Doubly High-Dimensional Contextual Bandits: An Interpretable Model for Joint Assortment-Pricing

Junhui Cai
Department of Information Technology, Analytics, and Operations, University of Notre Dame, jcai2@nd.edu

Ran Chen
Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, ran1chen@mit.edu

Martin J. Wainwright
Laboratory for Information and Decision Systems, Statistics and Data Science Center,
EECS & Mathematics, Massachusetts Institute of Technology, wainwrigwork@gmail.com

Linda Zhao
Department of Statistics and Data Science, University of Pennsylvania, lzhao@wharton.upenn.edu

Key challenges in running a business include deciding which products or services to present to consumers (the assortment problem), and how to price products (the pricing problem) to maximize revenue or profit. Instead of considering these problems in isolation, we address assortment-pricing *jointly* and tackle the intrinsic doubly high dimensionality—both actions and contextual vectors can take continuous value in high-dimensional spaces. We propose a doubly high-dimensional contextual bandit model to formulate this problem. To circumvent the curse of dimensionality, our model is simple yet flexible, capturing the interaction effects between covariates (context) and actions on the reward via a low-rank representation matrix. The resulting class of models is reasonably expressive while remaining interpretable through latent factors and includes various bandit and pricing models as special cases, making it suitable for applications involving simultaneous multiple decision-making beyond joint assortment-pricing. We develop a computationally tractable procedure that combines an exploration/exploitation protocol with an efficient low-rank matrix estimator. We provide a non-asymptotic instance-dependent regret bound involving dimensions and rank in addition to the time horizon. Simulations on standard bandit and pricing models—special cases of our model—demonstrate that our method yields lower regret than state-of-the-art methods. Real-world assortment-pricing case studies, from an industry-leading instant noodle manufacturer to an emerging beauty start-up, underscore the gains achievable using our method, showing at least three-fold gains in revenue/profit and the interpretability of the latent factor models that are learned.

Key words: contextual bandits; on-line decision-making; high-dimensional statistics; low-rank matrices; factor models.

---

[1] Authorship is in alphabetic order.

1. Introduction

In the modern business and healthcare landscape, it is now *status quo* to make use of online sequential decision-making algorithms that incorporate individual characteristics as well as micro- and macro-economic conditions. For example, retailers or manufacturers determine product offerings and pricing based on customer demographics, purchasing history, and seasonal demand; business managers allocate resources, such as staff and equipment, based on current operational conditions; and medical providers prescribe treatment and therapy combinations based on the patient's medical records.

In these settings, bandit algorithms are often deployed to learn the reward structure while optimizing performance by strategically "exploring" and "exploiting" potential actions. To make better decisions, decision-makers should consider the two main influences on a reward (e.g., revenue or profit): the possible actions and the exogenous factors, also called covariates or context, such as individual characteristics and micro- or macroeconomic conditions. Many bandit algorithms are limited to finite or low-dimensional action and context spaces, but in practice, both are often high-dimensional in nature. For instance, the action vector for an online retailer may include pricing and assortment information for dozens of products characterized by numerous attributes. Thus, we are led to consider the following question: can we develop useful models and efficient learning procedures for contextual bandits that are high-dimensional in both actions and covariates? Providing one affirmative answer to this open question and demonstrating the utility of the resulting model and algorithms for two real-world motivating case studies are the primary contributions of our work.

1.1. Background and Our Approach

The primary application that motivates our work is dynamic assortment and pricing. It is a central challenge for manufacturers and retailers, and using bandits for this problem is natural given the sequential nature of the decision-making. The assortment problem refers to deciding what products or services to offer to customers, whereas the pricing problem is to set selling prices for these products. Both assortment and pricing decisions share a common goal: maximizing a specific objective function, such as revenue or profit. Although both dynamic assortment optimization and pricing problems have been separately studied extensively in the literature, the *joint* assortment-pricing problem has received comparatively less attention.

The key to a successful assortment and pricing strategy lies in understanding the market response to the assortment-pricing decisions. A major challenge in modern assortment-pricing is the explosion in dimensionality of both the action and covariate spaces. Companies typically either consider large numbers ($\gg 100$) of products simultaneously or have the need to design new products based on product attributes. From the universe of existing and potential products, they must determine a large collection of products to offer and set appropriate prices. Such decision-making necessitates a high-dimensional and continuous action space. The problem is further complicated by the high-dimensional covariates: fueled by the rise of e-commerce, it is possible to measure many customer-specific or industry-specific features that can be

relevant to modeling demand and price sensitivity. As the action-covariate dimensions grow, 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 and Szepesvári 2020). Thus, it becomes essential to develop models with "low-dimensional structure" that explain important features of the data, while being amenable to statistically and computationally efficient algorithms.

A fortunate fact, and the starting point of our modeling, is that a small set of latent factors spanning a low-dimensional space often explains the bulk of the reward structure. In the context of revenue management, one deciding quantity for reward (revenue) is the demands. The demands for products that share similar features/attributes are influenced in common ways by underlying market conditions. Usually only a handful of the underlying product factors matter. For instance, there exists "color psychology" in marketing (Singh 2006) and customers' color preference in basic colors such as white, black, blue, and red (Madden et al. 2000). Similarly, the covariate vectors relevant for assortment-pricing can be explained by a few latent factors. At the individual level, much of the variance in consumer buying power can be captured by a mixture of demographic (e.g., income, education level) and geographic traits (see Pol (1991) and references therein); at the macro level, population purchasing preference, usually indicated by season, region, and other macroeconomic indices, significantly impact the overall demand (Estelami et al. 2001, Gordon et al. 2013, Kumar et al. 2014). As a result, the interaction effects between the action and covariates—a major source contributing to revenue—can be characterized by a few latent factors.

In summary, the low-dimensional structure captures the essence of the effect of actions and covariates on the reward function and often aligns with intuitive or interpretable factors. Accounting for the common latent factors further speeds up the reward learning regarding sample complexity, and low-dimensional models often improve computational efficiency. The interpretability and computational efficiency using latent factors turn the "curse of dimensionality" into a "blessing of dimensionality" (Li et al. 2018).

With these insights, we tackle the joint assortment-pricing problem by casting it as a doubly high-dimensional bandit problem and proposing a new model that captures interactions between the high-dimensional actions and covariates via an (approximately) low-rank matrix representation. Our goal is to offer a sequence of assortment and pricing decisions, which can be represented as a sequence of action vectors $\{\boldsymbol{a}_t\}_{t=1}^T$ that take values in (some subset of) $\mathbb{R}^{d_a}$, under the contexts, which can be represented as a sequence of covariate vectors $\{\boldsymbol{x}_t\}_{t=1}^T$ taking values in $\mathbb{R}^{d_x}$, such that the cumulative expected revenue over the time horizon T is maximized. Since both the action dimension d_a and the covariate dimension d_x can be large, our proposed model uses a low-rank matrix to take advantage of the low-dimensional structure. Specifically, our reward model takes the bilinear form: given an action vector $\boldsymbol{a} \in \mathbb{R}^{d_a}$ and a covariate vector $\boldsymbol{x} \in \mathbb{R}^{d_x}$, 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 $\mathbf{\Theta}^* \in \mathbb{R}^{d_a \times d_x}$ is an unknown representation matrix that is relatively low-rank—say with rank $r \ll \min\{d_a, d_x\}$—or more generally, well-approximated by a matrix with low rank.

The representation matrix $\mathbf{\Theta}^*$ captures important factors (e.g., interaction effects of action-covariate pairs) via its spectral structure, providing interpretability. Performing a singular value decomposition (SVD) on the matrix yields the latent structure, with the left (respectively right) singular 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.

Given this model structure, we further propose a new algorithm (`Hi-CCAB`) that combines low-rank estimation with an exploration/exploitation strategy. The approach is computationally efficient, involving only convex programs or simple problems admitting closed-form solutions in all phases. We prove a non-asymptotic bound on its expected regret, showing that it is also statistically efficient in terms of problem dimension and the low-rank structure. We also show our method not only can solve the joint assortment-pricing problem, but is actually general enough to encompass various bandit models as special cases.

1.2. Main Contributions

Let us summarize some of our main contributions:

1. **A general and interpretable model for joint assortment-pricing.** We propose a doubly high-dimensional contextual bandit model where both covariates and actions can be high-dimensional and continuous, leveraging the low-dimensional latent factors via a low-rank representation matrix.

   Our model is particularly powerful for tackling the dynamic *joint* assortment-pricing problem, simultaneously addressing two interrelated problems that have largely been studied separately. Our model captures the influence of both the decision—through product attributes and prices—and the contextual information, including their interactions, on demand or sales revenue, thereby naturally accounting for demand heterogeneity driven by action-context interactions. It can also design new products based on attributes, in contrast to most existing assortment models that are limited to a fixed product set. Furthermore, our model is applicable across any level of granularity, from individual customers to actual time intervals such as days or weeks.

   From a technical perspective, as we argue, an advantage of this low-rank model is its combination of a high degree of interpretability with predictive power. The low-rank matrix encapsulates the interaction between action-covariate pairs via its singular vectors, providing a form of dimension reduction and interpretability. Additionally, given the covariate, our model is able to predict the reward of an unseen action. Both interpretability and predictive power can be tremendously useful for decision-makers.

   Our model is general. It unifies a number of structured bandit and pricing models studied in past work; it can capture complex relationships between variables; and it is applicable to an array of applications involving multiple decision-making.

2. **A computationally efficient and adaptive online algorithm.** We propose an efficient online learning algorithm for our new model, termed the **Hi**gh-dimensional **C**ontextual and **Hi**gh-dimensional **C**ontinumm **A**rmed **B**andit (`Hi-CCAB`). It interleaves an estimation step, in which the low-rank representation matrix is estimated based on data observed thus far, with a policy learning step, in which new actions are selected by balancing exploration and exploitation. Both steps are computationally efficient. In addition, `Hi-CCAB` is adaptive to both rank r and time horizon T: it does not require prior knowledge of r or T yet performs well for all ranks and horizons relative to the intrinsic difficulty of the problem.
3. **A non-asymptotic and instance-dependent upper bound.** We measure the performance of our algorithm using the standard notion of expected regret, which is the average expected deficit in reward achieved by `Hi-CCAB` compared with an oracle that knows the low-rank representation matrix. We provide a non-asymptotic instance-dependent upper bound on the expected regret of `Hi-CCAB`. A technical challenge is that samples are highly dependent in a complicated way as the bandit protocol collects data based on all existing observations, making classic matrix theory results for i.i.d. data inapplicable. We overcome this challenge by proving a new tail bound for the low-rank matrix estimator by carefully constructing martingales to separate sources of randomness, developing (matrix-valued) martingale concentration results, analyzing non-standard distributions, and thereby giving a non-asymptotic upper bound on the expected regret. We further note that the bound holds for all T and r while the algorithm does not require prior knowledge of T or r. This adaptivity is of both theoretical interest and practical importance.
4. **Take-away insights for assortment-pricing practice.** We evaluate `Hi-CCAB` in simulation under various standard bandit and pricing models against state-of-art methods and apply it to real-world joint assortment–pricing problems faced by manufacturers. Simulations show that `Hi-CCAB` outperforms state-of-the-art methods in expected regret. We further demonstrate its practical value in revenue maximization through two case studies: one for a leading instant noodle producer and another for a manicure start-up. Both involve a large number of products and covariates, rendering existing methods inapplicable. `Hi-CCAB` successfully handles such doubly high-dimensionalities and provides joint assortment and pricing decisions. The assortment-pricing policy based on `Hi-CCAB` yields sales almost four times as high as the strategies in practice. Moreover, our model reveals insights for assortment and pricing such as the popularity of flavor (noodles) or color (manicure) under different contexts such as locations and seasons. Finally, our model is able to predict the revenue of a new product, which can guide new product designs.

1.3. Related Literature

So as to situate our work more broadly, let's discuss and summarize some related literature.

Dynamic Assortment and Pricing. In the field of operations research and revenue management, 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.

Beginning with dynamic assortment, Caro and Gallien (2007) was an early approach to formulate it as a multi-armed bandit problem, but assuming independent demand for each product. A popular alternative demand model is the multinomial logit (MNL) choice model, which uses a logistic model to estimate demand parameters (Rusmevichientong et al. 2010, Sauré and Zeevi 2013); more recent work has adapted multi-armed bandit techniques to the MNL model (Chen and Wang 2017, Agrawal et al. 2019, Chen et al. 2021, 2023, Shen et al. 2023). The MNL model can be further extended to personalized dynamic assortment by integrating personal information (Cheung and Simchi-Levi 2017, Chen et al. 2020, Miao and Chao 2022). Another MNL variant accounts for heterogeneity via customer segmentation (Bernstein et al. 2019, Kallus and Udell 2020). In particular, Kallus and Udell (2020) also adopt a low-rank matrix to model the interaction between product and customer types, but they only consider finite types of products and customers and does not account for product/customer attributes.

Despite the many merits of MNL models, they fall short in addressing several practical challenges faced by our collaborating companies, along with many other manufacturers and retailers. First, the unit of analysis in MNL models is at the individual customer level, whereas firms often operate on accounting periods defined by actual time (e.g., daily or weekly) due to operational constraints, platform restrictions, and concerns over brand perception (Cavallo 2018, Aparicio et al. 2023, Ferreira and Mower 2023). Extending MNL models to actual time settings requires overcoming multiple practical modeling challenges jointly: introducing customer arrival models, accounting for variations in arrival rates induced by assortment and pricing decisions, allowing for concurrent arrivals, and relaxing the one-purchase-per-customer assumption. While some MNL variants address some of these issues in isolation, none fully overcome all of them in one solution. In addition, MNL models require customer-level data, while in many settings, including our case studies, only aggregated sales information is accessible. Without customer arrival data, neither arrival rates nor MNL model can be estimated due to incomplete information. Although some studies have explored MNL-based demand estimation without customer arrival data (Vulcano et al. 2012, Abdallah and Vulcano 2021, Wang 2021), they introduce additional assumptions and focus on offline settings. In contrast, our model can handle any granularity and avoids all these limitations.

Second, existing MNL models typically do not incorporate both product attributes and contextual information, but at most one of the two. Our model integrates both and captures their interaction effects in an interpretable manner. Third, MNL-based methods typically select from a fixed finite set of products, while companies usually have large catalogs and frequently need to propose new products. Moreover, companies may need to make additional decisions, such as determining the layout (e.g., display order) of products. Our

flexible vector-based action encoding can represent not only which products to offer but also factors such as their order, enabling both product design and display arrangement.

Dynamic pricing has been another important stream in revenue management and the price-demand curve is often assumed to be linear (Kleinberg and Leighton 2003, Araman and Caldentey 2009, Besbes and Zeevi 2009, Broder and Rusmevichientong 2012, den Boer and Zwart 2014, Keskin and Zeevi 2014); the paper by Den Boer (2015) provides a helpful survey. Recent work has turned towards dynamic pricing based on customer characteristics (e.g., Ban and Keskin (2021), Chen and Gallego (2021), Bastani et al. (2022)) and/or product features (e.g., Qiang and Bayati (2016), Javanmard and Nazerzadeh (2019), Cohen et al. (2020), Miao et al. (2022), Fan et al. (2022)). Much of the pricing literature focuses on single-product settings, while work on multi-product pricing is more limited (Akçay et al. 2010, Gallego and Wang 2014) and typically ignores product features and customer characteristics.

While dynamic assortment and pricing problems have been studied extensively in isolation, research addressing the joint assortment-pricing problem is relatively sparse. Chen et al. (2022a) engaged with this issue in an offline setting. More recently, Miao and Chao (2021) offer a solution using the MNL choice model. However, their approach is hampered by the limitations inherent to the MNL model, as we mentioned above. Furthermore, their model does not incorporate contextual information nor product attributes. In contrast, our work integrate both using a new model that is inherently free from these limitations.

Product Design. The product design literature, rooted in marketing, spans single product and product line design. Most existing studies assume either a deterministic first-choice model, where customers select the product with the highest utility (McBride and Zufryden 1988, Green and Krieger 1985, Belloni et al. 2008, Bertsimas and Mišić 2019), or a probabilistic choice models such as the MNL model (Chen and Hausman 2000, Li et al. 2020). The first step estimates the utility function via conjoint analysis (Green and Krieger 1993), and the focus is on the second step—solving the optimization problem, often NP-hard, using the estimated parameters from the first step without updating parameters iteratively based on new observations (i.e., offline learning). In contrast, our work is capable of designing product, alongside pricing, in an online learning framework, where we simultaneously learn the model and optimize decision policies over time.

Multi-Armed and Continuum Armed Contextual Bandits. Most online decision-making problems, including dynamic assortment and pricing, can be modeled as particular instances of a bandit problem, with the latter dating back to the seminal work of Robbins (1952). At each round, a decision-maker chooses an action (arm) 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; we refer the reader to the comprehensive book by Lattimore and Szepesvári (2020) and references therein for more background.

More recently, the literature on high-dimensional bandit problems has been an active area; it exploits a relatively mature body of statistical tools for high-dimensional problems (e.g., see the book (Wainwright

2019) and references therein). There is a line of work on contextual bandits with high-dimensional covariates, including the LASSO bandit problem (Abbasi-Yadkori et al. 2012, Kim and Paik 2019, Bastani and Bayati 2020, Hao et al. 2020, Papini et al. 2021, Xu and Bastani 2021, Chen et al. 2022b), in which the mean reward is assumed to be a linear function of a sparse unknown parameter vector. As we describe in the sequel, these high-dimensional bandit models are special cases of the high-dimensional low-rank model studied in this paper. Other work exploit non-parametric methods—among them random forests, or neural networks—to estimate the reward function (Féraud et al. 2016, Zhou et al. 2020, Ban et al. 2022, Chen et al. 2022c, Xu et al. 2022). Such approaches are quite different in flavor from our model, and 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 setup in this paper. For example, one line of research focuses on representation learning in linear bandits, specifically for low-rank bandit models and multi-task learning where several bandits are played concurrently. The actions for each task are embedded in the same space and share a common low-dimensional representation (Kveton et al. 2017, Lale et al. 2019, Yang et al. 2020, Hu et al. 2021, Lu et al. 2021, Kang et al. 2022). However, this line of research does not consider contextual information, and often imposes case-specific assumptions on the action space. Among such papers, Kang et al. (2022) study a trace inner product bandit with a matrix of known (low) rank r, in which the action is matrix-valued.

Our algorithm and theory, in contrast, are designed explicitly for contextual problems, and we do not need to know the rank r of the target matrix. Our reward model is connected to but different from other papers that propose bilinear-type reward models (e.g., Jun et al. (2019), Kim and 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.

One class of models for continuum-action bandits takes the reward function to be "smooth" over the action space, with Lipschitz or Hölder smoothness (e.g., Agrawal (1995), Kleinberg (2004), Kleinberg et al. (2019)) being typical examples. Researchers have taken different approaches to such models, including reducing the problem to a finite action space via discretization, or using non-parametric methods to estimate the reward function; both approaches are drastically different from ours. Other work on contextual bandits with continuous states-action spaces imposes Lipchitz-type conditions on the reward function jointly over the action-covariate space (Lu et al. 2010, Slivkins 2011, Krishnamurthy et al. 2020); for these reasons, it is limited to relatively low-dimensional settings. There is also other work on high-dimensional models for contextual bandits (Turğay et al. 2020). Yet, these are rooted in different models, cater to different settings, employ disparate techniques, and lack the interpretability inherent in our low-rank bilinear model.

