T-1}\left(\mathbb{E}\left(\left\langle\frac{(\mathbf{\Theta}^* - \hat{\mathbf{\Theta}}_t)\boldsymbol{b}_t}{\|\mathbf{\Theta}^*\boldsymbol{b}_{t+1}\|_2} + \frac{\|\hat{\mathbf{\Theta}}_t\boldsymbol{b}_{t+1}\|_2 - \|\mathbf{\Theta}^*\boldsymbol{b}_{t+1}\|_2}{\|\hat{\mathbf{\Theta}}_t\boldsymbol{b}_{t+1}\|_2\|\mathbf{\Theta}^*\boldsymbol{b}_{t+1}\|_2}\hat{\mathbf{\Theta}}_t\boldsymbol{b}_{t+1}, \mathbf{\Theta}^*\boldsymbol{b}_{t+1}\right\rangle\mathbb{1}\{Q_t\}\right)\right. \\ & \qquad\qquad \left. + 2\|\mathbf{\Theta}^*\|_2\sqrt{\mathbb{E}[\|\boldsymbol{b}_{t+1}\|^2]}\left(\frac{3}{t} + \frac{2}{t^2} + \frac{2}{Lt} + \frac{2}{L^3t^3} + \frac{1}{t^{\frac{1}{3}}}\right)\right) \\ & \leq \sum_{t=B_{\text{init}}}^{T-1}\mathbb{E}\left[2\|\mathbf{\Delta}\|_2\|\boldsymbol{b}_{t+1}\|\mathbb{1}\{Q_t\}\right] + 2\sqrt{Ld_x}\|\mathbf{\Theta}^*\|_2\left(\frac{3}{t} + \frac{2}{t^2} + \frac{2}{Lt} + \frac{2}{L^3t^3} + \frac{1}{t^{\frac{2}{15}}}\right) \\ & \leq \sum_{t=B_{\text{init}}}^{T-1}\mathbb{E}\left[2\|\mathbf{\Delta}\|_F\|\boldsymbol{b}_{t+1}\|\mathbb{1}\{Q_t\}\right] + 2\sqrt{Ld_x}\|\mathbf{\Theta}^*\|_2\left(\frac{3}{t} + \frac{2}{t^2} + \frac{2}{Lt} + \frac{2}{L^3t^3} + \frac{1}{t^{\frac{2}{15}}}\right) \\ & \leq 2\sum_{t=B_{\text{init}}}^{T-1}\sqrt{Ld_x}\left(\frac{3}{t^{\frac{2}{15}}}\frac{\sigma\sqrt{d_x+1}}{\sqrt{L}h^2} + 6\lambda_0\frac{\sqrt{2d}}{h^2t^{\frac{1}{6}}}\right) + 2\sqrt{Ld_x}\|\mathbf{\Theta}^*\|_2\left(\frac{3}{t} + \frac{2}{t^2} + \frac{2}{Lt} + \frac{2}{L^3t^3} + \frac{1}{t^{\frac{2}{15}}}\right). \end{aligned} \tag{15}$$

Similar arguments also give

$$\mathbb{E}\left(\sum_{t=0}^{B_{\text{init}}-1}\sum_{l=1}^{L}\mathbb{E}\left(\boldsymbol{a}_t^{*\top}\mathbf{\Theta}^*\boldsymbol{x}_{t,l} - \boldsymbol{a}_t^{\top}\mathbf{\Theta}^*\boldsymbol{x}_{t,l}\right)\right) \leq B_{\text{init}} \times 2\sqrt{Ld_x}\|\mathbf{\Theta}^*\|_2. \tag{16}$$

Therefore, for $T \geq B_{\text{init}}$,

$$\mathcal{R}^{\pi}(T) \leq 2\sqrt{Ld_x}\|\boldsymbol{\Theta}^*\|_2 B_{\text{init}} T^{-1} + \frac{60}{13}\sqrt{Ld_x}\|\boldsymbol{\Theta}^*\|_2 T^{-\frac{2}{15}} + \frac{90}{13}\frac{\sigma(d_x+1)}{h^2}T^{-\frac{2}{15}} + \frac{72}{5}\lambda_0\frac{\sqrt{2rd_xL}}{h^2}T^{-\frac{1}{6}}. \tag{17}$$

A.1. Proof of Lemma A.2

Suppose $r_{t,l} = \boldsymbol{a}_{t,l}^T \boldsymbol{\Theta}^* \boldsymbol{x}_{t,l} + \sigma\varepsilon_{t,l}$, then

$$\begin{aligned}\nabla\mathcal{L}_T(\boldsymbol{\Theta}^*) &= \frac{\sigma}{LT}\sum_{t=1}^{T}\sum_{l=1}^{L} -\varepsilon_{t,l}\boldsymbol{x}_{t,l}\boldsymbol{a}_{t,l}^{\top} \\ &= \frac{\sigma}{LT}\sum_{l=1}^{L} -\varepsilon_{1,l}\boldsymbol{x}_{1,l}\boldsymbol{a}_{1,l}^{\top} + \frac{\sigma}{LT}\sum_{t=2}^{T}\sum_{l=1}^{L}\left(-\varepsilon_{t,l}\boldsymbol{x}_{t,l}\hat{\boldsymbol{a}}_t^{\top} - \varepsilon_{t,l}\boldsymbol{x}_{t,l}\boldsymbol{\delta}_t^{\top}\right).\end{aligned} \tag{18}$$