Factor Models and Low-rank Matrix Estimation. Factor models and methods for low-rank matrix estimation and prediction have been studied extensively in both statistics and machine learning (e.g., Srebro et al. (2005), Recht et al. (2010), Candes and Plan (2010), Negahban and Wainwright (2011), Udell et al. (2016), Cai and Zhang (2018), Chen et al. (2022a)) with a wide variety of practical applications ranging from psychology (Hotelling 1933), finance and economics (Fan et al. 2021), recommendation system (Bennett et al. 2007), and electronic health records (Schuler et al. 2016). Our `Hi-CCAB` algorithm uses least-squares with nuclear norm regularization (cf. Chapter 10 in Wainwright (2019)) in the estimation block; however, its precise design and analysis requires a number of technical innovations to address the highly dependent nature of bandit data collection.

1.4. Notation

We use bold lowercase for vectors and bold uppercase for matrices. We use $\|\boldsymbol{a}\|_2$ to denote the ℓ_2-norm of vector $\boldsymbol{a}$. For a matrix $\boldsymbol{A}$, we define its Frobenius norm $|||\boldsymbol{A}|||_{\mathrm{F}} := \sqrt{\sum_{i,j} a_{ij}^2}$; its ℓ_2-spectral norm $|||\boldsymbol{A}|||_{\mathrm{op}} := \sup_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{A}\boldsymbol{x}\|_2$; and its nuclear norm $|||\boldsymbol{A}|||_{\mathrm{nuc}} := \sum_{k=1}^{d} \sigma_k(\boldsymbol{A})$, where d is the rank and $\sigma_k(\boldsymbol{A})$ are the singular values of $\boldsymbol{A}$. We use $\langle \boldsymbol{a}, \boldsymbol{b} \rangle := \boldsymbol{a}^\top \boldsymbol{b}$ to denote the Euclidean inner product between two vectors, and $\langle\langle \boldsymbol{A}, \boldsymbol{B} \rangle\rangle := \mathrm{trace}(\boldsymbol{A}^\top \boldsymbol{B})$ the trace inner product between two matrices. We use the standard notation $\mathcal{O}(\cdot)$ and $\Omega(\cdot)$ to characterize the asymptotic growth rate of a function. We use $\mathcal{P}_{\mathcal{A}}(\boldsymbol{a})$ to denote projecting $\boldsymbol{a}$ to the set $\mathcal{A}$ with respect to Euclidean distance. We use $(a \vee b)$ to denote the maximum of $\{a, b\}$, and $(a \wedge b)$ the minimum of $\{a, b\}$.

1.5. Outline

The remainder of the paper is organized as follows. We begin in Section 2 by motivating our model with our real-world case study of a manufacturer that seeks to perform joint assortment-pricing on an e-commerce platform. Equipped with this motivation, we then formally describe the doubly high-dimensional contextual bandit model. Section 3 presents the `Hi-CCAB` algorithm for representation learning and regret minimization, whereas Section 4 provides a non-asymptotic instance-dependent bound on the expected regret. Finally, Section 5 describes a suite of empirical results on simulated data and real-world case studies. We compare the performance of `Hi-CCAB` with other pricing and bandit algorithms as well as apply `Hi-CCAB` to two case studies on real sales data from a food manufacturer and a start-up. We conclude with a summary and discussion of future research in Section 6. Proofs and additional empirical results are provided in the online appendices in the supplemental material.

2. Problem Motivation and Formulation

In this section, we begin by motivating the class of problems studied in this paper with a concrete example. We then provide a more precise formulation of the problem.

2.1. A Real-world Instance of a Doubly High-dimensional Bandit

We motivate our model by introducing the assortment-pricing problem faced by a market-leading instant noodle manufacturer in China. This company has 30 slots on their main page for displaying products with corresponding prices. In addition to its existing products, it frequently proposes new products; from March 1, 2021 to May 31, 2022, a total of 176 products appeared in its catalog. The task is to determine, at each time step (e.g., each day), which 30 products to offer from the set of existing and potential products and set prices. Even if we limit the decision to selecting from existing products without proposing new products and ignore pricing, the resulting combinatorial problem of $\binom{176}{30} \approx 6.4 \times 10^{33}$ is intractable without further modeling. This intractability motivates the need for an efficient representation of the action space. To address this challenge and enable product design, we leverage (possibly high-dimensional) product attributes and represent the entire assortment as one vector. Combined with pricing, the proposed representation for the joint assortment-pricing decision lives in a continuous and high-dimensional space.

Meanwhile, the company also has at its disposal a rich array of contextual information, including macro-environmental information such as season, location, and specific holidays. In addition, in certain cases, additional micro-level information is also available, such as users' profile information and historical data. Encoding this side information also leads to a high-dimensional context vector.

The company makes decisions and observes feedback in an iterative and sequential fashion. To be specific, at the beginning of each period (e.g., day, week, or year), the company jointly decides on the assortment and pricing. After doing so, it observes the revenue/profit, which we refer to as the reward, at the end of each period. The firm needs to learn the reward function with respect to different assortment and pricing given the contextual information on the fly and make the optimal assortment and pricing decision that maximizes the cumulative reward across the time horizon. This exploration-exploitation decision-making problem can be modeled as a contextual bandit problem, where both the arm (i.e., action/decision) and contextual vectors take continuous values in high-dimensional spaces.

High-dimensional actions and covariate arise in many applications. Without imposing additional structure on high-dimensional bandit problems, one cannot expect to obtain non-trivial guarantees (due to "no-free-lunch" theorems). Accordingly, it is essential to impose structure, and in this paper, we posit low-dimensional structure in the form of a small number of latent factors that control interaction effects between actions and covariates in determining the expected reward.

2.2. Formalizing the Model

With this intuition in place, let us formalize the class of models that we study in this paper. We consider a firm that makes assortment-pricing decisions over a period of T rounds, indexed by $t \in [T] := \{1, 2, \ldots, T\}$. Each can be of any predetermined granularity (e.g., by day, week, month, or by the arrival of one customer). There are a total of K product slots to be filled, indexed by $k \in [K] := \{1, \ldots, K\}$.

2.2.1. Action and Context Vectors. At each time $t \in [T]$, the product in slot $k \in [K]$ is associated with an m-dimensional attribute vector $\boldsymbol{f}_{t,k} \in \mathbb{R}^m$ along with a non-negative price $p_{t,k} \in [0, \infty)$. Features encoded by the vector $\boldsymbol{f}_{t,k}$ depend on the product, but might include color, flavor, material, and technical specifications, etc. Collecting together all the attribute vectors and prices across the K products, we obtain the *action vector* at time $t \in [T]$, given by

$$\boldsymbol{a}_t := \big(\boldsymbol{f}_{t,1}, p_{t,1}, \boldsymbol{f}_{t,2}, p_{t,2}, \cdots, \boldsymbol{f}_{t,k}, p_{t,k}, \cdots, \boldsymbol{f}_{t,K}, p_{t,K}, 1\big), \tag{1}$$

where $(\boldsymbol{f}_{t,k}, p_{t,k})$ denotes the product feature and price for slot k at time t. The special notation $(\boldsymbol{f}_{t,k} = \boldsymbol{0}, p_{t,k} = 0)$ indicates that the slot k being empty. Note that this action vector $\boldsymbol{a}_t$ has $d_a := K(m+1)+1$ components in total, and is thus high-dimensional for the values of (K, m) typical in practice.

It is worth mentioning that the order of $(\boldsymbol{f}_{t,k}, p_{t,k})$ matters as products on top of the list are likely to draw customer attention. Our model is capable of retaining this ordering information and takes actions accordingly. Generally, $\boldsymbol{a}_t$ is not confined to such specific form and can take value in an action set $\mathcal{A}_t \subseteq \mathbb{R}^{d_a}$ to tailor to any specific tasks.

For each period $t \in [T]$, the firm also observes side-information in the form of a *context vector* $\boldsymbol{x}_t \in \mathbb{R}^{d_x}$. This context vector $\boldsymbol{x}_t$ can include individual or aggregated customer information depending on the granularity or macro-environmental factors. Thus, the dimension d_x can be large. We further assume that $\boldsymbol{x}_t$ is independent of observations prior to t, including the firm's own decisions.

2.2.2. Reward Structure. The goal of the firm is to make assortment-pricing decisions, via their choice of the action vector $\boldsymbol{a}_t$ at each time t, so as to maximize revenue. At time t, we model the revenue in terms of its conditional expectation given an action-covariate pair $(\boldsymbol{a}_t, \boldsymbol{x}_t)$. In particular, we assume that the reward Y_t has a conditional mean function of the bilinear form

$$\mathbb{E}[Y_t \mid \boldsymbol{x}_t, \boldsymbol{a}_t] = \boldsymbol{a}_t^T \boldsymbol{\Theta}^* \boldsymbol{x}_t, \tag{2}$$

where $\boldsymbol{\Theta}^* \in \mathbb{R}^{d_a \times d_x}$ is an unknown representation matrix.

Such a bilinear structure is simple yet powerful. The matrix $\boldsymbol{\Theta}^*$ captures interactions between the action vector $\boldsymbol{a}_t$ and the covariate vector $\boldsymbol{x}_t$ in determining the expected reward. By formulating the action vector by concatenating the product attribute vectors (cf. equation (1)), we can take advantage of the similarity between different assortments as products with similar features often have similar rewards under similar context. The advantages of exploiting product and pricing features' predictive power in sales patterns are enormous in that it can reduce a possibly ultra high-dimensional problem (e.g., assuming all $\binom{N}{K} = \binom{176}{30} \approx 6.4 \times 10^{33}$ assortments are irrelevant) that forbids any analysis to a high-dimensional problem, and more importantly it allows the company to propose new products into the assortment—a crucial requirement for manufacturers. The bilinear reward structure also facilitates interpretability of the model, through analyzing the relationships between action-covariate-reward tuple characterized by $\boldsymbol{\Theta}^*$, as

elaborated in Section 2.4. We will demonstrate how the bilinear form not only unifies a number of structure bandit and pricing models but also can be generalized to capture non-linear effects in Section 2.5.

Our model can be generalized to a collection of $L \geq 2$ of different targets—say different platforms or geographic locations. Indexing the targets by $\ell \in [L] := \{1, 2, \ldots, L\}$, at each time t, we observe a collection of covariate vectors $\{\boldsymbol{x}_{t,\ell}\}_{\ell=1}^{L}$ and apply the same decision $\boldsymbol{a}_t$ to all targets. We then observe a batch of rewards $\{y_{t,\ell}\}_{\ell=1}^{L}$, and model the conditional mean $Y_{t,\ell}$ as follows

$$\mathbb{E}[Y_{t,\ell} \mid \boldsymbol{x}_{t,\ell}, \boldsymbol{a}_t] = \boldsymbol{a}_t^\top \boldsymbol{\Theta}^* \boldsymbol{x}_{t,\ell} \quad \text{for } \ell = 1, 2, \cdots, L. \tag{3}$$

For simplicity, we assume that the reward function of each target location is independent, but it is possible to extend our model to account for dependency.

2.3. Firm's Objective and Regret

The objective of the firm is to design a policy π that chooses a sequence of history-dependent actions $(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_T)$ so as to maximize the *expected cumulative revenue*

$$\mathbb{E}_\pi \left[\sum_{t=1}^{T} \sum_{\ell=1}^{L} \boldsymbol{a}_t^\top \boldsymbol{\Theta}^* \boldsymbol{x}_{t,\ell} \right]. \tag{4}$$

If the representation matrix $\boldsymbol{\Theta}^*$ is known a priori, then the firm can choose an optimal decision $\boldsymbol{a}_t^* \in \mathcal{A}_t$ that maximizes the sum of the reward functions (3) across L targets, i.e., $\boldsymbol{a}_t^* := \sup_{\boldsymbol{a} \in \mathcal{A}_t} \sum_{\ell=1}^{L} \boldsymbol{a}^\top \boldsymbol{\Theta}^* \boldsymbol{x}_{t,\ell}$. We call this optimal solution a *clairvoyant solution* and the clairvoyant revenue over the time horizon is given by $\sum_{t=1}^{T} \sum_{\ell=1}^{L} \boldsymbol{a}_t^{*\top} \boldsymbol{\Theta}^* \boldsymbol{x}_{t,\ell}$. Of course, this clairvoyant value is not attainable because $\boldsymbol{\Theta}^*$ is unknown in practice, but it serves as a useful benchmark for performance of any algorithm.

With this benchmark in place, we evaluate policy π through *cumulative regret*—that is, the gap between the expected cumulative revenue over the time horizon T between the revenue earned by implementing policy π, and the clairvoyant solution. Equivalently, we seek to minimize the *time-averaged regret*[1]—that is, the quantity

$$\mathcal{R}^\pi(T) := \frac{1}{T} \mathbb{E}_\pi \left[\sum_{t=1}^{T} \sum_{\ell=1}^{L} \boldsymbol{a}_t^{*\top} \boldsymbol{\Theta}^* \boldsymbol{x}_{t,\ell} - \boldsymbol{a}_t^\top \boldsymbol{\Theta}^* \boldsymbol{x}_{t,\ell} \right] = \frac{1}{T} \mathbb{E}_\pi \left[\sum_{t=1}^{T} \sum_{\ell=1}^{L} (\boldsymbol{a}_t^* - \boldsymbol{a}_t)^\top \boldsymbol{\Theta}^* \boldsymbol{x}_{t,\ell} \right]. \tag{5}$$

Since the representation matrix $\boldsymbol{\Theta}^*$ is unknown to us, we need to design an algorithm that simultaneously learns the representation matrix on the fly (exploration) and maximizes the total revenue (exploitation). This exploration–exploitation problem with high-dimensional action and covariate spaces, to which we refer as a *doubly high-dimensional contextual bandit*, is our focus.

2.4. Low-Rank Structure of $\boldsymbol{\Theta}^*$ and Its Implications

As argued previously, although actions and covariates are high-dimensional, the demand and sales are often driven by certain latent factors. Therefore, it is reasonable to impose a low-rank assumption on the representation matrix $\boldsymbol{\Theta}^*$.

[1] The time-averaged form of regret is rescaled by $1/T$ relative to the cumulative regret; we do so with the intent that our bounds can be stated in the form of standard consistency guarantees, with the error decreasing to zero as T increases.

To understand the meaning of such a low-rank condition, consider a matrix $\boldsymbol{\Theta}^*$ that is of rank $r \ll \min\{d_a, d_x\}$. It 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 (2) can be decomposed as

$$\mathbb{E}[Y_t \mid \boldsymbol{a}_t, \boldsymbol{x}_t] = \boldsymbol{a}_t^\top \boldsymbol{\Theta}^* \boldsymbol{x}_t = \sum_{j=1}^{r} s_j \langle \boldsymbol{a}_t, \boldsymbol{u}_j \rangle \cdot \langle \boldsymbol{v}_j, \boldsymbol{x}_t \rangle . \tag{6}$$

In other words, the mean reward is the summation of the products between the action projected on the left singular vector and the covariates projected on the right singular vector, weighted by the singular values. The low-rank condition on $\boldsymbol{\Theta}^*$ means that the expected reward is governed by a relatively small number of interactions between linear combinations of the action features and covariates. In this way, our model automatically explores the low-dimensional structure of the action and context vectors in terms of their effects on the reward via the left and right singular vectors; consequently, we can draw conceptual and modeling insights from the spectral structures of $\boldsymbol{\Theta}^*$. In the context of joint assortment-pricing, the left singular vectors $\boldsymbol{u}_j$ (respectively, the right singular vectors $\boldsymbol{v}_j$) can be thought of as weights associated with the latent products factor j, which loads on their attributes and prices (respectively, the latent covariate factor j).

We note that our empirical studies provide evidence for the suitability of the low-rank structure of $\boldsymbol{\Theta}^*$. For instance, in our instant noodle manufacturer example, for each product, there are 13 possible flavors along with the price, so a total of 14 attributes. In addition to including these attributes themselves, we also include their squares (so that we can model non-linear effects), for a total of 28 meta-attributes. We include these 28 meta-attributes for each of $K = 30$ possible product slots considered, leading to an action vector of dimension

$$d_a = 30K + 1 \;=\; 841,$$

where the additional one accounts for the presence of a constant offset term. In terms of covariates, we include 31 provinces, the year 2021/2022, 12 months, weekdays, an indicator of the annual sale events and an additional one, leading to a covariate vector of dimension $d_x = 50$. These vector representations of the decision and context account for the intercept, the main effects, and the interaction effects of decision and context. For this pair $(d_a, d_x) = (841, 50)$, our procedure learns a matrix $\widehat{\boldsymbol{\Theta}}$ with rank 4. See Section 5.3 for further discussion of the latent factors, and their real-world significance.

2.5. Relation to Other Models

In this section, we discuss how our model is related to other known bandit models and approaches to dynamic pricing and assortment (see Sections 2.5.1, 2.5.2 and 2.5.3, respectively).

2.5.1. Connection with Other Bandit Models. Let us summarize some connections to other bandit models that can be expressed as special cases of our reward model (3). In this section, we recycle the notation K, using it to represent the number of actions in the multi-action bandit by convention.[2]

1. A *multi-armed bandit* is defined by K independent actions (Robbins 1952). The k-th action can be represented by the unit basis vector $\boldsymbol{a}_k = (0, 0, \cdots, 1, \cdots, 0)^\top$, where the single 1 appears in the k-th entry. By setting $\boldsymbol{x} = 1$ and $\boldsymbol{\Theta}^* \in \mathbb{R}^{K\times 1}$ be the rank-one matrix with entries $\boldsymbol{\Theta}^*_{k1} = \mu_k$, we have $\boldsymbol{a}_k^\top \boldsymbol{\Theta}^* \boldsymbol{x} = \mu_k$ as a special case of our model. The *linear bandit* (e.g., Rusmevichientong and Tsitsiklis (2010), Dani et al. (2008), Auer (2002), Abbasi-Yadkori et al. (2011)) is a natural generalization of the multi-armed bandit, in which each of the K possible actions is associated with an arbitrary vector $\boldsymbol{a}_k$, and the reward function is a mapping $\boldsymbol{a} \mapsto \mu(\boldsymbol{a}) = \langle \boldsymbol{\theta}, \boldsymbol{a} \rangle$. Using $\boldsymbol{a}$ with $\boldsymbol{x} = 1$ as the "context", we can write this model in the form $\boldsymbol{\Theta}^* = \boldsymbol{\theta}$, again leading to a rank-one setting.
2. In *high-dimensional contextual* multi-armed bandits (Bastani and Bayati 2020), in addition to the K arms—each represented by a unit basis action vector $\boldsymbol{a}_k$ as above—we also have a (possibly high-dimensional) context vector $\boldsymbol{x} \in \mathbb{R}^{d_x}$. The reward associated with arm k is given by $\langle \boldsymbol{\beta}_k, \boldsymbol{x} \rangle$. By defining the matrix $\boldsymbol{\Theta}^* = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_K)^\top \in \mathbb{R}^{K\times d_x}$, we can represent this model in our bilinear form.
3. *Continuum-action bandits (without context).* Given a continuous action $b \in \mathbb{R}$, these models (e.g., Kleinberg et al. (2019)) use a general non-parametric reward function $b \mapsto \mu(b)$. Such models are actually non-parametric in nature, but can be approximated by linear bandits by lifting the action space. More precisely, since all continuous functions on a bounded interval can be approximated by polynomial functions to arbitrary precision, we can approximate the reward function using a polynomial of order at most N. Defining the augmented action vector $\boldsymbol{a} = (1, b, b^2, b^3, \cdots, b^N)$, we then have a linear bandit in dimension $N + 1$.