Now we consider the terms in (18) separately. Let

$$\begin{aligned}S_2 &= \frac{\sigma}{LT}\sum_{l=1}^{L} -\varepsilon_{1,l}\boldsymbol{x}_{1,l}\boldsymbol{a}_{1,l}^{\top} + \frac{\sigma}{LT}\sum_{t=2}^{T}\sum_{l=1}^{L} -\varepsilon_{t,l}\boldsymbol{x}_{t,l}\hat{\boldsymbol{a}}_t^{\top}. \\ S_3 &= \frac{\sigma}{LT}\sum_{t=2}^{T}\sum_{l=1}^{L} -\varepsilon_{t,l}\boldsymbol{x}_{t,l}\boldsymbol{\delta}_t^{\top}.\end{aligned} \tag{19}$$

Elementary Calculation show that

$$\mathbb{E}(\|S_2\|_F^4) \leq \frac{\sigma^4(d_x^2 + 2d_x)}{L^2T^2}. \tag{20}$$

Therefore,

$$P(\|S_2\|_F \geq \frac{\sigma\sqrt{d_x+1}}{\sqrt{LT}}T^{\frac{1}{30}}) \leq \frac{1}{T^{\frac{2}{15}}}. \tag{21}$$

For S_3, let G be an event defined as

$$\begin{aligned}G = \Big\{ &\max\{|\varepsilon_{t,l}| : 1 \leq t \leq T, 1 \leq l \leq L\} \leq 3\sqrt{\log TL}, \\ &\max\{\|\boldsymbol{x}_{t,l}\|^2 : 1 \leq t \leq T, 1 \leq l \leq L\} \leq 2d_x + 6\log LT, \\ &\max\{\|\boldsymbol{\delta}_t/h\|_2^2 : 1 \leq t \leq T\} \leq 2d_a + 6\log T\Big\}.\end{aligned} \tag{22}$$

Then elementary calculation shows that

$$P(G^c) \leq \frac{2}{T^3L^3} + \frac{1}{LT} + \frac{1}{T}. \tag{23}$$

Using Matrix Bernstein Inequality (Tropp, 2012) on event G, we have the operator norm of S_3 on G is bounded as follows

$$P\left(\left\{\left\|\frac{LT}{\sigma}S_3\right\|_2 \geq \alpha\right\} \cap G\right) \leq (d_x + d_a)\exp\left(\frac{-\alpha^2}{2\sigma_{S_3}^2 + 2D\alpha/3}\right), \tag{24}$$

where

$$\begin{aligned}\sigma_{S_3}^2 \geq \max\Bigg\{&\left\|\sum_{t=1}^{T}\mathbb{E}\left(\left(\sum_{l=1}^{L}\varepsilon_{t,l}\boldsymbol{x}_{t,l}\boldsymbol{\delta}_t^{\top}\right)\left(\sum_{l=1}^{L}\varepsilon_{t,l}\boldsymbol{x}_{t,l}\boldsymbol{\delta}_t^{\top}\right)^{\top}\right)\right\|_2, \\ &\left\|\sum_{t=1}^{T}\mathbb{E}\left(\left(\sum_{l=1}^{L}\varepsilon_{t,l}\boldsymbol{x}_{t,l}\boldsymbol{\delta}_t^{\top}\right)^{\top}\left(\sum_{l=1}^{L}\varepsilon_{t,l}\boldsymbol{x}_{t,l}\boldsymbol{\delta}_t^{\top}\right)\right)\right\|_2\Bigg\},\end{aligned} \tag{25}$$

and

$$D = \max_{t}\sup_{\text{event } G \text{ holds}}\|\sum_{l=1}^{L} -\varepsilon_{t,l}\boldsymbol{x}_{t,l}\boldsymbol{\delta}_t^{\top}\|_2 \leq 6Lh\sqrt{\log TL}\sqrt{(d_x + 3\log LT)(d_a + 3\log T)}. \tag{26}$$

Elementary calculation shows that taking

$$\sigma_{S_3}^2 = h^2 \lfloor T^{\frac{2}{3}} \rfloor L \max\{d_a, d_x\} \tag{27}$$

satisfies equation (25).

Taking

$$\begin{aligned} \alpha =& 2hT^{\frac{1}{3}} \log T \sqrt{L \max\{d_a, d_x\} \log{(d_a + d_x)}} + \\ & 8hL\sqrt{\log TL}\sqrt{(d_x + 3\log{(LT)})(d_a + 3\log T)}(\log{(d_x + d_a)} + 2\log T) \end{aligned} \tag{28}$$

$$P(\{\| \frac{LT}{\sigma} S_3 \|_2 \geq \alpha\} \cap G) \leq \frac{1}{T^2}. \tag{29}$$

Therefore, we have

$$\begin{aligned} & P\Big(\|S_3\|_2 \leq 2h\sigma T^{-2/3} \log T \sqrt{\frac{\max\{d_a, d_x\} \log{(d_a + d_x)}}{L}} + \\ & \frac{8h\sigma}{T}\sqrt{\log{(TL)}}\sqrt{(d_x + 3\log(LT))(d_a + 3\log T)}(\log(d_x + d_a) + 2\log T)\Big) \\ & \geq 1 - \frac{2}{L^3T^3} - \frac{1}{LT} - \frac{1}{T} - \frac{1}{T^2} \end{aligned} \tag{30}$$

Recalling that

$$|\langle \nabla \mathcal{L}_T(\Theta^*), \Delta \rangle| = |\langle S_2, \Delta \rangle + \langle S_3, \Delta \rangle| \leq \|S_2\|_F \|\Delta\|_F + \|S_3\|_2 \|\Delta\|_*, \tag{31}$$

we get the statement of the lemma.

A.2. Proof of Lemma A.1

Let $\boldsymbol{b}_t = \sum_{l=1}^{L} \boldsymbol{x}_{t,l}$. Let $\boldsymbol{\delta}_t = \mathbf{0}$ for exploitation rounds.

Then we know that

$$\begin{aligned} \mathcal{E}_T(\Delta) &= \frac{1}{2LT} \sum_{t=1}^{T} \sum_{l=1}^{L} (\boldsymbol{a}_{t,l}^\top \Delta \boldsymbol{x}_{t,l})^2 \\ &= \frac{1}{2LT} \sum_{t=1}^{T} \sum_{l=1}^{L} ((\frac{\boldsymbol{b}_t^\top \hat{\boldsymbol{\Theta}}_{t-1}^\top}{\|\boldsymbol{b}_t^\top \hat{\boldsymbol{\Theta}}_{t-1}^\top\|} + \boldsymbol{\delta}_t^\top) \Delta \boldsymbol{x}_{t,l})^2 \end{aligned} \tag{32}$$

Define

$$\begin{aligned} \mathcal{D}_T(\Delta) &= \frac{1}{2LT} \sum_{t=1}^{T} \sum_{l=1}^{L} \left((\frac{\boldsymbol{b}_t^\top \hat{\boldsymbol{\Theta}}_{t-1}^\top}{\|\boldsymbol{b}_t^\top \hat{\boldsymbol{\Theta}}_{t-1}^\top\|} \Delta \boldsymbol{x}_{t,l})^2 + (\boldsymbol{\delta}_t^\top \Delta \boldsymbol{x}_{t,l})^2 \right), \\ \mathcal{D}_{1,T}(\Delta) &= \frac{1}{2LT} \sum_{t=1}^{T} \sum_{l=1}^{L} (\frac{\boldsymbol{b}_t^\top \hat{\boldsymbol{\Theta}}_{t-1}^\top}{\|\boldsymbol{b}_t^\top \hat{\boldsymbol{\Theta}}_{t-1}^\top\|} \Delta \boldsymbol{x}_{t,l})^2 \\ \mathcal{D}_{2,T}(\Delta) &= \frac{1}{2LT} \sum_{t=1}^{T} \sum_{l=1}^{L} (\boldsymbol{\delta}_t^\top \Delta \boldsymbol{x}_{t,l})^2 \end{aligned} \tag{33}$$

Then

$$\mathcal{E}_T(\Delta) - \mathcal{D}_T(\Delta) = \frac{1}{LT} \sum_{t=1}^{T} \sum_{l=1}^{L} (\frac{\boldsymbol{b}_t^\top \hat{\boldsymbol{\Theta}}_{t-1}^\top}{\|\boldsymbol{b}_t^\top \hat{\boldsymbol{\Theta}}_{t-1}^\top\|} \Delta \boldsymbol{x}_{t,l})(\boldsymbol{\delta}_t^\top \Delta \boldsymbol{x}_{t,l}) \tag{34}$$

Elementary calculation shows that

$$\mathbb{E}(\mathcal{E}_T(\Delta) - \mathcal{D}_T(\Delta)) = 0, \tag{35}$$

and

$$\mathbb{E}(\mathcal{D}_{2,T}(\Delta)) \geq \frac{\lfloor T^{\frac{2}{3}} \rfloor}{2T} h^2 \|\Delta\|_F^2. \tag{36}$$

Now we proceed with proving that the following two bounds hold with high probability:

$$\begin{aligned}\inf_{\|\Delta\|_2>0}\frac{\mathcal{E}_T(\Delta)-\mathcal{D}_T(\Delta)}{\|\Delta\|_2^2} &\geq -7T^{-\frac{2}{3}}(h+h^2)\left(2d_x+2d_a+6\log T+6\log L\right)^2\log T\\ \inf_{\|\Delta\|_2>0}\frac{\mathcal{D}_{2,T}(\Delta)-\mathbb{E}(\mathcal{D}_{2,T}(\Delta))}{\|\Delta\|_2^2} &\geq -7T^{-\frac{2}{3}}(h+h^2)\left(2d_x+2d_a+6\log T+6\log L\right)^2\log T.\end{aligned} \tag{37}$$

Note that $\|\boldsymbol{x}_{t,l}\|_2^2\sim\chi^2_{d_x}$, $\|\boldsymbol{\delta}_t/h\|_2^2\sim\chi^2_{d_a}$. Therefore, we have that

$$\begin{aligned}&P(\sup_{t,l}\|\boldsymbol{x}_{t,l}\|_2^2\leq d_x+2\epsilon_1+2\sqrt{\epsilon_1 d_x},\sup_t\|\boldsymbol{\delta}_t/h\|_2^2\leq d_a+2\epsilon_2+2\sqrt{\epsilon_2 d_a})\\ &\geq 1-\left(LT\exp\left(-\epsilon_1\right)+T\exp\left(-\epsilon_2\right)\right).\end{aligned} \tag{38}$$

Let $\epsilon_1=2\log LT,\epsilon_2=2\log T$.