2.5.2. Compatibility with Pricing Models. The linear price-demand model plays a central role in dynamic pricing literature. This linear demand model is a special case of our bilinear reward model. We focus on a recent extension to the linear price-demand curve proposed by Ban and Keskin (2021) which considers the personalized pricing problem and assumes a personalized demand model whose parameters depend on the context vector. Specifically, they assume the demand model as

$$D_t = \boldsymbol{\alpha}^T \boldsymbol{x}_t + (\boldsymbol{\beta}^T \boldsymbol{x}_t) p_t + \epsilon_t \tag{7}$$

where $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^{d_x}$ are the unknown demand parameter vectors, and $p_t \in \mathbb{R}^+$ denotes the price, $\boldsymbol{x}_t \in \mathbb{R}^{d_x}$ the customer characteristics, and ϵ_t the noise. In this model, the inner product $\langle \boldsymbol{\alpha}, \boldsymbol{x}_t \rangle$ captures the "context-dependent customer taste and potential market size", whereas the inner product $\langle \boldsymbol{\beta}, \boldsymbol{x}_t \rangle$ captures the "context-dependent price sensitivity." Therefore, the expected revenue at time t is

$$\mathbb{E}[Y_t \mid \boldsymbol{x}_t, p_t] = p_t \big[\langle \boldsymbol{\alpha}, \boldsymbol{x}_t \rangle + \langle \boldsymbol{\beta}, \boldsymbol{x}_t \rangle \, p_t \big]. \tag{8}$$

[2] Please note, this should not be confused with the maximum number of slots in our general model set-up.

Note that the mean reward (8) is a special case of our model with action $\boldsymbol{a}_t = (p_t, p_t^2)$, covariate $\boldsymbol{x}_t$ being the same as in the demand model (7), and unknown parameter matrix $\boldsymbol{\Theta}^* = (\boldsymbol{\alpha}, \boldsymbol{\beta})^\top \in \mathbb{R}^{2\times d_x}$.

2.5.3. Connections to MNL models and Generalizations to Non-Linear Models. As we mentioned in Section 1.3 and demonstrated through our case study in Section 2.1, the MNL models have attracted much attention for the assortment problem, but have limitations that make them infeasible in practice. Our bilinear model is free from these limitations. In fact, the expected reward of MNL models can be expressed via a generalized bilinear model: $\mathbb{E}[Y_{t,\ell}|\boldsymbol{x}_{t,\ell}, \boldsymbol{a}_t] = g(\boldsymbol{a}_t^\top \boldsymbol{\Theta}^* \boldsymbol{x}_{t,\ell})$ with the following specification of action vector, covariate vector, parameter matrix, and link function g. Let $g(x) = \frac{1}{x}$, covariates $\boldsymbol{x}_{t,\ell}$ be a one hot vector with the ℓ-th element being 1 and all others being 0, and $\boldsymbol{\Theta}^*$ be an $N^2 \times N$ matrix:

$$\boldsymbol{\Theta}^{*\top} = \begin{pmatrix} \frac{1}{p_1} + \frac{1}{p_1\mu_1} & 0 & 0 & \dots & 0 & \vec{\mu}_1 & \vec{0} & \vec{0} \dots & \vec{0} \\ 0 & \frac{1}{p_2} + \frac{1}{p_2\mu_2} & 0 & \dots & 0 & \vec{0} & \vec{\mu}_2 & \vec{0} \dots & \vec{0} \\ \vdots & & \ddots & & \vdots & \vdots & & \ddots & \vdots \\ 0 & \dots & & \frac{1}{p_N} + \frac{1}{p_N\mu_N} & & \vec{0} & \dots & \vec{0} & \vec{\mu}_N \end{pmatrix}$$

where $\vec{\mu}_j = (\frac{\mu_1}{\mu_j p_j}, \frac{\mu_2}{\mu_j p_j}, \dots, \frac{\mu_{j-1}}{\mu_j p_j}, \frac{\mu_{j+1}}{\mu_j p_j}, \dots, \frac{\mu_N}{\mu_j p_j})$ and μ_j denotes the utility of product j. For action $\boldsymbol{a}_t$, denote $\boldsymbol{a}_t = (\boldsymbol{a}_{t,0}, \boldsymbol{a}_{t,1}, \boldsymbol{a}_{t,2}, \dots, \boldsymbol{a}_{t,N}) \in \{0,1\}^{N^2}$ where $\boldsymbol{a}_{t,0} \in \{0,1\}^N$ is an indicator vector of whether each product is in the assortment, i.e., i-th element is 1 if product i is in the assortment else 0; and $\boldsymbol{a}_{t,j} \in \{0,1\}^{N-1}$ corresponds to $\vec{\mu}_j$ for $j = 1, \dots, N$, whose element corresponding to $\frac{\mu_i}{\mu_j p_j}$ is 1 if both products i and j are in the assortment and 0 otherwise. Then

$$\sum_{\ell=1}^{N} g(\boldsymbol{a}_t^\top \boldsymbol{\Theta}^* \boldsymbol{x}_{t,\ell}) = \sum_{\ell=1}^{N} \mathbb{1}\{\ell \in \mathcal{S}\} \frac{p_\ell \mu_\ell}{1 + \sum_{\ell' \in \mathcal{S}} \mu_{\ell'}},$$

where the right-hand side corresponds to the expected return of MNL models.

The bilinear form is flexible and expressive—as shown above, it can be extended to generalized bilinear models that include MNL models. Even without the link function, it can capture non-linear effects and be further extended to Reproducing Kernel Hilbert Spaces (RKHS), as detailed in Section EC.1. We leave formal analysis of these models to future work, while our current analysis is general, offering ideas and results that can be carried to the aforementioned settings.

3. Algorithm

In this section, we describe our learning algorithm for the doubly high-dimensional contextual bandit problem. It involves two steps at each time period and is thus modular and generalizable. The first step learns a low-rank representation by constructing an estimate, $\widehat{\boldsymbol{\Theta}}_t$, using a penalized form of least-squares regression that involves action-covariate pairs $(\boldsymbol{a}_i, \boldsymbol{x}_{i,\ell})$ and responses $y_{i,\ell}$ for $i = 1, \dots, t$ and $\ell = 1, \dots, L$. In the second step, we use the estimated bilinear reward induced by $\widehat{\boldsymbol{\Theta}}_t$ to choose assortment-price actions within the action space $\mathcal{A}_t$. See Algorithm 1 for the full details.

Algorithm 1: The `Hi-CCAB` Algorithm.

Result: Actions $\boldsymbol{a}_{t_{init}+1}, \ldots, \boldsymbol{a}_T$.

Input: Initial step number t_{init}; set of possible actions $\mathcal{A}_{t_{init}}$, action vectors based on domain knowledge $\{\boldsymbol{a}_i\}_{i=1}^{t_{init}}$, covariate vectors $\{\boldsymbol{x}_{i,\ell}\}_{i=1}^{t_{init}}$, rewards $y_{i,\ell}$ for $\ell = 1, \ldots, L$, exploration parameter h and $\widehat{\boldsymbol{\Theta}}_{t_{init}}$.

Initialization: $\lambda_0 \leftarrow \frac{2}{t_{init} L} \| \sum_{i=1}^{t_{init}} \sum_{\ell=1}^{L} \left| \boldsymbol{a}_i^\top \widehat{\boldsymbol{\Theta}}_{t_{init}} \boldsymbol{x}_{i,\ell} - y_{i,\ell} \right| \boldsymbol{x}_{i,\ell} \boldsymbol{a}_i^\top \|_{\text{op}}$, and $t \leftarrow t_{init}$.

while $t < T$ **do**

$t \leftarrow t+1$; $\lambda_t \leftarrow \lambda_0/\sqrt{t}$;

Step 1: Low-rank representation learning:

$\widehat{\boldsymbol{\Theta}}_t \leftarrow \arg\min_{\boldsymbol{\Theta}} \left\{ \frac{1}{2tL} \sum_{i=1}^{t} \sum_{\ell=1}^{L} (\boldsymbol{a}_i^\top \boldsymbol{\Theta} \boldsymbol{x}_{i,\ell} - y_{i,\ell})^2 + \lambda_t \|\boldsymbol{\Theta}\|_{\text{nuc}} \right\}$;

Step 2: Policy learning (choosing action):

$\hat{\boldsymbol{a}}_{t+1} \leftarrow \arg\max_{\boldsymbol{a} \in \mathcal{A}_t} \left\{ \sum_{\ell=1}^{L} \boldsymbol{a}^\top \widehat{\boldsymbol{\Theta}}_t \boldsymbol{x}_{t+1,\ell} \right\}$ (take any one with the largest norm if the solution is not unique);

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: $\boldsymbol{a}_{t+1} \leftarrow \mathcal{P}_{\mathcal{A}_t}(\hat{\boldsymbol{a}}_{t+1} + \boldsymbol{\delta}_{t+1})$ where $\boldsymbol{\delta}_{t+1} \sim N(\boldsymbol{0}_{d_a}, h^2 \boldsymbol{I}_{d_a})$, update action space $\mathcal{A}_{t+1}$;

Apply action $\boldsymbol{a}_{t+1}$ and observe reward $y_{t+1,\ell}$ for $\ell = 1, \ldots, L$;

end while

3.1. Step 1: Low-rank Representation Learning.

The first step of the algorithm is to estimate the low-rank representation matrix $\boldsymbol{\Theta}^*$. As motivated in Section 2, it is reasonable to impose a low-rank condition on $\boldsymbol{\Theta}^*$. Disregarding computational issues, one might imagine estimating $\boldsymbol{\Theta}^*$ by imposing a rank constraint, or a penalty involving the rank. However, rank penalization results in a non-convex problem with associated computational challenges, so 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

$$\widehat{\boldsymbol{\Theta}}_t := \arg\min_{\boldsymbol{\Theta}} \left\{ \frac{1}{2Lt} \sum_{i=1}^{t} \sum_{\ell=1}^{L} \left(\boldsymbol{a}_i^\top \boldsymbol{\Theta} \boldsymbol{x}_{i,\ell} - y_{i,\ell} \right)^2 + \lambda_t \cdot \|\boldsymbol{\Theta}\|_{\text{nuc}} \right\}, \tag{9}$$

where $\lambda_t > 0$ is a regularization parameter. We update the parameter λ_t over the time periods with $\lambda_t = \frac{\lambda_0}{\sqrt{t}}$, where $\lambda_0 > 0$ is an initial choice. The decay rate $1/\sqrt{t}$ is chosen to match the typical standard deviation of the first data-dependent term: with L being constant, it is the sample average of t terms.

Now we turn to the choice of λ_0. While any $\lambda_0 > 0$ works, the most ideal choice matches the scale of the gradient of the first (data-dependent) term in Equation (9). When no historical data is available, we rely on domain knowledge and educated guesses to approximate this scale. When historical data is available, we determine λ_0 according to the Initialization step in Algorithm 1, based on a pre-specified initial estimator $\widehat{\boldsymbol{\Theta}}_{t_{init}}$. A simple yet effective choice of $\widehat{\boldsymbol{\Theta}}_{t_{init}}$ is $\boldsymbol{0}_{d_a \times d_x}$. Ideally, $\widehat{\boldsymbol{\Theta}}_{t_{init}}$ should be as close as possible to $\boldsymbol{\Theta}^*$,

though the performance gain from doing so is modest. One approach to specify a possibly closer $\widehat{\boldsymbol{\Theta}}_{t_{init}}$ is to select λ through cross-validation and then use the corresponding $\widehat{\boldsymbol{\Theta}}$ as $\widehat{\boldsymbol{\Theta}}_{t_{init}}$.

3.2. Step 2: Policy Learning (Choosing Actions).

Given an estimate of the low-rank matrix $\boldsymbol{\Theta}^*$, we can proceed to the action step, i.e., to select the assortment and pricing for time t. 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 at later steps to achieve higher long-term rewards. Specifically, given the estimate $\widehat{\boldsymbol{\Theta}}_t$ and the covariate vectors $\boldsymbol{x}_{t+1,\ell}$ for $\ell \in [L]$, we look for an action $\hat{\boldsymbol{a}}_{t+1}$ in the action space $\mathcal{A}_t$ that maximizes the expected total rewards across L objects:

$$\hat{\boldsymbol{a}}_{t+1} := \arg\max_{\boldsymbol{a}\in\mathcal{A}_t} \left\{ \sum_{\ell=1}^{L} \boldsymbol{a}^\top \widehat{\boldsymbol{\Theta}}_t \boldsymbol{x}_{t+1,\ell} \right\}. \tag{10}$$

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: $\boldsymbol{a}_{t+1} = \mathcal{P}_{\mathcal{A}_t}(\hat{\boldsymbol{a}}_{t+1} + \boldsymbol{\delta}_{t+1})$ where $\boldsymbol{\delta}_{t+1} \sim N(\mathbf{0}_{d_a}, h^2 \boldsymbol{I}_{d_a})$ and h is a tuning parameter. The choice of h in practice and alternative perturbation strategies will be discussed shortly. 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 less 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 affects trade-offs between different terms in the regret, as discussed later in Remarks 7 and 8.

The form of randomness in exploration, in addition to the intensity h, is another tuning parameter of the algorithm. Suppose $\hat{\boldsymbol{\tau}}_t$ is a vector with each element being the coordinate-wise squared standard error of the past actions up to time t, i.e., $\{\boldsymbol{a}_i\}_{i=1}^t$. For each exploration step at time t, in practice, one can explore isotropically by setting $h = c_0 \sqrt{\frac{1}{d_a}\sum_{j=1}^{d_a} \hat{\boldsymbol{\tau}}_t}$, where $c_0 > 0$ is a small positive constant depending on the company's risk tolerance. The more risk-averse the company is, the smaller c_0 should be. While c_0 can occasionally be as large as 1 (e.g., when the assortment and pricing change frequently), it is typically very small (e.g., ~ 0.01). Alternatively, it is preferable to explore anisotropically by letting $\boldsymbol{\delta}_{t+1} \sim N(\mathbf{0}_{d_a}, c_0^2 \operatorname{diag}(\hat{\boldsymbol{\tau}}_t))$, which scales all directions appropriately. Finally, for the action space $\mathcal{A}_{t+1}$, we can either hold it fixed throughout as a hard constraint set or update it accordingly. For example, if $\mathcal{A}_t \in \mathbb{R}^{d_a}$ can be defined by an upper limit $\bar{\boldsymbol{a}}_t$ and a lower limit $\underline{\boldsymbol{a}}_t$, we simply expand the action space by pushing the boundary of each coordinate to $\hat{\boldsymbol{a}}_{t+1,j} + \boldsymbol{\delta}_{t+1,j}$ if it falls outside the interval $[\underline{\boldsymbol{a}}_{t,j}, \bar{\boldsymbol{a}}_{t,j}]$ for $j = 1, \ldots, d_a$.

Remark 1 (Initialization). The initial step number t_{init} and actions $\{\boldsymbol{a}_i\}_{i=1}^{t_{init}}$ depends on the availability of historical data. When there exists historical data, t_{init} is the number of steps in the historical data, and the actions are the corresponding real actions. The real actions are often guided by "domain knowledge," such as market research, past experience, and heuristics. Historical data is often available in real applications

so we design our algorithm to take advantage of all the data available. In the absence of historical data, actions can either be domain-informed or randomly selected within the action set for a reasonably small number of steps.

Remark 2 (Practicality: Adaptivity). Our algorithm is adaptive w.r.t. time T and rank r—it does not have T-dependent or r-dependent tuning parameters, yet the performance adapts to the difficulty of the problem dictated by T and r (e.g., the average regret converges to zero as $T \to \infty$ and is small for small r as shown in Theorem 1). This adaptivity contrasts with common approaches such as explore-then-commit (ETC) and confidence-bound (e.g., UCB) type algorithms. The former requires a T-dependent parameter to specify the exploration length. While the doubling trick (Lattimore and Szepesvári 2020, Besson and Kaufmann 2018) can, in theory, make ETC anytime, it suffers from practical limitations: each restart discards previously collected data, leading to both inefficiency and spiky regret; and the exponentially increasing length of each exploration phase results in progressively longer periods of poor performance, which can be catastrophic in practice (see Remark 4). Confidence-bound algorithms rely on the analysis of estimation error, making them inherently r-dependent. Such adaptivity is of practical importance, as companies desire algorithms that consistently perform well across time horizons and ranks that are not known a priori.

Remark 3 (Practicality: Robustness). Our algorithm is robust, in the sense that it continues to perform well even when model assumptions are violated, which is highly valued in practice. Our algorithm keeps exploring and remains effective even if the underlying $\mathbf{\Theta}^*$ changes slightly over time (e.g., if $\mathbf{\Theta}^*$ changes abruptly at round $\sqrt{T}$, the averaged regret of our algorithm still converges to zero following the same proof idea of Theorem 1). On the contrary, ETC-type algorithms stop exploring after the initial exploration phase, usually of an order way smaller than $O(T)$, making them far off when $\mathbf{\Theta}^*$ changes. Moreover, our algorithm is robust to misspecification of r, a common challenge in practice. It can adjust to r while the UCB-type algorithms fail. In general, UCB-type algorithms are more sensitive to assumptions because they are built upon analysis that is heavily assumption-laden.

Remark 4 (Practicality: No consecutive random exploration). Our algorithm avoids prolonged consecutive random exploration, which can be risky and even survival-threatening in practice. Theoretically, when T is known, front-loading exploration is appealing since it allows the collected information to inform decision-making for the longest possible period of time and simplifies analysis due to independent samples. However, such strategy lacks adaptivity (Remark 2) and poses significant risk: the resulting extended consecutive periods of poor performance can jeopardize firm survival, even for the large ones. Our algorithm spreads out exploration over time to gain adaptivity and mitigate the risk of consecutive explorations. Additionally, we adopt an *informed* exploration by adding a small perturbation to the best action found so far, thereby further minimizing the negative impact of explorations.

Remark 5 (Interpretability). To take advantage of the interpretability of our model, we can further explore the structure of the $\widehat{\mathbf{\Theta}}_t$. Specifically, we can apply singular value decomposition (SVD) on $\widehat{\mathbf{\Theta}}_t$ to

explore the underlying latent structure of the arms (respectively, covariates) from the left (respectively, right) singular vectors. One can further rotate the singular vectors using techniques in factor analysis such as Varimax (Kaiser 1958, Rohe and Zeng 2023) so as to obtain a sparse/simplified loading structure for easier interpretation.