Denote

$$U_1=d_x+2\epsilon_1+2\sqrt{\epsilon_1 d_x},U_2=d_a+2\epsilon_2+2\sqrt{\epsilon_2 d_a}. \tag{39}$$

And let the event O be

$$O=\{\sup_{t,l}\|\boldsymbol{x}_{t,l}\|_2^2\leq U_1,\sup_t\|\boldsymbol{\delta}_t/h\|_2^2\leq U_2\}. \tag{40}$$

For the following, we restrict our attention to event O.

Note that

$$\inf_{U_0/1.1\|\Delta\|_2\leq U_0}\mathcal{E}_T(\Delta)-\mathcal{D}_T(\Delta)\geq\inf_{U_0/1.1\|\Delta\|_2\leq U_0,\hat{\boldsymbol{\Theta}}_{t-1}^\top\neq\mathbf{0}\text{ for }1\leq t\leq T}\mathcal{E}_T(\Delta)-\mathcal{D}_T(\Delta), \tag{41}$$

also at most $\lfloor T^{\frac{2}{3}}\rfloor$ terms in the sum of $\mathcal{E}_T(\Delta)-\mathcal{D}_T(\Delta)$ are not zero, and for any term in the exploration round $\left(\sup_{U_0/1.1\leq\|\Delta\|_2\leq U_0,\hat{\boldsymbol{\Theta}}_{t-1}^\top\neq\mathbf{0}\text{ for }1\leq t\leq T}(\frac{\boldsymbol{b}_t^\top\hat{\boldsymbol{\Theta}}_{t-1}^\top}{\|\boldsymbol{b}_t^\top\hat{\boldsymbol{\Theta}}_{t-1}^\top\|}\Delta\boldsymbol{x}_{t,l})(\boldsymbol{\delta}_t^\top\Delta\boldsymbol{x}_{t,l})\right)-\left(\inf_{U_0/1.1\leq\|\Delta\|_2\leq U_0,\hat{\boldsymbol{\Theta}}_{t-1}^\top\neq\mathbf{0}\text{ for }1\leq t\leq T}(\frac{\boldsymbol{b}_t^\top\hat{\boldsymbol{\Theta}}_{t-1}^\top}{\|\boldsymbol{b}_t^\top\hat{\boldsymbol{\Theta}}_{t-1}^\top\|}\Delta\boldsymbol{x}_{t,l})(\boldsymbol{\delta}_t^\top\Delta\boldsymbol{x}_{t,l})\right)\leq 2U_1\sqrt{U_2}hU_0^2$

Therefore, through Functional Hoeffding theorem (Theorem 3.26 in Wainwright (2019))), we have

$$P(\mathcal{E}_T(\Delta)-\mathcal{D}_T(\Delta)\leq-\gamma_1|O)\leq\exp\left(-\frac{\frac{T^2}{\lfloor T^{\frac{2}{3}}\rfloor}\gamma_1^2}{16U_1^2U_2h^2U_0^4}\right) \tag{42}$$

for $\gamma_1>0$.

Similarly, for the exploration rounds in $\mathcal{D}_{2,T}(\Delta)$, we have

$$\left(\sup_{U_0/1.1\leq\|\Delta\|\leq U_0}(\boldsymbol{\delta}_t^\top\Delta\boldsymbol{x}_{t,l})^2\right)-\left(\inf_{U_0/1.1\leq\|\Delta\|\leq U_0}(\boldsymbol{\delta}_t^\top\Delta\boldsymbol{x}_{t,l})^2\right)\leq U_1U_2U_0^2h^2. \tag{43}$$

Again, according to Functional Hoeffding theorem, we have

$$P(\mathcal{D}_{2,T}(\Delta)-\mathbb{E}\left(\mathcal{D}_{2,T}(\Delta)\right)\leq-\gamma_2|O)\leq\exp\left(-\frac{\frac{T^2}{\lfloor T^{\frac{2}{3}}\rfloor}\gamma_2^2}{4U_1^2U_2^2U_0^4h^4}\right) \tag{44}$$

Take $\gamma_1=\gamma_2=7T^{-\frac{2}{3}}(h+h^2)\left(2d_x+2d_a+6\log T+6\log L\right)^2\|\Delta\|_2^2\log T$.

Therefore,

$$\begin{aligned}
&P\left(\mathcal{E}_T(\Delta) - \mathbb{E}(\mathcal{D}_{2,T}) \leq -14T^{-\frac{2}{3}}(h+h^2)\left(2d_x + 2d_a + 6\log T + 6\log L\right)^2 \|\Delta\|_2^2 \log T\right) \\
&\leq P\left(\mathcal{E}_T(\Delta) - \mathcal{D}_T(\Delta) \leq -7T^{-\frac{2}{3}}(h+h^2)\left(2d_x + 2d_a + 6\log T + 6\log L\right)^2 \|\Delta\|_2^2 \log T | O\right) \\
&+ P\left(\mathcal{D}_{1,T}(\Delta) \leq 0 | O\right) \\
&+ P\left(\mathcal{D}_{2,T} - \mathbb{E}(\mathcal{D}_{2,T}) \leq -7T^{-\frac{2}{3}}(h+h^2)\left(2d_x + 2d_a + 6\log T + 6\log L\right)^2 \|\Delta\|_2^2 \log T | O\right) \\
&+ P(O^c) \\
&\leq \frac{1}{LT} + \frac{2}{T} + \frac{1}{T^2}
\end{aligned} \tag{45}$$

Hence with probability at least $1 - \frac{1}{LT} - \frac{2}{T} - \frac{1}{T^2}$,

$$\mathcal{E}_T(\Delta) \geq \frac{\lfloor T^{\frac{2}{3}} \rfloor}{2T} h^2 \|\Delta\|_F^2 - 14T^{-\frac{2}{3}}(h+h^2)\left(2d_x + 2d_a + 6\log T + 6\log L\right)^2 \|\Delta\|_2^2 \log T. \tag{46}$$

B. Details on the simulation study

In this section, we detail the tuning parameters of each algorithm we used for the simulation study.

`Hi-CCAB`. There are three tuning parameters for `Hi-CCAB`: we set the steps for initialization $t_1 = 100$, the initialized penalization parameter $\lambda_0 = \|\frac{1}{2t_1 L}\sum_{i=1}^{t_1}\sum_{j=1}^{L} |\boldsymbol{a}_i^\top \widehat{\boldsymbol{\Theta}}_{t_1} \boldsymbol{x}_{i,j} - y_{i,j}| \boldsymbol{x}_{i,j} \boldsymbol{a}_i^\top \|_2$, and the exploration parameter $h = .1$.

LinUCB (Li et al., 2010). We apply the LinUCB with disjoint linear models and set multiplier for the upper confidence bound $\alpha = 1 + \sqrt{\ln(2/\delta)/2}$ with $\delta = .05$ as suggested in the paper.

Lasso Bandit (Bastani & Bayati, 2020). There are a couple of tuning parameters in the original algorithm including h for the set of "near-optimal arms", q for the force-sample set, and λ_1 and $\lambda_{2,0}$ as the regularization parameters for the "forced sample estimate" and "all-sample estimate". We follow the original paper and set $h = 5$, $\lambda_1 = \lambda_{2,0} = 0.05$. We set $q = 2$ so that the size of initialized forced sample set is close to that we used for `Hi-CCAB`.

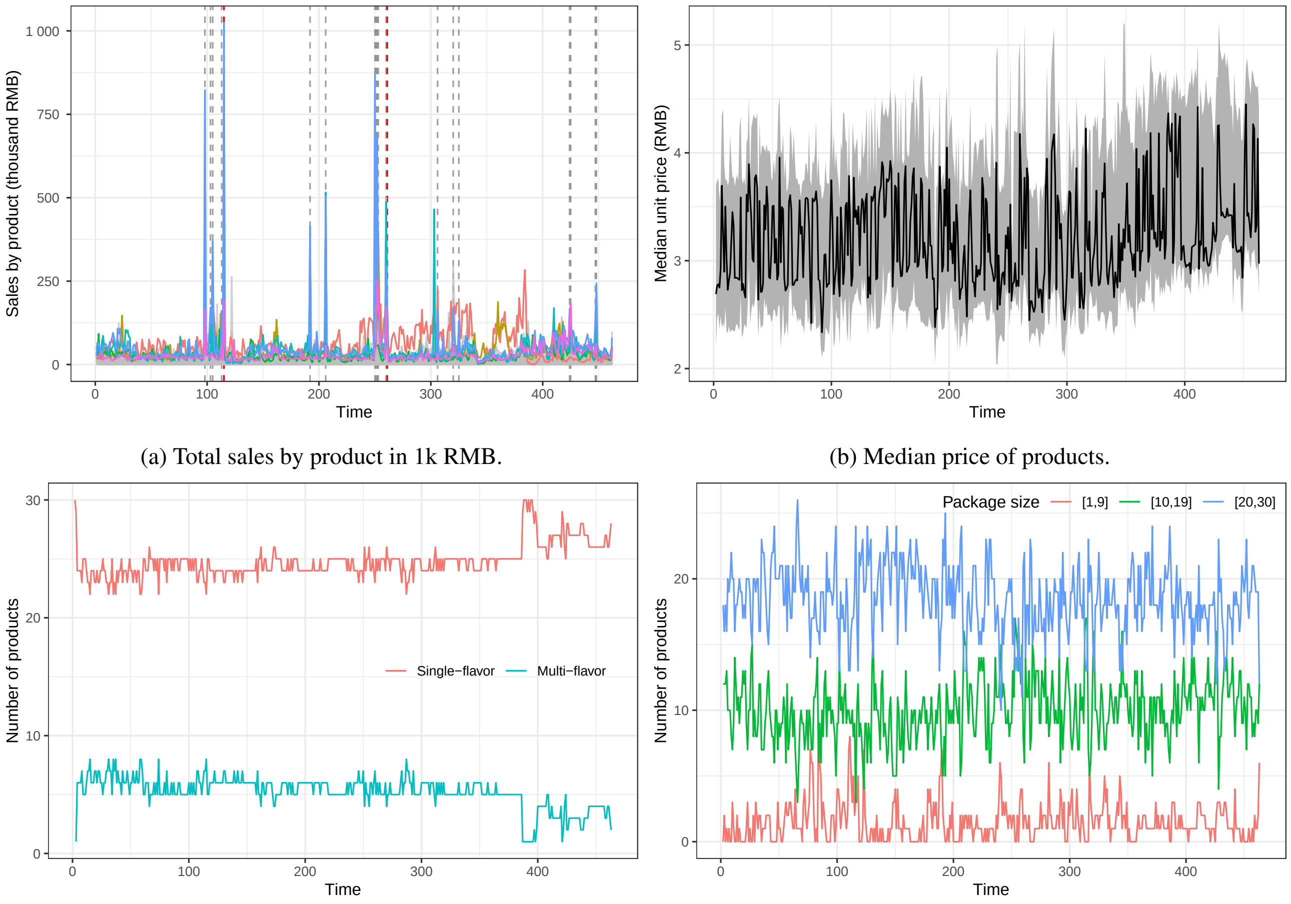
NeuralUCB (Zhou et al., 2020). The tuning parameters of NeuralUCB include the confidence parameter as in all UCB-based algorithm, the size of neural network, as well as the step size, regularization parameter for gradient descent to train the neural network. We adapted the code from `https://github.com/uclaml/NeuralUCB` and used the default settings.