Remark 6 (Computational efficiency). Our algorithm is computationally efficient in both Step 1 and Step 2. For Step 1, the penalized least-squares problem (9) is convex and can be computed efficiently with precise theoretical justifications (Chen 2022, Chapter 4). For Step 2, the optimization problem (10) is either convex or can be decomposed into a series of univariate optimization problems involving polynomial functions with convex constraints, each of which can be solved efficiently. Moreover, for the case where one selects from a fixed pool of products rather than proposing new products, the procedure is also computationally efficient: for each slot, we only need to search over the product pool and select the one whose attribute vector has the largest inner product with the corresponding segment of the action vector obtained in Step 2.

4. Regret Analysis

We now turn to some theoretical analysis of our procedure, beginning in Section 4.1 with the statement of our main theorem, and with the following Section 4.2 devoted to proofs.

4.1. Instance-dependent Regret Bound

We begin by stating a non-asymptotic instance-dependent bound on the expected time-averaged regret incurred by Algorithm 1. It shows that for any problem and for any dimension, the expected time-averaged regret decays to zero at least as fast as $\tilde{\mathcal{O}}(T^{-1/6})$.

Our analysis applies to an instantiation of Algorithm 1 with no historical data and actions randomly chosen according to the exploration protocol $\boldsymbol{a}_{t+1} = \mathcal{P}_{\mathcal{A}_t}(\hat{\boldsymbol{a}}_{t+1} + \boldsymbol{\delta}_{t+1})$, where $\boldsymbol{\delta}_{t+1} \sim N(\mathbf{0}_{d_a}, h^2 \boldsymbol{I}_{d_a})$ for a pre-specified $h > 0$, implemented at each time instant

$$t \in \{\lfloor w^{\frac{3}{2}} \rfloor \mid w = 1, 2, 3, \ldots\}. \tag{11}$$

The constraint set for this instance is $\mathcal{A}_t = \{\boldsymbol{a} \in \mathbb{R}^{d_a} : \|\boldsymbol{a}\| \leq 1\}$ for all $t \geq 1$, and the first action is selected randomly. We also pick a $\lambda_0 > 0$ beforehand. As our analysis involves an additional assumption on the reward error, we introduce a short-hand notation for the reward error,

$$\varepsilon_{t,\ell} = y_{t,\ell} - \boldsymbol{a}_t^T \boldsymbol{\Theta}^* \boldsymbol{x}_{t,\ell}. \tag{12}$$

Finally, our statement involves a burn-in period B_{init}, which is a function of h, L, λ_0, d_x, and d_a. It has an upper bound, as stated in Section 4.2.1 and detailed in Section EC.2.1.

Theorem 1. *Suppose that the ground truth $\boldsymbol{\Theta}^*$ has rank r, we observe covariates $\boldsymbol{x}_{t,\ell} \overset{i.i.d}{\sim} N(\mathbf{0}_{d_x}, \boldsymbol{I}_{d_x})$, and the reward errors $\varepsilon_{t,\ell} \overset{i.i.d}{\sim} N(0, \sigma^2)$. Then, there are universal constants $\{c_j\}_{j=1}^3$ such that for any $h, \lambda_0 > 0$ and for all $T \geq B_{init}$, the expected time-averaged regret is bounded as*

$$\mathcal{R}^{\pi}(T) \leq \frac{c_1}{T}\sqrt{L}\|\boldsymbol{\Theta}^*\|_F B_{init} + \frac{c_2 \log T}{T}\sqrt{L}\|\boldsymbol{\Theta}^*\|_F + \frac{1 \wedge 2h\sqrt{d_a}}{T^{1/3}}\sqrt{L}\|\boldsymbol{\Theta}^*\|_F$$

$$+ \frac{c_3\sqrt{L}}{v(d_a,h)} T^{-1/6} \left(\left(\sqrt{d_x \log TL} + 2\log TL \right) \sigma + \lambda_0 \sqrt{r} \right), \quad (13)$$

where $v(d_a,h) := \frac{1}{d_a-1} \wedge h^2 \wedge (d_a-1)h^4 \mathbb{1}\left\{ h \geq \frac{1}{4(d_a-1)} \right\}$.

Remark 7 (Comments on T and dimension dependence). In rough terms, Theorem 1 guarantees that the expected regret converges to zero at least as quickly as $\frac{\log T}{T^{1/6}}$ 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 (11). This exponent can be further tuned, for example to $\frac{4}{3}$, to obtain a faster convergence rate to $\frac{\log T}{T^{1/4}}$. However, the optimal choice of the exponent deviates from our main goal, and we leave it to future work.

What is most important about our convergence guarantee is that the product $d_x d_a$ of the state and action dimensions *does not* appear in the bound: rather, any dimension factors are multiplied only by the rank r, which we expect to be far lower than the dimensions. Thus, to the best of our knowledge, our result stands as the first convergence result with non-trivial dimension scaling (i.e., $T \ll d_x d_a$) for doubly high-dimensional contextual bandits.

As we argued, our model is more general and expressive than many existing bandit models, making it intrinsically more complex, and our algorithm takes many practical concerns into consideration, both of which requires analysis from scratch. That being said, we provide a non-asymptotic, instance-dependent bound. Moreover, our algorithm does not require prior knowledge of T and r as mentioned in Remark 2, and our bound also holds consistently for all T and r, which we will further discuss in Remark 10. Establishing tighter bounds for our algorithm in various specific (well-studied) settings, i.e., special cases of our model, requires separate analyses, which deviates from our main goal. Nevertheless, our simulation shows that our method outperforms the state-of-the-art methods for these specific settings in Sections 5.1 and 5.2.

Remark 8 (Burn-in term). The first term in the bound (13) is a burn-in term, where the algorithm is gaining knowledge of $\mathbf{\Theta}^*$ from scratch. We do not impose any assumptions on these starting steps so that we have a relatively conservative burn-in term. Therefore, we only provide an upper bound on this burn-in term in Section EC.2.1. In practice, we can leverage historical data to obtain an initial estimate of $\mathbf{\Theta}^*$ so that the burn-in term can be much smaller.

The order of the burn-in term depends on the exponent—currently 3/2—used to specify the exploration frequency (11). Smaller exponents lead to more exploration, and hence a smaller burn-in term. As noted, it would be interesting to determine optimal choices of the exponent.

Remark 9 (Exploration-Exploitation Trade-off and Explanation of Terms). The third term can be considered as the cost of exploration steps. Each exploration round introduces a bias, regardless of estimation accuracy, primarily due to the projection onto the constraint set. Consequently, more frequent exploration increases the third term. The fourth term comes from estimation inaccuracies—more exploration reduces the fourth term. Therefore, there is an exploration-exploitation trade-off in terms of the exploration

frequency. Moreover, $(1 \wedge 2h\sqrt{d_a})$ in the third term and $v(d_a, h)$ in the fourth term are related to the variance of the perturbation (after projection). Their specific forms may change if we change the form of noise. However, the general message remains—larger variance typically leads to higher values in these two quantities. Therefore, there is also an exploitation-exploration trade-off in terms of the intensity of exploration in each exploration round. On a side note, the second term is a technical remaining term due to estimation inaccuracy, which always stays small.

Remark 10 (Adaptivity). Algorithm 1 is adaptive because it does not require knowing r or T a priori (except the ending point) as noted in Remark 2; moreover, the non-asymptotic bounds hold for all r and T. This adaptivity is of both theoretical interest and practical importance. Adaptivity regarding T overcomes the limitations of the traditional bandit framework, which possibly favors good performance at a specific T at the expense of other values. These limitations lead to algorithms involving T-dependent tuning parameters. In practice, it is preferable to have algorithms that do not require such tuning yet consistently perform well across all T. This important and desirable adaptivity property, unfortunately, often comes at the cost of the rate, as shown by Cai and Guo (2017).

Remark 11 (Choice of λ_0). The influence of λ_0 on the right-hand side is in the first term through B_{init} and the last term. B_{init} decreases as $\lambda_0 > 0$ increases from 0_+, but it no longer decreases once λ_0 is of the order $\sigma\sqrt{d_x \log(d_a + d_x)}$. The last term always increases with λ_0. Therefore, a sweet point for λ_0 is around the order $\sigma\sqrt{d_x \log(d_a + d_x)}$.

Remark 12 (Assumptions). To convey the main idea in a simple way, we have chosen to enforce relatively stringent assumptions. However, neither the normality assumptions of covariates and reward errors, nor the shape of the constraint set, is essential to the core structure of the proof.

Remark 13 (Technical novelty in analysis and possibility for further improvement). Due to the complexity of our model and our priority on practical algorithm design, we must confront the challenges posed by strong dependencies and non-classical distributions from scratch in our analysis. A core difficulty lies in analyzing the accuracy of $\widehat{\mathbf{\Theta}}_t$, which is based on observations that are highly dependent in intricate ways. However, most existing tools developed for high-dimensional matrix estimation rely critically on independence assumptions and are thus not directly applicable in our setting. To address these dependencies, we carefully construct martingales through conditional expectations and derive new concentration inequalities and bounds for the martingales. However, since martingales are less well-understood than independent sums, there remains room for tightening some intermediate results to improve the convergence rate. We leave such refinements for future research, as our primary focus is not on deriving optimal bounds.

Remark 14 (Lower Bounds). While some rough lower bounds for our model can be relatively easily derived, establishing informative lower bounds in our setting requires conceptual innovation and new tools.

As mentioned in Section 2.5, our model is general and encompasses the classical assortment, pricing, and bandit models, among others. Therefore, lower bounds for those models automatically apply to our model with straightforward modifications (e.g., $\sqrt{T}$ for linear bandit). However, establishing a more informative lower bound for our case is significantly more complicated. An informative lower bound for our doubly high-dimensional setting, where we cannot assume $T \gg \max\{d_a, d_x\}$, should account for both the dimensions (and the rank) and T, with greater emphasis on the former. Matrix completion literature typically accounts for optimality in terms of dimensions and rank (Negahban and Wainwright 2012, Cai and Zhou 2016), but they focus on i.i.d. observations. Bandit literature, on the other hand, addresses the dependency structure but not other complex structures in the data, such as high-dimensionality and the non-trivial dual complications in actions and contexts. The examples used by these two lines of literature for constructing lower bounds, however, are less compatible with each other. Apart from the difficulty of establishing meaningful lower bounds for the doubly high-dimensional contextual bandit model, it is important to recognize that many desirable algorithm properties, such as adaptivity, avoidance of excessive consecutive random explorations, and robustness, are not inherent to the model itself and thus cannot be reflected in the lower bound. Consequently, a comprehensive investigation of lower bounds would require a series of bounds under different assumptions, and is an important direction for future research.

4.2. Proof Sketch

At a high level, the proof of Theorem 1 consists of two major steps: Section 4.2.1 provides the high-probability bound on the estimation error of the low-rank matrix estimator $\widehat{\mathbf{\Theta}}_t$; Section 4.2.2 provides a non-asymptotic upper bound for the expected time-averaged regret $\mathcal{R}^\pi(T)$. Here we provide a sketch of each step, referring the reader to Appendix EC.2 for all the technical details.

4.2.1. Bounding the Estimation Error. An accurate estimate of the matrix $\mathbf{\Theta}^*$ is required to obtain good actions, so that our first step is to bound this estimation error. We introduce the shorthand $\mathbf{\Delta}_t := \widehat{\mathbf{\Theta}}_t - \mathbf{\Theta}^*$ for the error of the estimate $\widehat{\mathbf{\Theta}}_t$ at round t. Our first auxiliary result provides a high-probability bound on the Frobenius norm error $\|\mathbf{\Delta}_t\|_{\mathrm{F}}$. Before stating the proposition, we introduce an upper bound on the burn-in term B_{init}: $\overline{B_{\text{init}}} := C_L \left(d_x^{9/2} d_a^{9/2} (\log(d_a + d_x))^6 \left(\mathbb{1}\{h \geq \frac{1}{\sqrt{d_a - 1}}\} + \frac{\mathbb{1}\{h \geq \frac{1}{4(d_a-1)}\}}{d_a^{9/2} h^9} + \frac{\mathbb{1}\{h < \frac{1}{4(d_a-1)}\}}{d_a^{3/2} h^3} \right) + 1 \vee (\frac{\sigma\sqrt{d_x}}{\lambda_0})^6 (d_a^3 h^6 \wedge 1) \log^{12}\left(\frac{\sigma\sqrt{(d_a h^2 \wedge 1)}}{\lambda_0} + d_a + d_x \right) \right)$, where $C_L > 0$ is a constant depending on L only.

PROPOSITION 1. *There exist a function of L, d_x, d_a, h, σ, and λ_0, denoted as B_{init}, and $B_{init} \leq \overline{B_{init}}$. For any time $t \geq B_{init}$, we have*

$$\|\mathbf{\Delta}_t\|_F \leq C_f \frac{(\sqrt{d_x \log(tL)} + 2\log(tL))\sigma + \lambda_0\sqrt{r}}{t^{\frac{1}{6}} \left(\frac{1}{d_a - 1} \wedge h^2 \wedge (d_a - 1) h^4 \mathbb{1}\{h \geq \frac{1}{4(d_a-1)}\} \right)}, \tag{14a}$$

with probability at least

$$\phi(t) := 1 - \frac{5}{t} - \frac{2}{Lt} - \frac{2}{t^2} - \frac{2}{L^3 t^3}, \tag{14b}$$

where $C_f > 0$ is an absolute constant.

The technical challenge in establishing this result lies in the adaptive nature of the data: actions are chosen based on past data, and, in turn, affect future data, resulting in a highly non-i.i.d. dataset. For this reason, the summands in the empirical loss function are strongly dependent, so that known results for matrix completion, based on i.i.d. or weakly dependent data, are no longer applicable. Herein lies the need for careful analysis and technical innovation to handle the adaptive nature of bandit data collection.

The proof proceeds in three main steps. In the first step, we use the optimality conditions that define the estimator to derive a basic inequality, which we then re-arrange via a Taylor series into a more amenable form. In Steps 2 and 3, we carefully take conditional expectations to construct martingales, develop concentration results for (matrix-valued) martingales, and analyze non-standard distributions to derive high-probability upper bounds on different components of this inequality. We conclude the proof by combining these results.

1. First, since $\widehat{\boldsymbol{\Theta}}_t$ minimizes the function $\boldsymbol{\Theta} \mapsto \mathcal{L}_t(\boldsymbol{\Theta}) + \lambda_t |||\boldsymbol{\Theta}|||_{\text{nuc}}$, we have the basic inequality

$$\mathcal{L}_t(\widehat{\boldsymbol{\Theta}}_t) + \lambda_t |||\widehat{\boldsymbol{\Theta}}_t|||_{\text{nuc}} \leq \mathcal{L}_t(\boldsymbol{\Theta}^*) + \lambda_t |||\boldsymbol{\Theta}^*|||_{\text{nuc}},$$

   By performing a first-order Taylor series expansion of the loss function around $\boldsymbol{\Theta}^*$, this inequality implies that

$$e_t(\boldsymbol{\Delta}_t) \leq -\langle\langle \nabla \mathcal{L}_t(\boldsymbol{\Theta}^*),\, \boldsymbol{\Delta}_t \rangle\rangle + \lambda_t \left(|||\boldsymbol{\Theta}^*|||_{\text{nuc}} - |||\boldsymbol{\Theta}^* + \boldsymbol{\Delta}_t|||_{\text{nuc}} \right), \tag{15}$$

   where we have defined the Taylor series error function

$$e_t(\boldsymbol{\Delta}) := \mathcal{L}_t(\boldsymbol{\Theta}^* + \boldsymbol{\Delta}) - \mathcal{L}_t(\boldsymbol{\Theta}^*) - \langle\langle \nabla \mathcal{L}_t(\boldsymbol{\Theta}^*),\, \boldsymbol{\Delta} \rangle\rangle.$$

   The remainder of our analysis focuses on the three terms in Inequality (15). We establish a lower bound on the left-hand side term $e_t(\boldsymbol{\Delta}_t)$, and upper bounds on the two terms on the right-hand side.

2. Beginning with the left-hand side, we prove the following lower bound:

   LEMMA 1. *Under the assumptions of Theorem 1, for any $t \geq 4$, we have*

$$e_t(\boldsymbol{\Delta}) \geq \frac{\lfloor t^{2/3} \rfloor \tilde{c}_6 \left(\frac{1}{d_a - 1} \wedge h^2 \wedge (d_a - 1) h^4 \mathbb{1}\{h \geq \frac{1}{4(d_a-1)}\} \right)}{2t} |||\boldsymbol{\Delta}|||_F^2$$
$$- \frac{c}{t^{2/3}} (d_x + 3\log(tL) + d_x \log\log(tL)) \sqrt{d_a d_x \log(t)} \left(1 \wedge h \sqrt{d_a + 3\log(t) + \frac{d_a}{2} \log\log t} \right) |||\boldsymbol{\Delta}|||_F^2$$

   with probability at least $1 - \frac{1}{Lt} - \frac{2}{t} - \frac{1}{t^2}$, where $c > 0$ and $\tilde{c}_6 > 0$ are absolute constants.

   Note that the first term on the right-hand side scales as $t^{-1/3} |||\boldsymbol{\Delta}|||_{\text{F}}^2$, whereas the second term scales as $t^{-2/3} \log t |||\boldsymbol{\Delta}|||_{\text{F}}^2$. Therefore, we have established a lower bound on $e_t(\boldsymbol{\Delta})$ that scales as $t^{-1/3} |||\boldsymbol{\Delta}|||_{\text{F}}^2$ for large t, along with a pre-factor that depends on (h, d_x, d_a, L).

3. Our next lemma provides a high-probability bound on the quantity $\left| \langle\langle \nabla \mathcal{L}_t(\boldsymbol{\Theta}^*),\, \boldsymbol{\Delta} \rangle\rangle \right|$, which appears as the first term on the right-hand side of the Inequality (15). It involves the two pre-factors:

$$\phi_1(t) := \frac{2\sigma}{\sqrt{t}} \left(\sqrt{d_x \log(tL)} + 2\log(tL) \right) \quad \textit{and}$$
$$\phi_2(t) := \frac{\sigma}{t^{2/3}} \frac{12\sqrt{2}}{\sqrt{L}} \sqrt{\log(tL)(d_x + 3\log(Lt)) \Big(\big(d_a + 3\log(t)\big) h^2 \wedge 2 \Big) \big(\log(d_a + d_x) + 2\log t \big)}.$$

LEMMA 2. *Under the assumptions of Theorem 1, uniformly over all matrices* $\mathbf{\Delta} \in \mathbb{R}^{d_a \times d_x}$*, we have*

$$\left|\langle\langle \nabla \mathcal{L}_t(\mathbf{\Theta}^*),\, \mathbf{\Delta} \rangle\rangle\right| \leq \phi_1(t)\, |||\mathbf{\Delta}|||_F + \phi_2(t)\, |||\mathbf{\Delta}|||_{nuc}, \tag{16}$$

with probability at least $1 - \frac{2}{L^3 t^3} - \frac{1}{Lt} - \frac{3}{t} - \frac{1}{t^2}$.

By examining the prefactors $\phi_1(t)$ and $\phi_2(t)$ and considering their scalings in $(t, \mathbf{\Delta})$, we see that $|\langle\langle \nabla \mathcal{L}_t(\mathbf{\Theta}^*),\, \mathbf{\Delta} \rangle\rangle|$ is upper bounded by a quantity scaling $\frac{\log(t)}{\sqrt{t}} |||\mathbf{\Delta}|||_{\mathrm{F}}$ for sufficiently large t.

With these two lemmas in place, let us sketch out the remainder of the proof, deferring the full argument to Appendix EC.2.1. For a rank-r matrix $\mathbf{\Theta}^*$, a spectral decomposition argument can be used to show that

$$|||\mathbf{\Theta}^*|||_{\mathrm{nuc}} - |||\mathbf{\Theta}^* + \mathbf{\Delta}_t|||_{\mathrm{nuc}} \leq \sqrt{2r} |||\mathbf{\Delta}_t|||_{\mathrm{F}}. \tag{17}$$

We use this inequality to control the remaining term in the bound (15).

As noted in our discussion following Steps 2 and 3, for sufficiently large t, we have established the scaling relations $e_t(\mathbf{\Delta}_t) \gtrsim \frac{1}{t^{1/3}} |||\mathbf{\Delta}_t|||_{\mathrm{F}}^2$, and $|\langle\langle \nabla \mathcal{L}_t(\mathbf{\Theta}^*),\, \mathbf{\Delta} \rangle\rangle| \precsim \frac{\log(t)}{\sqrt{t}} |||\mathbf{\Delta}|||_{\mathrm{F}}$. Combining these scaling relations with the bound (17), our choice $\lambda_t \sim \frac{1}{\sqrt{t}}$, and substituting into the Inequality (15), we have

$$\frac{1}{t^{1/3}} |||\mathbf{\Delta}_t|||_{\mathrm{F}}^2 \precsim \frac{\log(t)}{\sqrt{t}} |||\mathbf{\Delta}_t|||_{\mathrm{F}} + \frac{1}{\sqrt{t}} \sqrt{2r} |||\mathbf{\Delta}_t|||_{\mathrm{F}}.$$

Consequently, we conclude that $|||\mathbf{\Delta}_t|||_{\mathrm{F}} \lesssim \frac{1}{t^{1/6}} \log(t)$ with high probability. Again, we refer the reader to Appendix EC.2.1 for all the technical details, including careful tracking of the lower order terms.

4.2.2. Bounding the Expected Regret. At each round t, we define the event

$$\mathcal{E}_t := \left\{ |||\mathbf{\Delta}_t|||_{\mathrm{F}} \leq C_f \frac{(\sqrt{d_x \log(tL)} + 2\log(tL))\sigma + \lambda_0 \sqrt{r}}{t^{\frac{1}{6}} \left(\frac{1}{d_a - 1} \wedge h^2 \wedge (d_a - 1) h^4 \mathbb{1}\{h \geq \frac{1}{4(d_a-1)}\} \right)} \right\}, \tag{18}$$

where C_f comes from Proposition 1, which guarantees that for large t, $\mathbb{P}(\mathcal{E}_t^c) \leq \frac{5}{t} + \frac{2}{Lt} + \frac{2}{t^2} + \frac{2}{L^3 t^3}$. By analyzing the expected regret on the events $\mathcal{E}_t$ and $\mathcal{E}_t^c$ separately, we show that both terms vanish with t at a polynomial rate.

5. Experimental Studies

This section is devoted to some experimental studies of the behavior of the proposed algorithm in different settings, both via controlled simulations and applications to two real-world datasets.

In Sections 5.1 and 5.2, we compare the performance of `Hi-CCAB` with other bandit and pricing algorithms. In all cases, we assume the reward error (12) follow a normal distribution with mean zero and variance σ^2,

$$\varepsilon_{t,\ell} \overset{i.i.d.}{\sim} N(0, \sigma^2) \quad \text{for } t = 1, \ldots, T \text{ and } \ell = 1, \ldots, L. \tag{19}$$

We then revisit the instant noodle joint assortment-pricing case study in Section 5.3 where we find that `Hi-CCAB` can boost cumulative sales by a factor larger than 4. Moreover, examination of the learned representation matrix $\widehat{\mathbf{\Theta}}$ provides insight into the latent factors of actions and covariates that influence revenue. Finally, in Section 5.4, we provide a real-world case study analysis of the assortment-pricing problem faced by a manicure start-up. Additional technical details are deferred to Appendix EC.3 of the supplementary material.

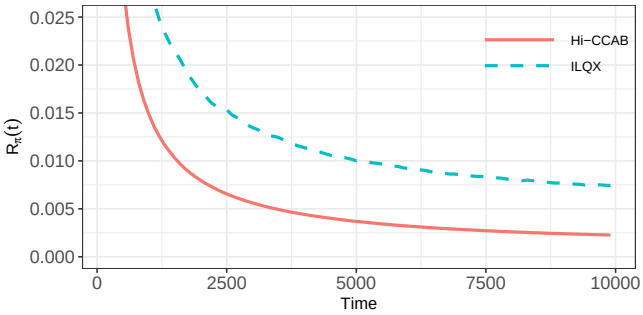


Figure 1 **Time-averaged regret for `Hi-CCAB` and the iterated Lasso-regularized quasi-likelihood regression (ILQX) proposed by Ban and Keskin (2021).**

5.1. Simulation Experiment I: Pricing Models

We follow the simulation set-up introduced by Ban and Keskin (2021), where demands and rewards are generated according to the demand model (7) and revenue model (8), with parameter vectors

$$\alpha := [1.1, -0.1, 0, 0.1, 0, 0.2, 0, 0.1, -0.1, 0, 0, 0.1, -0.1, 0.2, -0.2], \quad \text{and}$$
$$\beta := (-1)[0.5, 0.1, -0.1, 0, 0, 0, 0, 0.2, 0.1, 0.2, 0, 0.2, -0.1, -0.2, 0],$$

and the noise $\epsilon_t \overset{i.i.d.}{\sim} N(0, 10^{-4})$. As noted in Section 2.5.2, model (8) is a special case of our model (2).

We compare the time-averaged regret $\mathcal{R}^\pi(t)$ of `Hi-CCAB` with that of ILQX (iterated lasso-regularized quasi-likelihood regression with price experimentation) proposed by Ban and Keskin (2021). The basic idea of ILQX is to use LASSO to estimate the unknown α and β, and meanwhile to conduct price experiments for at least an order of $\sqrt{t}$ times.

Figure 1 compares the performance, measured by the time-averaged regret, of `Hi-CCAB` and ILQX. It is evident that `Hi-CCAB` converges faster than ILQX. As shown in Ban and Keskin (2021), ILQX converges faster than the greedy iterated least squares (Keskin and Zeevi 2014, Qiang and Bayati 2016), which decides the price based on the least square estimate of the unknown α and β at each iteration without experiments. We thus conclude that `Hi-CCAB` has better performance than various dynamic pricing algorithms, and is competitive in a continuum armed bandit problem.

5.2. Simulation Experiment II: Bandit Models

In this simulation study, we consider a multi-armed contextual bandit, which corresponds to a special case of our model with representation matrix $\boldsymbol{\Theta}^* = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_m)^\top$ as discussed in Section 2.5.1. In particular, each row of $\boldsymbol{\Theta}^*$ is the parameter vector of each arm for the multi-armed contextual bandit. We set the number of arms $d_a = \{10, 30, 50\}$, the dimension of covariates $d_x = 100$, and the consider both sparse and non-sparse $\boldsymbol{\Theta}^*$. For the non-sparse $\boldsymbol{\Theta}^*$, 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 diagonal entries $(1, 0.9, 0.9, 0.8, 0.5)$. The matrix $\boldsymbol{V}$ is generated by first drawing entries i.i.d. from $N(0, 1)$, and then orthonormalizing the columns via Gram-Schmidt. The matrix $\boldsymbol{U}$ is generated similarly, but further multiplied by $\sqrt{d_a}$ at the end so that the rewards are comparable across

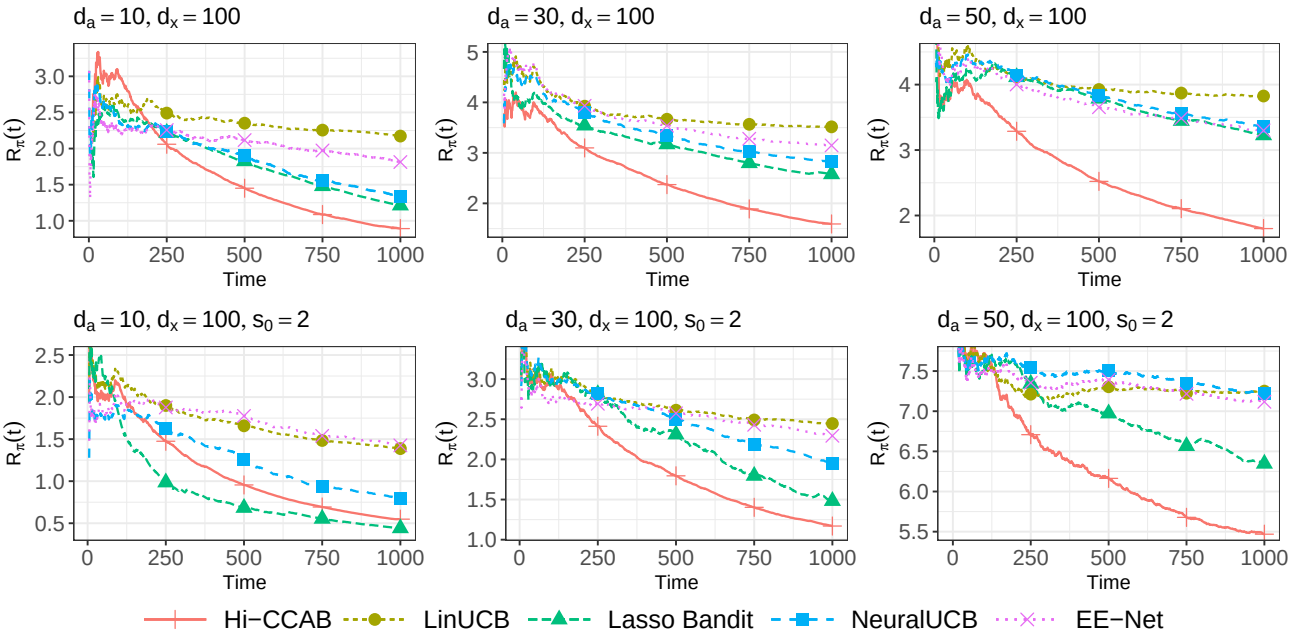


Figure 2 Time-averaged regret in a multi-armed contextual bandit setup for `Hi-CCAB` and other bandit algorithms. The upper row corresponds to a non-sparse case and the lower one corresponds to a sparse case.

different d_a's. For the sparse $\boldsymbol{\Theta}^*$, each of its rows—the coefficients for each arm—is set to zeros except for $s_0 = 2$ randomly selected elements i.i.d. 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 our model (3) with error variance (19) $\sigma^2 = 0.01$.

We compare `Hi-CCAB` against *a*) the LinUCB (Li et al. 2010), which is an extension of the traditional Upper Confidence Bound (UCB) algorithm to the contextual multi-armed bandit settings; *b*) a Lasso Bandit for high-dimensional contextual bandits (Bastani and Bayati 2020); *c*) NeuralUCB, a neural-network-based method for contextual bandits (Zhou et al. 2020) and *d*) EE-Net (Ban et al. 2022), which uses two separate neural networks for exploration and exploitation. Details of the tuning parameters of each algorithm are provided in Appendix EC.3.1.

Figure 2 shows the time-averaged regret $\mathcal{R}^\pi(T)$ averaged over 50 simulations. In the non-sparse settings (upper row), `Hi-CCAB` converges faster than all competing methods, with the performance gap widening as the arm dimension (i.e., number of arms) increases. This phenomenon highlights the benefit of leveraging low-rank structure, particularly in high-dimensional regimes. We expect that as the dimensions of the action and covariate spaces grow further, the advantage of `Hi-CCAB` will continue to increase. In the sparse settings (lower row), which are not to the advantage of `Hi-CCAB`, when the arm dimension is relatively small ($d_a = 10$), Lasso Bandit converges faster, though the margin over `Hi-CCAB` is modest. As the arm dimension increases, `Hi-CCAB` outperforms all competing methods. This phenomenon is surprising because, in this sparse setup, the $\boldsymbol{\Theta}^*$ is nearly full-rank. A possible explanation is that, as the arm dimension increases, the leading singular vectors capture a greater proportion of the variability, making $\boldsymbol{\Theta}^*$ approximately low/moderate-rank. In

particular, the top 20% singular vectors explain around 40% of the variance for $d_a = 10$, and increase to nearly 60% when $d_a = 50$; the top 50% singular vectors explain around 70% of the variance for $d_a = 10$, and increase to nearly 90% when $d_a = 50$. As mentioned earlier, our algorithm is adaptive to the rank of $\boldsymbol{\Theta}^*$, making it possible for the algorithm to work well in an approximately low-rank setting, which happens to echo the sparse setting in the high-dimensional regime. This phenomenon corroborates the adaptivity and robustness of our algorithm.

5.3. Case Study I: Instant Noodle Manufacturer

In this section, we revisit the instant noodle case study first introduced in Section 2.1, providing more detailed descriptions of the data, experiment setup, and results. Our algorithm provides joint assortment-pricing decisions that quadruple the cumulative sales revenue, and offers insightful interpretations of customer behavior through latent factors drawn from the estimated representation matrix $\widehat{\boldsymbol{\Theta}}$.

Data Description. The original data records daily sales across $L = 31$ provinces (covering 369 cities) in China over the time period from March 1st, 2021 to May 31st, 2022, for a total of $T = 456$ days. Throughout this period, new products were frequently introduced, gradually expanding the catalog, which eventually included 176 products (SKUs). Each product consisted of noodle packages of either a single flavor (13 possible options) or an assortment of flavors in varying quantities. Assortments and prices changed daily but were identical across all locations. The homepage could display at most $K = 30$ products. The manufacturer must decide which products to offer, from both the existing pool and potential new products, and set their prices accordingly.

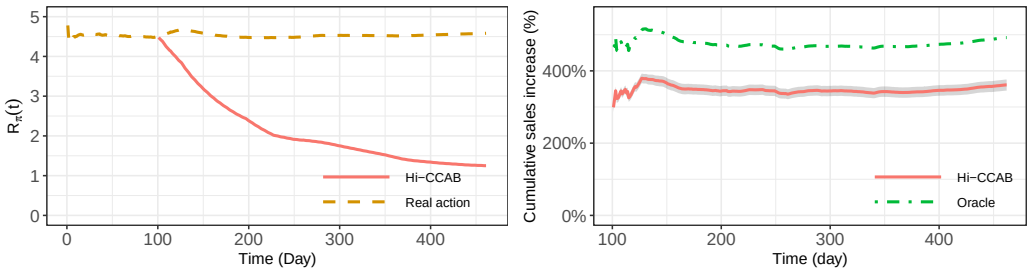
Experiment Setup and Results. To apply `Hi-CCAB`, we specify the action vectors $\boldsymbol{a}_t$ and covariate vectors $\{\boldsymbol{x}_{t,\ell}\}_{\ell=1}^{L}$ with $L = 31$ at given time t following the setup in Section 2.1. The action vector takes the form

$$\boldsymbol{a} = (\boldsymbol{f}_1, \boldsymbol{f}_1^2, p_1, p_1^2, \cdots, \boldsymbol{f}_K, \boldsymbol{f}_K^2, p_K, p_K^2, 1) \in \mathbb{R}^{2(m+1)K+1=841},$$

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, and $\boldsymbol{f}_k^2$ denotes the vector formed by squaring each component of $\boldsymbol{f}_k$. The context vector $\boldsymbol{x}_\ell \in \mathbb{R}^{50}$ includes the intercept, 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. See Appendix EC.3.2 for complete details.

In order to run simulations using the dataset, we first create a *pseudo-ground-truth* model by estimating $\boldsymbol{\Theta}^*$ and the reward error variance σ^2 using the full dataset. With slight abuse of notation, we still use $(\boldsymbol{\Theta}^*, \sigma^2)$ to denote the pseudo-ground-truth for simplicity.

To assess the validity of bilinear reward assumption (3), we evaluate the out-of-sample sales prediction accuracy. The prediction is done through a leave-one-out (LOO) approach: recursively over the index i, we compute an estimate $\widehat{\boldsymbol{\Theta}}$ of the true representation matrix without the i^{th} sample, and use this fit to predict the i^{th} reward. We then measure the performance of these LOO predictions relative to the actual rewards; doing so yields an out-of-sample prediction error rate of approximately 7%. See Appendix EC.3.2 for details.



(a) Time-averaged regret (in 1 million RMB). (b) Percentage cumulative sales gain.

Figure 3 Performance comparison between `Hi-CCAB` and real actions in terms of the time-averaged regret and the percentage gain in cumulative revenue. The boundaries of the shadow are the 5th and 95th quantiles.

Initializing `Hi-CCAB` with the initial step number $t_{init} = 100$ and λ_0 according to Algorithm 1, we then run 100 trials, in each of which we iterate from $t_{init} = 100$ to $T - 1 = 455$. At each iteration t, we first estimate $\widehat{\boldsymbol{\Theta}}_t$ and make an assortment-pricing decision $\boldsymbol{a}_{t+1}$ that maximizes the total sales given the covariate $\boldsymbol{x}_{t+1}$ according to Algorithm 1, and then generate a reward based on the pseudo-ground-truth model $(\boldsymbol{\Theta}^*, \sigma^2)$.

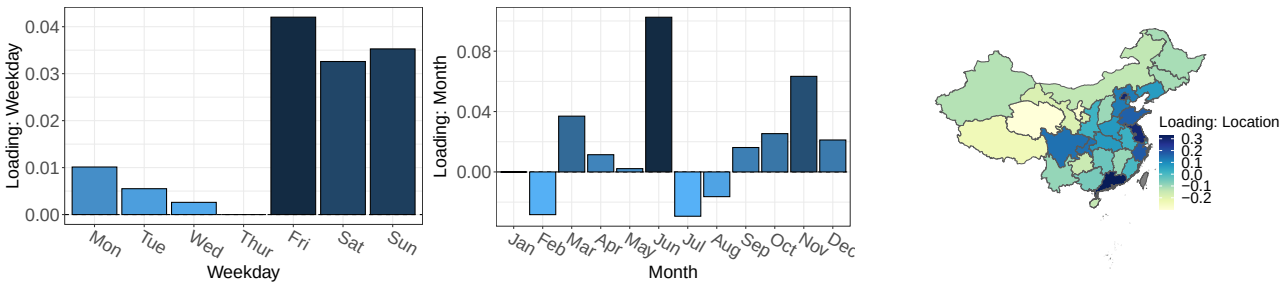
We evaluate `Hi-CCAB`'s performance in terms of the time-averaged regret (5) and the percentage gain in the cumulative sales revenue by comparing with the actions taken by the manufacturer, since no existing bandit algorithms are applicable to this joint assortment-pricing problem with contextual information.