EE-Net (Ban et al., 2022). EE-Net involves tuning parameters for gradient descent to train the exploitation network, exploration network, and the decision-maker network. We adapted the code from `https://github.com/banyikun/EE-Net-ICLR-2022` and used the default settings.

C. More details on the case study and additional numerical results

In this section, we provides more background information on the case study and additional interpretations of the represetation matrix $\boldsymbol{\Theta}$ and numerical results.

Figure 3a shows the daily sales by product and each color represents one product (only products that appeared more than 95% of the days are colored; the rest are colored as grey). The days corresponding to the vertical dashed grey lines are days with promotion. The two red vertical lines correspond to the annual sales events. The variation between products was large and one product dominated the rest most of the time. The sales were also driven by the promotion – the sales went up when there is a promotion. Figure 3b shows the median unit price across time with the 25th and 75th quantiles as the boundaries of the grey area. The median unit price was around 3.2 RMB and there were variations in unit price among products. Figure 3c shows the number of single-flavor and multi-flavor products. Three-quarters of the products were single-flavored. Note that products with the same flavor can have different package sizes. Figure 3d shows the number of products with different package sizes. The package size of about 60% of the products is larger than 20 with 30% having package sizes between 10 and 20 and the rest less than 10.



(a) Total sales by product in 1k RMB.

(b) Median price of products.

(c) Number of products with various number of flavors.

(d) Number of products with various package sizes.

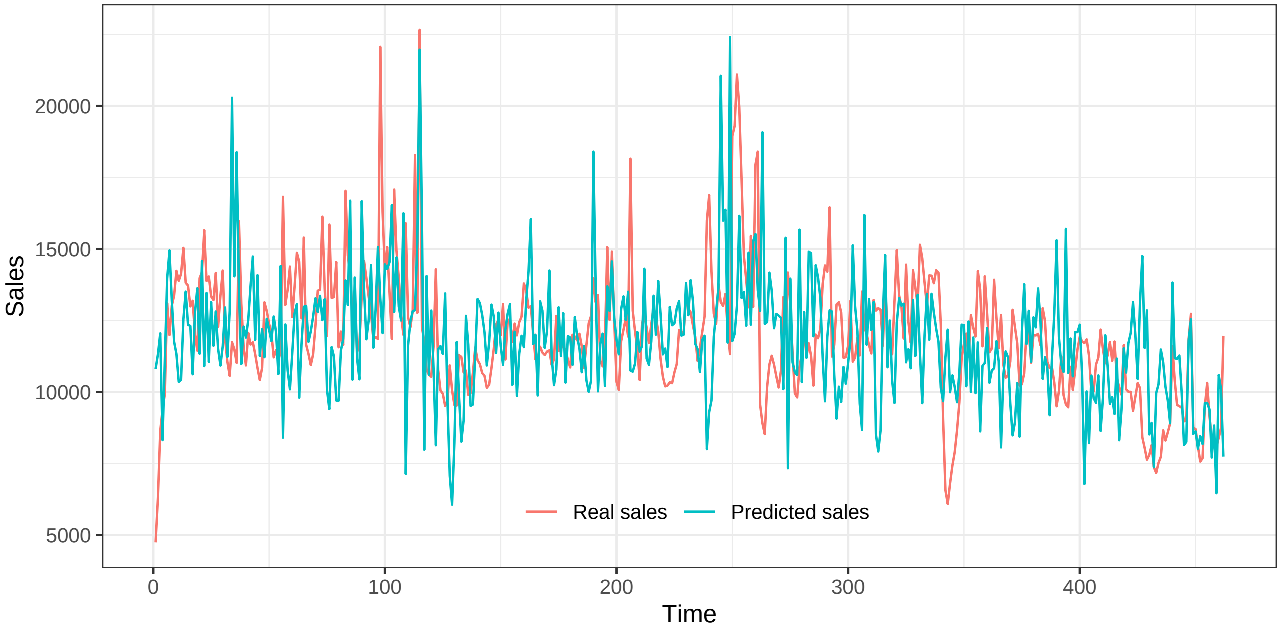
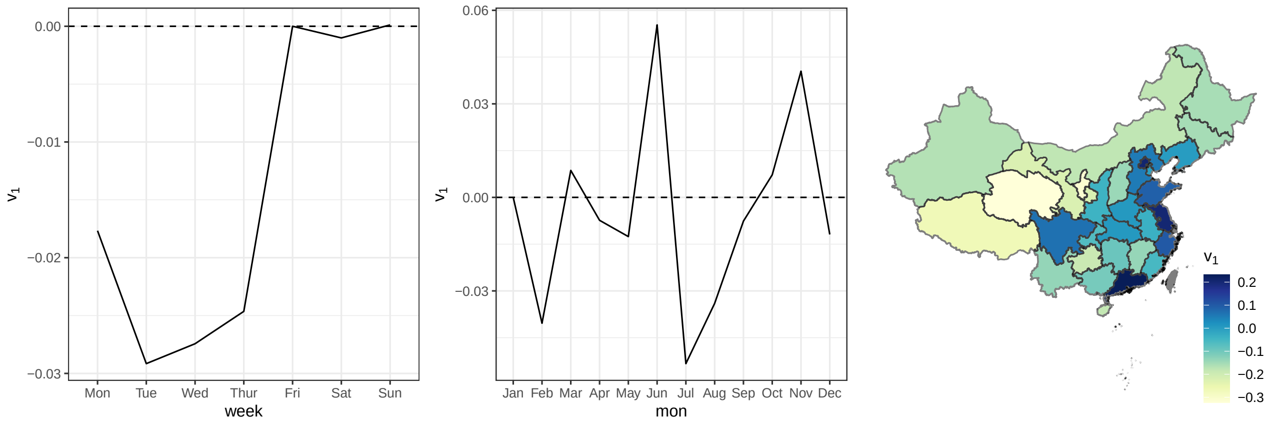