Figure 3a shows the time-averaged regret and Figure 3b shows the percentage gain in cumulative sales compared to the real sales (averaged over 100 simulations). The expected time-averaged 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 to almost 4 times. The time-averaged sales by `Hi-CCAB` is also converging to that obtained by the oracle solution.

Interpretation from the Representation Matrix $\boldsymbol{\Theta}^*$. One advantage of our model is the interpretability that allows us to gain insights from the latent factors of the representation matrix $\boldsymbol{\Theta}^*$. Specifically, our model is able to discover the underlying factors of the effect of arm-covariate pairs on the reward. In the following, we examine the pseudo ground truth $\boldsymbol{\Theta}^*$ we obtained using all the data.

The rank of $\boldsymbol{\Theta}^*$ is 4 with the singular values being $(1.9, 0.2, 0.02, 0.00003)$. The leading singular value dominates the rest, making the corresponding singular vectors the most important in explaining the effect on the reward. We, therefore, focus on them in what follows.

Figure 4 shows the loadings for different covariates (i.e., the leading right singular vector). Our algorithm captures interpretable patterns in effects on the reward: within a week, the effects are drastically different between weekdays and weekends; across months, the effects show different patterns during the promotion months (June and November) from other months; across locations, the effects of the southeast-coastal provinces are different from the rest, which exactly corresponds to the regional economic development levels in China. In sum, our model can exploit the underlying structure of the covariates and provide insights into purchasing behavior and seasonality.



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

Figure 4 Loadings of the leading right singular vectors for the covariates.

	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	0.11	0.00	0.00	0.00	0.05	0.23	0.01	0.13	0.50	0.25	0.03	0.51
$\tilde{u}_1$ (quadratic)	1.00	0.07	0.00	0.00	0.00	0.03	0.17	0.01	0.09	0.41	0.19	0.02	0.42

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

On the other hand, Table 1 explores the (scaled) 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 $\boldsymbol{x}_\ell$ for $\ell = 1, \ldots, L$ on this day). 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 the months of May. 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.[3] As in Table 1, on May 29th 2022, flavor 1 (denoted $F1$) has the largest effect, followed by flavors 13, 10, 11, 7, and 9. Therefore, our model learns the values of the flavors (per unit).

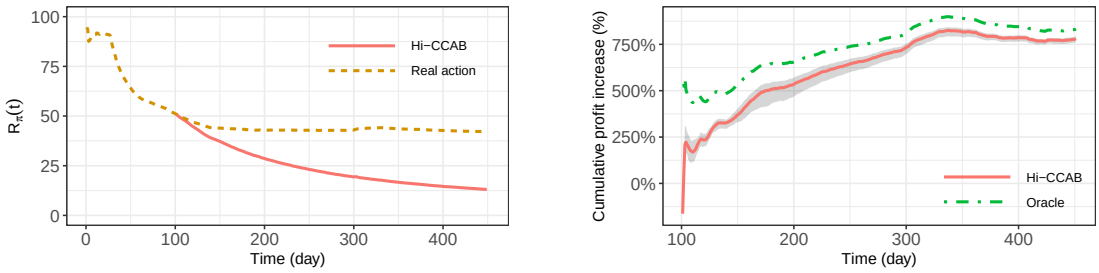
Limitations. We would like to point out that the increase in cumulative sales in real life could be less than four times as claimed because there will be additional constraints on the action space due to constraints on production or supply chain.

5.4. Case Study II: Manicure Start-up

In this section, we apply our method to a joint assortment-pricing problem faced by a manicure start-up. As a fast fashion brand, this start-up updates its product line quite frequently, and seeks to determine which colors and styles to prioritize in design, as well as how to optimize discounts and promotions. Accordingly, we encode the action differently by using the aggregated product features and discount rate, thereby demonstrating the flexibility of our model. As we show, our method not only boosts profit but also provides insightful interpretations.

Data Description. The start-up provided transaction-level data over the period from February 1st, 2020 to April 26th, 2021, for a total of $T = 451$ days. Over this period, the product line was updated on a regular

[3] The correlation of sales and the linear-term loadings is 0.95 and that of the quadratic-term loadings is 0.97.



(a) Time-averaged regret (in 1 thousand USD) (b) Percentage cumulative profit gain

Figure 5 Performance comparison between `Hi-CCAB` and real actions in terms of the time-averaged regret and the percentage gain in cumulative profit. The boundaries of the shadow are the 5^{th} and 95^{th} quantiles.

basis, with a maximum number of products available online being 74 during October 2020, and a total of 84 products (SKU) available at some point during the entire time horizon. Each product can be described by its texture (glossy versus matted), transparency, and colors (solid or multiple colors). The store also collaborates with designers and we measure their popularity by the number of their Instagram followers. The price of the products is fixed for designer vs non-designer and the cost of each product is known. Being a start-up company in a growth phase, they provide discount promotions on a regular basis to attract more customers. For each transaction, the data contains the purchased product, order total price, discount, shipping address, and an indicator of accepting marketing or not.

Experiment Setup and Results. As the start-up is primarily interested in tracking trends in colors and styles, it must also decide on the number of available products, designer items, and promotional offers on a daily basis. To capture these dynamics, we construct the action vector using the daily aggregated product information and the covariates vector based on the daily aggregated customer information as well as time indicators. Specifically, the action vector $\boldsymbol{a}$ includes the count of different colors used in all the manicures (black, white, gray, red, orange, yellow, green, blue, indigo, and violet), styles (the proportions of glossy, transparent and designer manicures, the total number of Instagram followers of the designers), the discount rate, as well as the quadratic terms of all the above, along with an additional one, leading to $d_a = 31$. The covariate vector $\boldsymbol{x}$ includes the intercept, location (percentages of purchase from Midwest, Northeast, South, and West), demographic proxies (average of median income and racial distribution by ZIP code), and the proportions of customers accepting marketing of last period, along with dummy variables for the 12 months, weekdays and public holidays. The dimension of the covariate vector is $d_x = 30$. Given that costs are known, we use profit rather than total sales revenue as our reward y.

Similar to our case study described in Section 5.3, we first create a pseudo-ground-truth model by estimating $\boldsymbol{\Theta}^*$ and σ using all data and check our model assumptions in Appendix EC.3.3. We then run the simulation 100 times with initialization $t_{init} = 100$.

Figure 5 shows the performance comparison between `Hi-CCAB` and actions taken by the start-up in terms of the time-averaged regret and percentage gain in cumulative profit (averaged over 100 simulations). The

findings are similar to those in Section 5.3: in particular, the time-averaged regret of `Hi-CCAB` converges to zero, whereas that of the real actions stays bounded away from zero. In addition, `Hi-CCAB` boosts the cumulative profit to more than 7 times.

Interpretation from the Representation Matrix. The rank of the pseudo-ground-truth representation matrix $\mathbf{\Theta}^*$ is 5, which is low compared to its dimensions 31 by 30. Its ordered singular values are $(0.66, 0.54, 0.22, 0.16, 0.0008)$. Since the last singular value is negligible compared to the first four, we focus our discussion on the singular vectors associated with the first four singular values.

Figures 6 and 7 illustrate (respectively) the loadings of the covariates and actions. At a high level, the four factors capture the Western market $(\boldsymbol{u}_1, \boldsymbol{v}_1)$, Northeastern/Southern market $(\boldsymbol{u}_2, \boldsymbol{v}_2)$, income effect $(\boldsymbol{u}_3, \boldsymbol{v}_3)$, and time effect $(\boldsymbol{u}_4, \boldsymbol{v}_4)$ respectively. The covariate loadings associated with the factors are well-separated, which facilitates interpretations for the actions. Let us first look at color and style. For color, we examine the quadratic terms since the color action vector represents the absolute count of each color's appearance in the manicures so that the quadratic terms dominate. Comparing the Western market ($\boldsymbol{u}_1$ and $\boldsymbol{v}_1$) and Northeastern/Southern market ($\boldsymbol{u}_2$ and $\boldsymbol{v}_2$), we observe distinct geographical preferences in color choices: blue is more popular in the Western market while black is preferred in the Northeastern/Southern market; gray does not sell well in the west while white is lackluster in the Northeast and South. As for the income factor $(\boldsymbol{u}_3, \boldsymbol{v}_3)$, the color preferences are reflected in the linear terms (since the quadratic terms are negligible): white and indigo are more sought-after among higher-income customers while gray, orange, and green are less favorable. For the time factor $(\boldsymbol{u}_4, \boldsymbol{v}_4)$, red stands out as a festive favorite, aligning with holiday trends. In terms of style, their loadings concentrate in the income factor $(\boldsymbol{u}_3, \boldsymbol{v}_3)$. Customers with higher incomes show less interest in designer and glossy products and care less about designers' popularity.

Next, we investigate the effects of discounts. Loadings of discount rate are concentrated in the vector $\boldsymbol{u}_3$, associated with the income factor tracked by the pair $(\boldsymbol{u}_3, \boldsymbol{v}_3)$. Generally, for high-income customers, higher discount rates yield lower profits. One plausible explanation is: the demand function (7) is controlled by the market size (via the linear term) and the price sensitivity (via the quadratic term), both of which depend on income. In our case, the market size in the linear term dominates due to relatively small discount rates (median: 0.067; 5th and 95th quantiles: 0.03 and 0.18) and the fact that the linear term in $\boldsymbol{u}_3$ outweighs the quadratic term. Therefore, the negative linear term suggests that profits will be lower with higher discount rates for our customer base whose household income is of mid-to-high levels (ranging from \$95K to \$110K).

As income increases, so does the market size, particularly for hedonic purchases such as manicures among our customer base. Consequently, offering discounts to higher-income customers may lead to greater profit loss. On the other hand, as a start-up, customer expansion and retention are vital for long-term growth, and discounts can serve as effective incentives in this regard. This unique dynamics of a start-up, however, can only be revealed by a longer sequence of data. That being said, this longer-term aspect of decision-making, while beyond the scope of our current case study, is an important direction for future research.

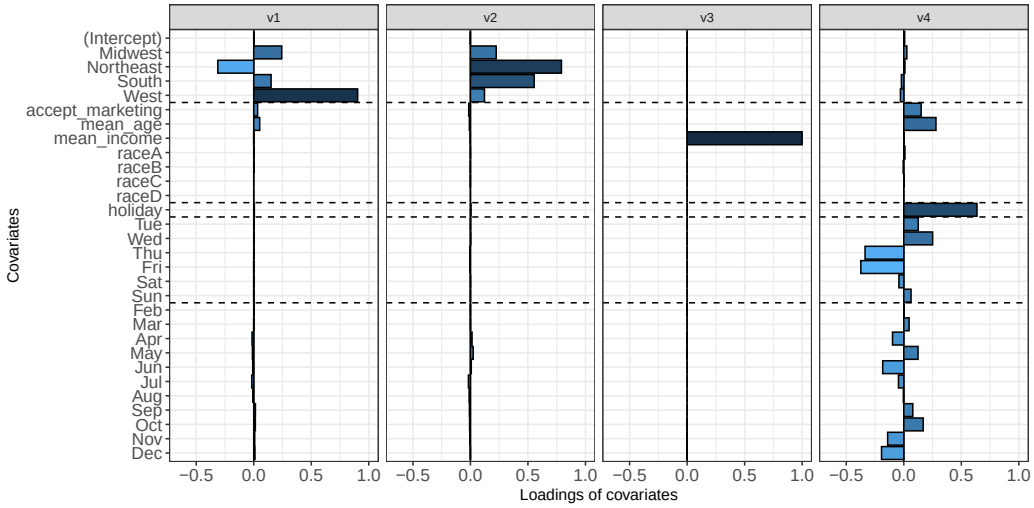


Figure 6 **Loadings associated with context vectors, as illustrated by stem plots of the singular vectors v_1, v_2, v_3 and v_4. Dashed lines separate the covarites into locations, demographic proxies, a holiday indicator, weekdays, and months.**

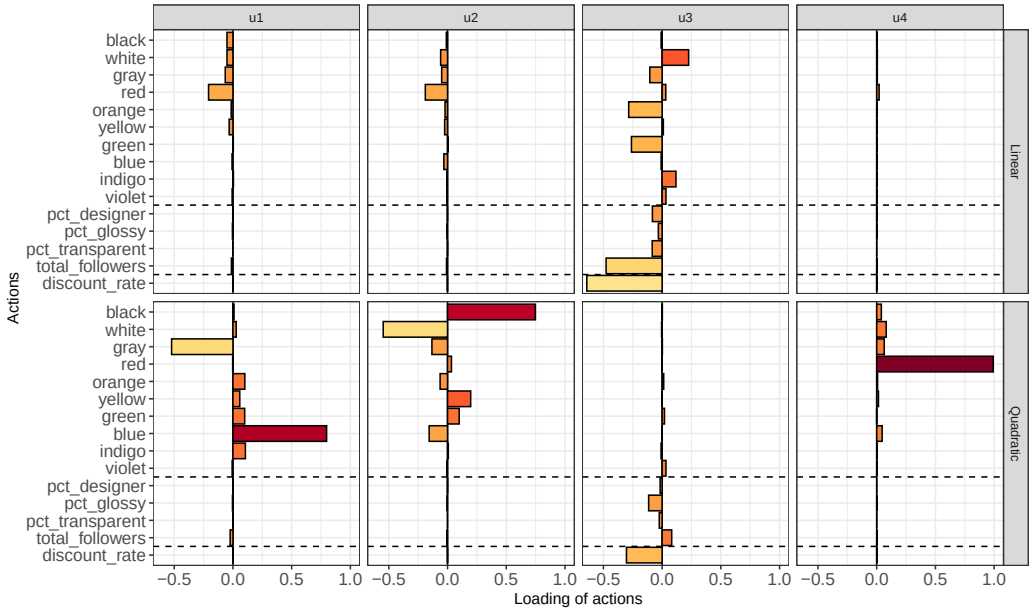


Figure 7 **Loadings associated with the action vectors, as illustrated by stem plots of the left singular vectors u_1, u_2, u_3, and u_4. Dashed lines separate the actions into colors, styles, and discount rate.**

6. Conclusions and Future Directions

The growing need for online sequential data-driven decision-making has led to increased interest in bandit models among both theorists and practitioners. Nonetheless, to date, we are not aware of any work on contextual bandits in which both the covariate and action spaces are high-dimensional. Our work is motivated by real-world problems that admit a bandit formulation—among them the joint assortment-pricing problem—that have this "doubly" high-dimensional nature. We proposed a bilinear model with low-rank matrix

to capture interaction effects of action-covariate pairs in determining the reward function. This model is reasonably general, including a number of structured bandit and pricing models as special cases; and it is also highly interpretable via the spectral structure of the representation matrix.

We proposed an efficient algorithm, `Hi-CCAB`, that iteratively interleaves low-rank matrix estimation with exploration/exploitation, and we established a non-asymptotic upper bound on its time-averaged regret. Our method is capable of addressing the *joint* assortment-pricing problem, each of which has been studied extensively in operations research and marketing, but not jointly. In real case studies with the largest instant noodle manufacturer and a manicure start-up, our method not only boost sales/profits but also provides insights into how actions and contexts influence the sales revenue (e.g., revealing purchasing behaviors).

We conclude by discussing some future research directions. The first is on the theoretical side. While we established a non-asymptotic upper bound on the regret for our new algorithm targeting our novel doubly high-dimensional contextual bandit model, the tight analysis of regret is left to the future work, which is feasible and important, given the oustanding performance of `Hi-CCAB` compared to other bandit algorithms. A matching lower bound is also along this direction. Another direction is on generality of the model in terms of mathematical expressiveness, i.e., when the covariate and action vectors are specified as particular feature maps, such as the generalized bilinear models, MNL models, or other formulations as explained in Appendix EC.1, with a particular focus on theory.

Application wise, the generality and flexibility of model allow for applications to other multiple decision-making problems in diverse sectors, such as other business settings and healthcare. In the realm of business, our model can incorporate other quantifiable actions for joint decision-making, particularly in marketing and operations. Furthermore, it offers flexibility in the objective, allowing it to be tailored to suit different outcomes, such as social benefits for social enterprises and the Environmental, Social, and Governance (ESG) performance for responsible investment. In healthcare, especially personalized healthcare, our model holds high potential. For instance, health monitoring systems that recommend actions based on individual health conditions align naturally with our framework. Actions like suggestions regarding sleeping schedules, exercise regimens, social media usage, and dietary choices can be represented as high-dimensional and continuous action vectors. Meanwhile, health outcome also highly depends on contextual variables such as age, gender, weight, height, basic health status, and compliance level. Traditional bandit models do not suffice for such doubly high-dimensional contextual settings, but our bilinear model with a low-rank $\boldsymbol{\Theta}^*$ for mean reward fits well, as health effects of actions and user characteristics can be typically captured by a few latent factors.

Finally, our case studies partially relied on simulations to evaluate the efficacy of our method. However, real operational settings often impose additional constraints on the action space. In order to gauge and improve the real-world performance of our methods, we anticipate further collaboration with companies in carrying out live deployments.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24.

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

Abdallah, T. and Vulcano, G. (2021). Demand estimation under the multinomial logit model from sales transaction data. *Manufacturing & Service Operations Management*, 23(5):1196–1216.

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

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

Akçay, Y., Natarajan, H. P., and Xu, S. H. (2010). Joint dynamic pricing of multiple perishable products under consumer choice. *Management Science*, 56(8):1345–1361.

Aparicio, D., Eckles, D., and Kumar, M. (2023). Algorithmic pricing and consumer sensitivity to price variability. *Available at SSRN 4435831*.

Araman, V. F. and Caldentey, R. (2009). Dynamic pricing for nonperishable products with demand learning. *Operations Research*, 57(5):1169–1188.

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

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

Ban, Y., Yan, Y., Banerjee, A., and He, J. (2022). EE-net: Exploitation-exploration neural networks in contextual bandits. In *The Tenth International Conference on Learning Representations*.

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

Bastani, H., Simchi-Levi, D., and Zhu, R. (2022). Meta dynamic pricing: Transfer learning across experiments. *Management Science*, 68(3):1865–1881.

Belloni, A., Freund, R., Selove, M., and Simester, D. (2008). Optimizing product line designs: Efficient methods and comparisons. *Management Science*, 54(9):1544–1552.

Bennett, J., Lanning, S., et al. (2007). The Netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35. New York.

Bernstein, F., Modaresi, S., and Sauré, D. (2019). A dynamic clustering approach to data-driven assortment personalization. *Management Science*, 65(5):2095–2115.

Bertsimas, D. and Mišić, V. V. (2019). Exact first-choice product line optimization. *Operations Research*, 67(3):651–670.

Besbes, O. and Zeevi, A. (2009). Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57(6):1407–1420.

Besson, L. and Kaufmann, E. (2018). What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*.

Broder, J. and Rusmevichientong, P. (2012). Dynamic pricing under a general parametric choice model. *Operations Research*, 60(4):965–980.

Cai, T. T. and Guo, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, 45(2):615 – 646.

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

Cai, T. T. and Zhou, W.-X. (2016). Matrix completion via max-norm constrained optimization.