Figure 4: Real sales vs predicted sales.



(a) Weekday (b) Month (c) Location

Figure 5: Loadings of the leading right singular vectors for the covariates.

To check our model assumption (1) on the data, Figure 4 shows the hold-out-sample prediction of the sales versus the real sales. The predicted sale at each time point t is the predicted total sales across $L = 31$ locations based on $\widehat{\mathbf{\Theta}}_{-t}$ estimated from all the data except for data at time t, i.e.,

$$\hat{y}_{t,j} = \boldsymbol{a}_t^\top \widehat{\mathbf{\Theta}}_{-t} \boldsymbol{x}_{t,j}$$

where $\widehat{\mathbf{\Theta}}_{-t} = \arg\min_{\mathbf{\Theta}} \sum_{i=1, i \neq t}^{T} \sum_{j=1}^{L} \left(\boldsymbol{a}_i^\top \mathbf{\Theta} \boldsymbol{x}_{i,j} - r_{i,j} \right)^2 + \lambda \|\mathbf{\Theta}\|_*$. As shown in Figure 4, the real sales and the out-of-sample predicted sales follow quite closely across time, which indicates that both our model and estimation are reasonable.

Structure of the representation matrix $\mathbf{\Theta}$. One advantage of model is the interpretability which allows us to gain insights from the representation matrix $\mathbf{\Theta}$. Specifically, our model is able to discover the underlying factors of the effect of both arms and covariates on the reward. In the following, we will examine the pseudo ground truth $\mathbf{\Theta}$ we obtained using all the data.

The rank of $\mathbf{\Theta}$ is 5 with the singular values being $(2.5, 0.3, 0.2, 0.02, 0.002)$. The leading singular value dominates the rest and thus the leading left and right singular vectors are the most important ones in explaining the effect on the reward and we focus on the leading singular vectors in what follows.

Figure 5 shows the loadings for different covariates (i.e., the leading right singular vector) and our algorithm is able to learn interpretable patterns of the effects on the reward – for weekday, the effects are drastically different during the weekend and during the weekend; for months, the effects show different patterns during the promotion month (June and November) from other months; for location, the effects of the coastal provinces are different from the rest, which exactly corresponds to the levels of economic development of different regions in China. In sum, our model can exploit the underlying structure of the covariates and provide insights into purchasing behavior and seasonality.

On the other hand, Table 1 explores the loadings for the arm on May 29th 2022, the last Sunday in our data (i.e., the leading left singular vectors multiplied with $\langle \boldsymbol{v}_1, \bar{\boldsymbol{x}} \rangle$ where $\bar{\boldsymbol{x}}$ is the average of $\boldsymbol{x}_j$ for $j = 1, \ldots, L$ on May 29th 2022). Specifically, we investigate the effect of flavors on the reward given the context. We take the average of the loadings of the linear and quadratic terms for each flavor in all 30 products and compare with the total sales of each flavor across all Sundays in Mays. For ease of comparison, we further scale the sales and the loadings by their corresponding largest numbers. The loadings and sales are closely related to each other.[2] As in Table 1, on May 29th 2022, flavor 1 (F1) has the largest effect, followed by flavor 10, 13, 7, 9 and 11. Therefore, our model learns the values of the flavors (per unit).