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

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

Cavallo, A. (2018). More amazon effects: online competition and pricing behaviors. Technical report, National Bureau of Economic Research.

Chen, K. D. and Hausman, W. H. (2000). Mathematical properties of the optimal product line selection problem using choice-based conjoint analysis. *Management Science*, 46(2):327–332.

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

Chen, R. (2022). *Estimation and Inference for Convex Functions and Computational Efficiency in High Dimensional Statistics*. PhD thesis, University of Pennsylvania.

Chen, X., Krishnamurthy, A., and Wang, Y. (2023). Robust dynamic assortment optimization in the presence of outlier customers. *Operations Research*.

Chen, X., Owen, Z., Pixton, C., and Simchi-Levi, D. (2022a). A statistical learning approach to personalization in revenue management. *Management Science*, 68(3):1923–1937.

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

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

Chen, X., Wang, Y., and Zhou, Y. (2020). Dynamic assortment optimization with changing contextual information. *The Journal of Machine Learning Research*, 21(1):8918–8961.

Chen, Y., Wang, Y., Fang, E. X., Wang, Z., and Li, R. (2022b). Nearly dimension-independent sparse linear bandit over small action spaces via best subset selection. *Journal of the American Statistical Association*, pages 1–13.

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

Cheung, W. C. and Simchi-Levi, D. (2017). Thompson sampling for online personalized assortment optimization problems with multinomial logit choice models. *Available at SSRN 3075658*.

Cohen, M. C., Lobel, I., and Paes Leme, R. (2020). Feature-based dynamic pricing. *Management Science*, 66(11):4921–4943.

Dani, V., Hayes, T. P., and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*.

Den Boer, A. V. (2015). Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in Operations Research and Management Science*, 20(1):1–18.

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

Estelami, H., Lehmann, D. R., and Holden, A. C. (2001). Macro-economic determinants of consumer price knowledge: A meta-analysis of four decades of research. *International Journal of Research in Marketing*, 18(4):341–355.

Fan, J., Guo, Y., and Yu, M. (2022). Policy optimization using semiparametric models for dynamic pricing. *Journal of the American Statistical Association*, pages 1–29.

Fan, J., Li, K., and Liao, Y. (2021). Recent developments in factor models and applications in econometric learning. *Annual Review of Financial Economics*, 13:401–430.

Féraud, R., Allesiardo, R., Urvoy, T., and Clérot, F. (2016). Random forest for the contextual bandit problem. In *Artificial Intelligence and Statistics*, pages 93–101. PMLR.

Ferreira, K. J. and Mower, E. (2023). Demand learning and pricing for varying assortments. *Manufacturing & Service Operations Management*, 25(4):1227–1244.

Foupouagnigni, M. and Mouafo Wouodjié, M. (2020). On multivariate Bernstein polynomials. *Mathematics*, 8(9):1397.

Gallego, G. and Wang, R. (2014). Multiproduct price optimization and competition under the nested logit model with product-differentiated price sensitivities. *Operations Research*, 62(2):450–461.

Gordon, B. R., Goldfarb, A., and Li, Y. (2013). Does price elasticity vary with economic growth? a cross-category analysis. *Journal of Marketing Research*, 50(1):4–23.

Green, P. E. and Krieger, A. M. (1985). Models and heuristics for product line selection. *Marketing Science*, 4(1):1–19.

Green, P. E. and Krieger, A. M. (1993). Conjoint analysis with product-positioning applications. *Handbooks in operations research and management science*, 5:467–515.

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

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417.

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

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

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

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

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

Kang, Y., Hsieh, C.-J., and Lee, T. C. M. (2022). Efficient frameworks for generalized low-rank matrix bandit problems. In *Advances in Neural Information Processing Systems*, volume 35, pages 19971–19983.

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

Kim, G.-S. and Paik, M. C. (2019). Doubly-robust LASSO bandit. In *Advances in Neural Information Processing Systems*, volume 32.

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

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

Kleinberg, R. and Leighton, T. (2003). The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 594–605. IEEE.

Kleinberg, R., Slivkins, A., and Upfal, E. (2019). Bandits and experts in metric spaces. *Journal of the ACM*, 66(4).

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

Kumar, V., Umashankar, N., Kim, K. H., and Bhagwat, Y. (2014). Assessing the influence of economic and customer experience factors on service purchase behaviors. *Marketing Science*, 33(5):673–692.

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

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

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

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338.

Li, H., Webster, S., and Yu, G. (2020). Product design under multinomial logit choices: Optimization of quality and prices in an evolving product line. *Manufacturing & Service Operations Management*, 22(5):1011–1025.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670.

Li, Q., Cheng, G., Fan, J., and Wang, Y. (2018). Embracing the blessing of dimensionality in factor models. *Journal of the American Statistical Association*, 113(521):380–389.

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

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

Madden, T. J., Hewett, K., and Roth, M. S. (2000). Managing images in different cultures: A cross-national study of color meanings and preferences. *Journal of International Marketing*, 8(4):90–107.

McBride, R. D. and Zufryden, F. S. (1988). An integer programming approach to the optimal product line selection problem. *Marketing Science*, 7(2):126–140.

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

Miao, S. and Chao, X. (2022). Online personalized assortment optimization with high-dimensional customer contextual data. *Manufacturing & Service Operations Management*, 24(5):2741–2760.

Miao, S., Chen, X., Chao, X., Liu, J., and Zhang, Y. (2022). Context-based dynamic pricing with online clustering. *Production and Operations Management*, 31(9):3559–3575.

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

Negahban, S. and Wainwright, M. J. (2012). Restricted strong convexity and (weighted) matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13:1665–1697.

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

Pol, L. G. (1991). Demographic contributions to marketing: an assessment. *Journal of the Academy of Marketing Science*, 19:53–59.

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

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

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

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

Rohe, K. and Zeng, M. (2023). Vintage Factor Analysis with Varimax Performs Statistical Inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkad029.

Rusmevichientong, P., Shen, Z.-J. M., and Shmoys, D. B. (2010). Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations Research*, 58(6):1666–1680.

Rusmevichientong, P. and Tsitsiklis, J. N. (2010). Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411.

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

Schuler, A., Liu, V., Wan, J., Callahan, A., Udell, M., Stark, D. E., and Shah, N. H. (2016). Discovering patient phenotypes using generalized low rank models. In *Biocomputing 2016: Proceedings of the Pacific Symposium*, pages 144–155. World Scientific.

Shen, S., Chen, X., Fang, E., and Lu, J. (2023). Combinatorial inference on the optimal assortment in multinomial logit models. *Available at SSRN 4371919*.

Singh, S. (2006). Impact of color on marketing. *Management Decision*, 44(6):783–789.

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

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

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

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

Udell, M., Horn, C., Zadeh, R., Boyd, S., et al. (2016). Generalized low rank models. *Foundations and Trends® in Machine Learning*, 9(1):1–118.

Vulcano, G., Van Ryzin, G., and Ratliff, R. (2012). Estimating primary demand for substitutable products from sales transaction data. *Operations Research*, 60(2):313–334.

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

Wang, R. (2021). Consumer choice and market expansion: Modeling, optimization, and estimation. *Operations Research*, 69(4):1044–1056.

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

Xu, P., Wen, Z., Zhao, H., and Gu, Q. (2022). Neural contextual bandits with deep representation and shallow exploration. In *The Tenth International Conference on Learning Representations*.

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

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

Supplementary material

EC.1. Nonlinear Relationship Representations and Extensions to Reproducing Kernel Hilbert Spaces

In this section, we describe how our bilinear form can capture complicated non-linear effects in the actions and covariates. Suppose that the original (primitive) action vector is given by $\boldsymbol{a}_{base} = (a_1, a_2, \ldots, a_{d_1}) \in \mathbb{R}^{d_1}$. We then define an augmented action vector by including all polynomial terms of $\boldsymbol{a}_{base}$ up to order N_1—that is

$$\begin{aligned}\boldsymbol{a} := (1, a_1, a_2, \ldots, a_{d_1}, a_1^2, a_2^2, \ldots, a_{d_1}^2, a_1a_2, a_1a_3, \ldots, a_1a_{d_1}, a_2a_3, \ldots, a_2a_{d_1}, \ldots, a_{d_1-1}a_{d_1}, \ldots \\ \ldots, a_1^{N_1}, \cdots, a_{d_1}^{N_1}, a_1^{N_1-1}a_2, a_1^{N_1-1}a_3, \ldots, a_1^{N_1-1}a_{d_1}, \ldots, a_{d_1}^{N_1-1}a_{d_1-1}, \ldots, a_{d_1-N_1+1}a_{d_1-N_1+2}\cdots a_{d_1}). \quad \text{(EC.1)}\end{aligned}$$

Similarly, let $\boldsymbol{x}_{base} \in \mathbb{R}^{d_2}$ denote the original (primitive) context vector. We then define an extended context vector $\boldsymbol{x}$ by including all the polynomial terms of $\boldsymbol{x}_{base}$ up to order N_2.

With these definitions, the expected reward takes the form

$$\boldsymbol{a}^\top \boldsymbol{\Theta} \boldsymbol{x} = \sum_{\substack{k_i, \ell_j \geq 0 \\ k_1+k_2+\ldots+k_{d_1} \leq N_1, \\ \ell_1+\ell_2+\ldots+\ell_{d_2} \leq N_2}} h(k_1, k_2, \ldots, k_{N_1}, \ldots, \ell_1, \ell_2, \ldots, \ell_{N_2}) \cdot a_1^{k_1} a_2^{k_2} \cdots a_{d_1}^{k_{d_1}} \cdot x_1^{\ell_1} x_2^{\ell_2} \cdots x_{d_2}^{\ell_{d_2}}, \quad \text{(EC.2)}$$

where $h(k_1, k_2, \ldots, k_{N_1}, \ldots, \ell_1, \ell_2, \ldots, \ell_{N_2})$ is an element in $\boldsymbol{\Theta}$.

It is evident that the function $g(\boldsymbol{a}_{base}, \boldsymbol{x}_{base}) = \boldsymbol{a}^\top \boldsymbol{\Theta} \boldsymbol{x}$ as in Equation (EC.2) can capture nonlinear relationships in terms of $(\boldsymbol{a}_{base}, \boldsymbol{x}_{base})$. Furthermore, by suitably choosing the orders (N_1, N_2) of lifting, such a function g can be used to approximate any continuous functions $(\boldsymbol{a}_{base}, \boldsymbol{x}_{base}) \mapsto f(\boldsymbol{a}_{base}, \boldsymbol{x}_{base})$ to arbitrary accuracy on any compact set $\mathcal{C}$. (E.g., see the paper Foupouagnigni and Mouafo Wouodjié (2020) for results on the approximation (multivariate) functions with Bernstein polynomials). Note that besides polynomial type basis (EC.1) that yields the form (EC.2), we can also use other bases such as Fourier or Haar.

Sometimes we may want to use the basis in an RKHS defined with kernel $\mathcal{K}_{\boldsymbol{x}_{base}}(\cdot, \cdot)$: $\phi_1(\cdot), \phi_2(\cdot), \phi_3(\cdot), \ldots$, for covariates, and the basis associated with kernel $\mathcal{K}_{\boldsymbol{a}_{base}}(\cdot, \cdot)$: $\psi_1(\cdot), \psi_2(\cdot), \psi_3(\cdot), \ldots$, for actions. The reasons for this choice are numerous, including the vanishing contribution to the reward of large primitive covariates (or primitive action), the need for transforming the domain of the primitive covariates (or primitive action), etc. One difficulty in using the basis in RKHS is that we usually are not able to write out the eigenfunctions by the order of eigenvalues (see Chapter 12 in Wainwright (2019) for basic properties of RKHS). But this difficulty can be dealt by mapping $\boldsymbol{a}_{base}$ to a lifted vector of the form

$$\boldsymbol{x}_{base} \mapsto \boldsymbol{x} := (\mathcal{K}_{\boldsymbol{x}_{base}}(\mathbb{x}_1, \boldsymbol{x}_{base}), \mathcal{K}_{\boldsymbol{x}_{base}}(\mathbb{x}_2, \boldsymbol{x}_{base}), \mathcal{K}_{\boldsymbol{x}_{base}}(\mathbb{x}_3, \boldsymbol{x}_{base}), \cdots, \mathcal{K}_{\boldsymbol{x}_{base}}(\mathbb{x}_{N_x}, \boldsymbol{x}_{base})), \quad \text{(EC.3)}$$

where $\mathbb{x}_1, \mathbb{x}_2, \ldots, \mathbb{x}_{N_x}$ are "characteristic indices" satisfying 1. $\mathcal{K}(\mathbb{x}_i, \mathbb{x}_j) = 0$ for all $i \neq j$, and 2. $P_{\boldsymbol{x}_{base}}(\mathcal{K}(\mathbb{x}_i, \boldsymbol{x}_{base}) \neq 0) > 0$ for all i. In this way, we include the basis that form the relevant space and the matrix $\boldsymbol{\Theta}$ can learn the right structure.

Lifting the space of $\boldsymbol{a}$ and $\boldsymbol{x}$, either by explicitly expanding the vector through basis or by using RKHS, significantly increases the expressiveness and flexibility of our model. Despite the increasing complexity in space, the learning task does not necessarily become more complex thanks to our low-rank assumptions. Efficient algorithms designed for low-rank structures can easily handle the increasing space complexity.

EC.2. Technical Details for Proofs

In this section, we provide the full technical details of the proofs that were deferred from the main body. We begin with details for the proof of Proposition 1, which bounds the error in estimating the representation matrix $\boldsymbol{\Theta}^*$. We then turn to the proof of Theorem 1. The two lemmas involved in the proof of Proposition 1 stated in the main text are proved in Section EC.2.3 and Section EC.2.4.

EC.2.1. Proof of Proposition 1

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

Introducing the shorthand notation $\boldsymbol{\Delta}_{t,\perp} := \boldsymbol{U}_\perp \boldsymbol{U}_\perp^\top \boldsymbol{\Delta}_t \boldsymbol{V}_\perp \boldsymbol{V}_\perp^\top$, this construction ensures that $|||\boldsymbol{\Theta}^* + \boldsymbol{\Delta}_{t,\perp}|||_{\text{nuc}} = |||\boldsymbol{\Theta}^*|||_{\text{nuc}} + |||\boldsymbol{\Delta}_{t,\perp}|||_{\text{nuc}}$. This fact, along with some applications of the triangle inequality, yields the lower bound

$$\begin{aligned}
|||\boldsymbol{\Theta}^* + \boldsymbol{\Delta}_t|||_{\text{nuc}} &\geq |||\boldsymbol{\Theta}^* + \boldsymbol{\Delta}_{t,\perp}|||_{\text{nuc}} - |||\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_{t,\perp}|||_{\text{nuc}} \\
&= |||\boldsymbol{\Theta}^*|||_{\text{nuc}} + |||\boldsymbol{\Delta}_{t,\perp}|||_{\text{nuc}} - |||\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_{t,\perp}|||_{\text{nuc}} \\
&\geq |||\boldsymbol{\Theta}^*|||_{\text{nuc}} + |||\boldsymbol{\Delta}_{t,\perp}|||_{\text{nuc}} - \sqrt{2r}|||\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_{t,\perp}|||_{\text{F}},
\end{aligned}$$

where the final inequality follows from the fact that the matrix $\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_{t,\perp}$ has rank at most $2r$.

Combining this inequality with Inequality (15) yields the upper bound

$$e_t(\boldsymbol{\Delta}_t) \leq \left|\langle\langle \nabla \mathcal{L}_t(\boldsymbol{\Theta}^*),\, \boldsymbol{\Delta}_t \rangle\rangle\right| + \lambda_t \left(\sqrt{2r}|||\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_{t,\perp}|||_{\text{F}} - |||\boldsymbol{\Delta}_{t,\perp}|||_{\text{nuc}} \right). \tag{EC.4}$$

Next, we apply the results of Lemma 1 and Lemma 2 to Equation (EC.4). Doing so guarantees that, with probability at least $1 - \frac{5}{t} - \frac{2}{Lt} - \frac{2}{t^2} - \frac{2}{L^3 t^3}$, we have

$$\begin{aligned}
&\left\{ \frac{\lfloor t^{2/3} \rfloor}{2t} \tilde{c}_6 \Big(\frac{1}{d_a - 1} \wedge h^2 \wedge (d_a - 1) h^4 \mathbb{1}\{h \geq \frac{1}{4(d_a - 1)}\} \Big) \right. \\
&\left. - \frac{c}{t^{2/3}} (d_x + 3\log(tL) + d_x \log\log(tL)) \sqrt{d_a d_x \log(t)} \left(1 \wedge h \sqrt{d_a + 3\log(t) + \frac{d_a}{2}\log\log t} \right) \right\} |||\boldsymbol{\Delta}_t|||_{\text{F}}^2 \\
&\leq |||\boldsymbol{\Delta}_t|||_{\text{F}} \frac{2\sigma}{\sqrt{t}} \big(\sqrt{d_x \log(tL)} + 2\log(tL) \big) + \\
&\frac{\sigma}{t^{2/3}} \frac{12\sqrt{2}}{\sqrt{L}} \sqrt{\log(tL)(d_x + 3\log(Lt)) \Big(\big(d_a + 3\log(t)\big) h^2 \wedge 2 \Big) \big(\log(d_a + d_x) + 2\log t \big)} \left(|||\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_{t,\perp}|||_{\text{nuc}} + |||\boldsymbol{\Delta}_{t,\perp}|||_{\text{nuc}} \right) \\
&+ \lambda_0 \frac{\sqrt{t}}{t} \sqrt{2r} |||\boldsymbol{\Delta}_t|||_{\text{F}} - \lambda_0 \frac{\sqrt{t}}{t} |||\boldsymbol{\Delta}_{t,\perp}|||_{\text{nuc}}.
\end{aligned} \tag{EC.5}$$

Next, we update Inequality (EC.5) by applying the bound $\|\!|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_{t,\perp}|\!\|_{\text{nuc}} \le \sqrt{2r}\|\!|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_{t,\perp}|\!\|_{\text{F}}$, dividing both sides by $\|\!|\boldsymbol{\Delta}_t|\!\|_{\text{F}}$, and multiplying both sides with $3t^{\frac{1}{3}}/\Big(\tilde{c}_6(\frac{1}{d_a-1} \wedge h^2 \wedge (d_a-1)h^4\mathbb{1}\{h \ge \frac{1}{4(d_a-1)}\})\Big)$. Suppose B_{init} is the smallest integer such that for all $t \ge B_{\text{init}}$, the following inequalities hold

$$t \ge 8, \tag{EC.6a}$$

$$t^{1/3} \ge \frac{8c(d_x + 3\log(tL) + d_x \log\log(tL))\sqrt{d_a d_x \log(t)}\left(1 \wedge h\sqrt{d_a + 3\log(t) + \frac{d_a}{2}\log\log t}\right)}{\tilde{c}_6(\frac{1}{d_a-1} \wedge h^2 \wedge (d_a-1)h^4\mathbb{1}\{h \ge \frac{1}{4(d_a-1)}\})}, \tag{EC.6b}$$

$$\lambda_0 \ge \sigma t^{-1/6}\frac{12\sqrt{2}}{\sqrt{L}}\sqrt{\log(tL)(d_x + 3\log(Lt))\Big(\big(d_a + 3\log(t)\big)h^2 \wedge 2\Big)\big(\log(d_a + d_x) + 2\log t\big)}. \tag{EC.6c}$$

Clearly, the burn-in period B_{init} is well-defined because there is a constant C_L depending on L only such that for any

$$t \ge C_L\Bigg(d_x^{9/2}d_a^{9/2}(\log(d_a + d_x))^6(\mathbb{1}\{h \ge \frac{1}{\sqrt{d_a - 1}}\} + \frac{\mathbb{1}\{h \ge \frac{1}{4(d_a-1)}\}}{d_a^{9/2}h^9} + \frac{\mathbb{1}\{h < \frac{1}{4(d_a-1)}\}}{d_a^{3/2}h^3}) + 1 \vee (\frac{\sigma\sqrt{d_x}}{\lambda_0})^6(d_a^3h^6 \wedge 1)\log^{12}(\frac{\sigma\sqrt{(d_a h^2 \wedge 1)}}{\lambda_0} + d_a + d_x)\Bigg),$$

Inequality (EC.6) holds. Therefore, we have an upper bound

$$B_{\text{init}} \le C_L\Bigg(d_x^{9/2}d_a^{9/2}(\log(d_a + d_x))^6(\mathbb{1}\{h \ge \frac{1}{\sqrt{d_a}}\} + \frac{\mathbb{1}\{h \ge \frac{1}{4(d_a-1)}\}}{d_a^{9/2}h^9} + \frac{\mathbb{1}\{h < \frac{1}{4(d_a-1)}\}}{d_a^{3/2}h^3}) + 1 \vee (\frac{\sigma\sqrt{d_x}}{\lambda_0})^6(d_a^3h^6 \wedge 1)\log^{12}(\frac{\sigma\sqrt{(d_a h^2 \wedge 1)}}{\lambda_0} + d_a + d_x)\Bigg).$$

Then we have for $t \ge B_{\text{init}}$,

$$\|\!|\widehat{\boldsymbol{\Theta}}_t - \boldsymbol{\Theta}^*|\!\|_{\text{F}} = \|\!|\boldsymbol{\Delta}_t|\!\|_{\text{F}} \le C_f\frac{(\sqrt{d_x\log(tL)} + 2\log(tL))\sigma + \lambda_0\sqrt{r}}{t^{\frac{1}{6}}\Big(\frac{1}{d_a-1} \wedge h^2 \wedge (d_a-1)h^4\mathbb{1}\{h \ge \frac{1}{4(d_a-1)}\}\Big)}, \tag{EC.7}$$

where $C_f > 0$ is an absolute constant.