More on simulation with additional numerical results. We first detail how we ran the simulation and then provide more simulation results. To be specific, we first use $t_1 = 100$ for the initialization step to estimate $\widehat{\mathbf{\Theta}}_{t_1}$; and then at each time $t = t_1 + 1, \ldots, T$, we follow Algorithm 1 to decide on the action $\boldsymbol{a}_t$ for assortment and pricing. After determining $\boldsymbol{a}_t$, we generate the sales $\boldsymbol{r}_t$ according to (1) using the pseudo true $\mathbf{\Theta}$ and σ. We further compare the performance of the

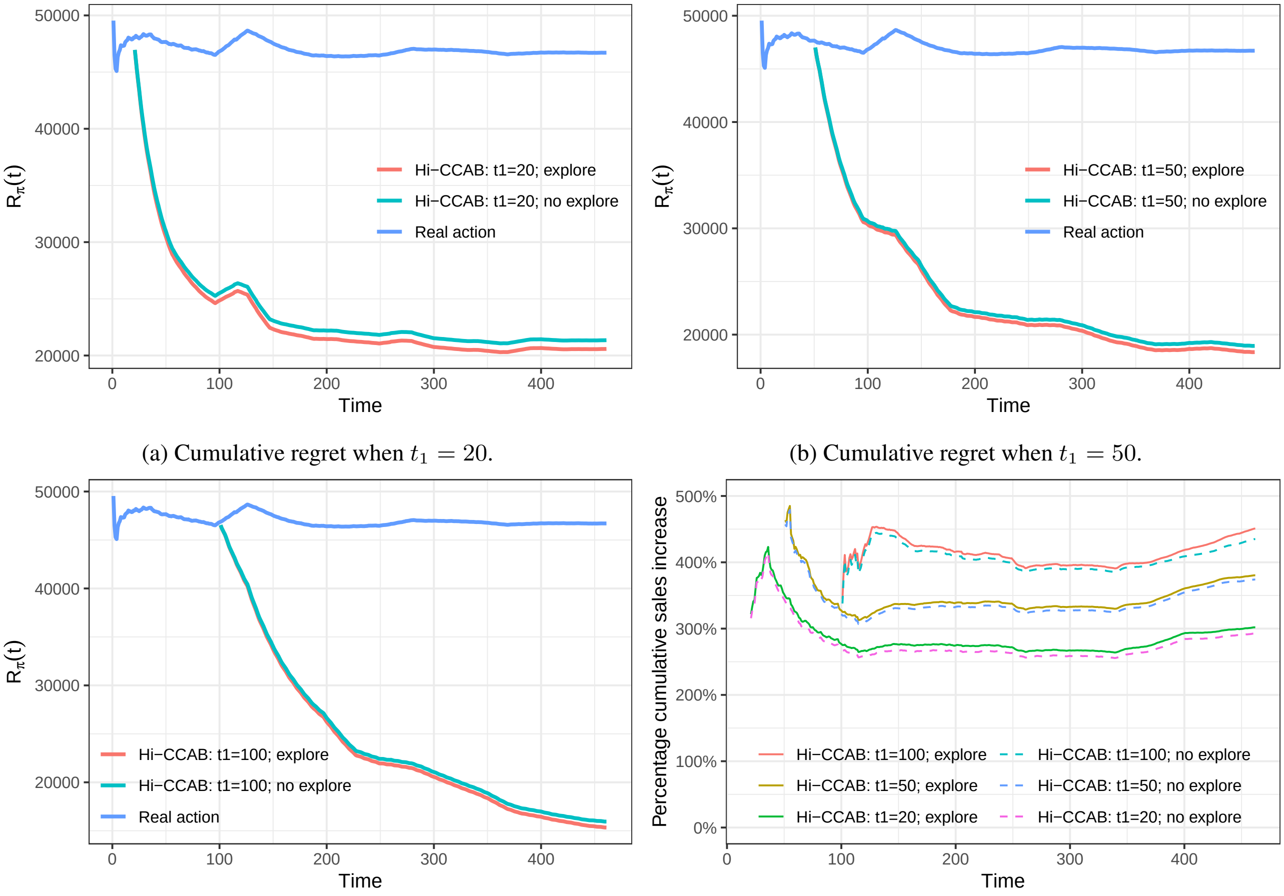
[2] The correlation of sales and the linear-term loadings is 0.91 and that of the quadratic-term loadings is 0.97.

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13
Sales	1.00	0.05	0.00	0.00	0.00	0.03	0.19	0.00	0.08	0.19	0.18	0.00	0.38
$\tilde{u}_1$ (linear)	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\tilde{u}_1$ (quadratic)	1.00	0.12	-0.00	0.00	0.00	0.03	0.19	0.00	0.15	0.39	0.16	0.03	0.33

Table 1: Total sales and loadings of the linear and quadratic terms (scaled) of the 13 flavors.

assortment-pricing policy with exploration and without exploration and with different initialization time t_1. Each setup is simulated 100 times.

Figures 6a-6b show cumulative regret and Figures 6d show percentage gain in cumulative sales when $t_1 = 20, 50, 100$ with exploration and without exploration. `Hi-CCAB` with exploration performs better then without exploration. As expected, longer initialization steps provide a better initial estimation of the $\mathbf{\Theta}$ and thus helps with the performance in a short time windows. As time goes by, all of the expected regrets converge to zero and the percentage gain in cumulative sales should converge.



(a) Cumulative regret when $t_1 = 20$.

(b) Cumulative regret when $t_1 = 50$.

(c) Cumulative regret when $t_1 = 100$.

(d) Percentage cumulative sales gain.

Figure 6: Performance of `Hi−CCAB` with different initialization times t_1 and with exploration and without exploration.