EC.2.2. Proof of Theorem 1

Let the oracle optimal action at time t be $\boldsymbol{a}_t^*$ and $\boldsymbol{b}_t = \sum_{\ell=1}^L \boldsymbol{x}_{t,\ell}$. We can decompose the cumulative regret $T\mathcal{R}^\pi(T)$ as the sum $B_T + D_T$, where we define: (i) the cumulative regret until time B_{init}

$$B_T := \mathbb{E}\left(\sum_{t=0}^{\min(B_{\text{init}}-1, T-1)}\sum_{\ell=1}^{L}\left(\boldsymbol{a}_{t+1}^{*\top}\boldsymbol{\Theta}^*\boldsymbol{x}_{t+1,l} - \boldsymbol{a}_{t+1}^{\top}\boldsymbol{\Theta}^*\boldsymbol{x}_{t+1,l}\right)\right),$$

and (ii) the cumulative regret after the time B_{init}:

$$D_T := \mathbb{E}\left(\sum_{t=\min(B_{\text{init}},T)}^{T-1}\sum_{\ell=1}^{L}\left(\boldsymbol{a}_{t+1}^{*\top}\boldsymbol{\Theta}^*\boldsymbol{x}_{t+1,l} - \boldsymbol{a}_{t+1}^{\top}\boldsymbol{\Theta}^*\boldsymbol{x}_{t+1,l}\right)\right).$$

When $B_{\text{init}} \ge T$, $D_T = 0$ as it does not include any terms in the summation. We use the same convention for all the summations introduced later: when the index of the beginning of the summation is larger than that of the end, then the summation is set to 0.

In what follows, we derive bounds for D_T and B_T.

EC.2.2.1. Bounding D_T: We begin by bounding D_T. In order to do so, we let $\mathcal{E}_t$ denote the event that Inequality (EC.7) holds, and $\mathcal{E}_t^c$ be its complement. By Proposition 1, we have the bound

$$\mathbb{P}(\mathcal{E}_t^c) \le \frac{5}{t} + \frac{2}{Lt} + \frac{2}{t^2} + \frac{2}{L^3 t^3} \qquad \text{for all } t \ge B_{\text{init}}.$$

Moreover, the event $\mathcal{E}_t^c$ and the estimation of $\widehat{\boldsymbol{\Theta}}_t$ are jointly independent of the random vector $\boldsymbol{b}_{t+1}$.

Now, by definition, we have

$$\boldsymbol{a}_{t+1}^* = \frac{\boldsymbol{\Theta}^* \boldsymbol{b}_{t+1}}{\|\boldsymbol{\Theta}^* \boldsymbol{b}_{t+1}\|_2}. \tag{EC.8}$$

We introduce a lemma that we will prove later.

LEMMA EC.1. *For $t \ge 0$, the conditional expectation of the action $\boldsymbol{a}_{t+1}$ generated by Algorithm 1 satisfies*

$$\mathbb{E}(\boldsymbol{a}_{t+1} \mid \{\boldsymbol{x}_{i,\ell}\}_{j \le t+1, \ell \le L}, \{\boldsymbol{a}_i\}_{i \le t}, \{y_{i,\ell}\}_{i \le t, \ell \le L}) = \begin{cases} \frac{\widehat{\boldsymbol{\Theta}}_t \boldsymbol{b}_{t+1}}{\|\widehat{\boldsymbol{\Theta}}_t \boldsymbol{b}_{t+1}\|_2}, & \text{for } t \notin \{\lfloor w^{\frac{3}{2}} \rfloor : w \in \mathbb{Z}_+\} \\ \alpha_h \frac{\widehat{\boldsymbol{\Theta}}_t \boldsymbol{b}_{t+1}}{\|\widehat{\boldsymbol{\Theta}}_t \boldsymbol{b}_{t+1}\|_2}, & \text{for } t \in \{\lfloor w^{\frac{3}{2}} \rfloor : w \in \mathbb{Z}_+\} \end{cases}, \tag{EC.9}$$

where

$$\alpha_h = \mathbb{E}\left(\frac{1 + h\varepsilon_1}{1 \vee \sqrt{(1+h\varepsilon_1)^2 + h^2 \sum_{i=2}^{d_a} \varepsilon_j^2}}\right) \text{ for } \varepsilon_i \overset{i.i.d.}{\sim} N(0,1). \tag{EC.10}$$

In addition,

$$0 \le 1 - \alpha_h \le h\left(1 + \frac{\sqrt{d_a}}{2}\right) \wedge 1. \tag{EC.11}$$

Next we simplify D_T by first taking expectation conditioned on $\{\boldsymbol{x}_{\tau,\ell}\}_{\tau \le t+1, \ell \le L}$, $\{\boldsymbol{a}_\tau, y_{\tau,\ell}\}_{\tau \le t, \ell \le L}$ for each term in the sum and substitute the two relations from Equation (EC.8) and Equation (EC.9), then upper bound it with Inequality (EC.11). For simplicity of notation, let

$$\alpha_{t,h} = \begin{cases} 1, & t-1 \notin \{\lfloor w^{\frac{3}{2}} : w \in \mathbb{Z}_+ \rfloor\} \\ \alpha_h, & t-1 \in \{\lfloor w^{\frac{3}{2}} : w \in \mathbb{Z}_+ \rfloor\} \end{cases}.$$

Therefore,

$$\begin{aligned} D_T &= \mathbb{E}\left(\sum_{t=\min(B_{\text{init}},T)}^{T-1} \left\langle \frac{\boldsymbol{\Theta}^* \boldsymbol{b}_{t+1}}{\|\boldsymbol{\Theta}^* \boldsymbol{b}_{t+1}\|_2} - \frac{\widehat{\boldsymbol{\Theta}}_t \boldsymbol{b}_{t+1}}{\|\widehat{\boldsymbol{\Theta}}_t \boldsymbol{b}_{t+1}\|_2} + (1-\alpha_{t,h}) \frac{\widehat{\boldsymbol{\Theta}}_t \boldsymbol{b}_{t+1}}{\|\widehat{\boldsymbol{\Theta}}_t \boldsymbol{b}_{t+1}\|_2}, \boldsymbol{\Theta}^* \boldsymbol{b}_{t+1} \right\rangle\right) \\ &\overset{(i)}{\le} \mathbb{E}\left(\sum_{t=\min(B_{\text{init}},T)}^{T-1} \left\langle \frac{\boldsymbol{\Theta}^* \boldsymbol{b}_{t+1}}{\|\boldsymbol{\Theta}^* \boldsymbol{b}_{t+1}\|_2} - \frac{\widehat{\boldsymbol{\Theta}}_t \boldsymbol{b}_{t+1}}{\|\widehat{\boldsymbol{\Theta}}_t \boldsymbol{b}_{t+1}\|_2}, \boldsymbol{\Theta}^* \boldsymbol{b}_{t+1} \right\rangle\right) + \frac{(1-\alpha_h)}{T^{1/3}} \sqrt{L} \|\boldsymbol{\Theta}^*\|_{\text{F}} \mathbb{1}\{T > B_{\text{init}}\} \\ &\le \underbrace{\mathbb{E}\left(\sum_{t=\min(B_{\text{init}},T)}^{T-1} \left\langle \frac{\boldsymbol{\Theta}^* \boldsymbol{b}_{t+1}}{\|\boldsymbol{\Theta}^* \boldsymbol{b}_{t+1}\|_2} - \frac{\widehat{\boldsymbol{\Theta}}_t \boldsymbol{b}_{t+1}}{\|\widehat{\boldsymbol{\Theta}}_t \boldsymbol{b}_{t+1}\|_2}, \boldsymbol{\Theta}^* \boldsymbol{b}_{t+1} \right\rangle\right)}_{\overline{D}_T} + \frac{h(1+\frac{\sqrt{d_a}}{2}) \wedge 1}{T^{1/3}} \sqrt{L} \|\boldsymbol{\Theta}^*\|_{\text{F}} \mathbb{1}\{T > B_{\text{init}}\}, \end{aligned} \tag{EC.12}$$

where step (i) follows from Cauchy-Schwarz Inequality; and we recall that $D_T = 0$ when $B_{\text{init}} \ge T$.

Next, we use the pair of events $\mathcal{E}_t$ and $\mathcal{E}_t^c$ to write the decomposition $\overline{D}_T = D_T^A + D_T^B$, where

$$D_T^A := \sum_{t=B_{\text{init}} \wedge T}^{T-1} \mathbb{E}\left(\left\langle \frac{\boldsymbol{\Theta}^* \boldsymbol{b}_{t+1}}{\|\boldsymbol{\Theta}^* \boldsymbol{b}_{t+1}\|_2} - \frac{\widehat{\boldsymbol{\Theta}}_t \boldsymbol{b}_{t+1}}{\|\widehat{\boldsymbol{\Theta}}_t \boldsymbol{b}_{t+1}\|_2}, \boldsymbol{\Theta}^* \boldsymbol{b}_{t+1} \right\rangle \mathbb{1}\{\mathcal{E}_t\}\right), \quad \text{and}$$

$$D_T^B := \sum_{t=B_{\text{init}} \wedge T}^{T-1} \mathbb{E}\left(\left\langle \frac{\boldsymbol{\Theta}^* \boldsymbol{b}_{t+1}}{\|\boldsymbol{\Theta}^* \boldsymbol{b}_{t+1}\|_2} - \frac{\widehat{\boldsymbol{\Theta}}_t \boldsymbol{b}_{t+1}}{\|\widehat{\boldsymbol{\Theta}}_t \boldsymbol{b}_{t+1}\|_2}, \boldsymbol{\Theta}^* \boldsymbol{b}_{t+1} \right\rangle \mathbb{1}\{\mathcal{E}_t^c\}\right).$$

We analyze each of these two terms in turn.

Analysis of D_T^A: By adding and subtracting terms, we can write

$$\begin{aligned}
D_T^A &= \sum_{t=(B_{\text{init}}\wedge T)}^{T-1}\left(\mathbb{E}\left(\left\langle\frac{(\boldsymbol{\Theta}^*-\widehat{\boldsymbol{\Theta}}_t)\boldsymbol{b}_{t+1}}{\|\boldsymbol{\Theta}^*\boldsymbol{b}_{t+1}\|_2}+\frac{\|\widehat{\boldsymbol{\Theta}}_t\boldsymbol{b}_{t+1}\|_2-\|\boldsymbol{\Theta}^*\boldsymbol{b}_{t+1}\|_2}{\|\widehat{\boldsymbol{\Theta}}_t\boldsymbol{b}_{t+1}\|_2\|\boldsymbol{\Theta}^*\boldsymbol{b}_{t+1}\|_2}\widehat{\boldsymbol{\Theta}}_t\boldsymbol{b}_{t+1},\boldsymbol{\Theta}^*\boldsymbol{b}_{t+1}\right\rangle\mathbb{1}\{\mathcal{E}_t\}\right)\right.\\
&\le \sum_{t=(B_{\text{init}}\wedge T)}^{T-1}2\mathbb{E}\left(\|\boldsymbol{\Delta}_t\boldsymbol{b}_{t+1}\|\mathbb{1}\{\mathcal{E}_t\}\right)=\sum_{t=(B_{\text{init}}\wedge T)}^{T-1}2\mathbb{E}\left(\mathbb{E}\left(\|\boldsymbol{\Delta}_t\boldsymbol{b}_{t+1}\|\mathbb{1}\{\mathcal{E}_t\}\mid\{\boldsymbol{a}_\tau,\boldsymbol{x}_{\tau,\ell},y_\tau\}_{\tau<t,\ell\le L}\right)\right)\\
&\overset{(a)}{\le} \sum_{t=(B_{\text{init}}\wedge T)}^{T-1}2\sqrt{L}\mathbb{E}\left(|||\boldsymbol{\Delta}_t|||_{\text{F}}\right)\\
&\le 2\sqrt{L}\sum_{t=(B_{\text{init}}\wedge T)}^{T-1}C_f\frac{(\sqrt{d_x\log(tL)}+2\log(tL))\sigma+\lambda_0\sqrt{r}}{t^{\frac{1}{6}}\left(\frac{1}{d_a-1}\wedge h^2\wedge(d_a-1)h^4\mathbb{1}\{h\ge\frac{1}{4(d_a-1)}\}\right)}\\
&\le \frac{\frac{12}{5}C_f\sqrt{L}}{\frac{1}{d_a-1}\wedge h^2\wedge(d_a-1)h^4\mathbb{1}\{h\ge\frac{1}{4(d_a-1)}\}}T^{5/6}((\sqrt{d_x\log TL}+2\log TL)\sigma+\lambda_0\sqrt{r}),
\end{aligned}$$

where step (a) follows from dropping the event indicator, using Cauchy-Schwarz Inequality.

Analysis of D_T^B: We have

$$\begin{aligned}
D_T^B &\overset{(i)}{\le} \sum_{t=B_{\text{init}}\wedge T}^{T-1}2\mathbb{E}\left(\|\boldsymbol{\Theta}^*\boldsymbol{b}_{t+1}\|_2\mathbb{1}\{\mathcal{E}_t^c\}\right)\overset{(ii)}{=}\sum_{t=B_{\text{init}}\wedge T}^{T-1}2\mathbb{E}\left(\|\boldsymbol{\Theta}^*\boldsymbol{b}_{t+1}\|_2\right)\mathbb{E}\left(\mathbb{1}\{\mathcal{E}_t^c\}\right)\\
&\overset{(iii)}{\le} \sum_{t=B_{\text{init}}\wedge T}^{T-1}2\sqrt{\mathbb{E}\left(\mathbb{E}\Big(\boldsymbol{\Theta}^*\boldsymbol{b}_{t+1}\boldsymbol{b}_{t+1}^\top{\boldsymbol{\Theta}^*}^\top\mid\{\boldsymbol{a}_\tau,\boldsymbol{x}_{\tau,\ell},y_{\tau,\ell}\}_{\tau<t,\ell\le L}\Big)\right)}\left(\frac{5}{t}+\frac{2}{Lt}+\frac{2}{t^2}+\frac{2}{L^3t^3}\right)\\
&\overset{(iv)}{<} 16|||\boldsymbol{\Theta}^*|||_{\text{F}}\sqrt{L}\log(T),
\end{aligned}$$

where step (i) follows from $\left\|\frac{\boldsymbol{\Theta}^*\boldsymbol{b}_{t+1}}{\|\boldsymbol{\Theta}^*\boldsymbol{b}_{t+1}\|_2}-\frac{\widehat{\boldsymbol{\Theta}}_t\boldsymbol{b}_{t+1}}{\|\widehat{\boldsymbol{\Theta}}_t\boldsymbol{b}_{t+1}\|_2}\right\|_2\le 2$; step (ii) follows from $\left(\mathcal{E}_t^c,\widehat{\boldsymbol{\Theta}}_t\right)\perp\!\!\!\perp\boldsymbol{b}_{t+1}$; step (iii) uses the Cauchy–Schwarz inequality, conditional expectation, and $\mathbb{P}(\mathcal{E}_t^c)\le\frac{5}{t}+\frac{2}{Lt}+\frac{2}{t^2}+\frac{2}{L^3t^3}$ for $t\ge B_{\text{init}}$; and step (iv) follows from elementary calculation.

Putting together the pieces: Combining our bounds on D_T^A, D_T^B, and Inequality (EC.12) we find that

$$\begin{aligned}
D_T < 16|||\boldsymbol{\Theta}^*|||_{\text{F}}\sqrt{L}\log T+\frac{\frac{12}{5}\sqrt{L}C_f}{\frac{1}{d_a-1}\wedge h^2\wedge(d_a-1)h^4\mathbb{1}\{h\ge\frac{1}{4(d_a-1)}\}}T^{5/6}(\sqrt{d_x\log TL}+2\log TL+\lambda_0\sqrt{r})\\
+\frac{h(1+\frac{\sqrt{d_a}}{2})\wedge 1}{T^{1/3}}\sqrt{L}|||\boldsymbol{\Theta}^*|||_{\text{F}}. \quad \text{(EC.13)}
\end{aligned}$$

Proof of Lemma EC.1 For $t\notin\{\lfloor w^{\frac{3}{2}}\rfloor:w\in\mathbb{Z}_+\}$, the statement holds trivially.

Now we turn to $t\in\{\lfloor w^{\frac{3}{2}}\rfloor:w\in\mathbb{Z}_+\}$. Let $U(v)$ be a rotation matrix that rotates $(1,0,\ldots,0)^\top$ to $\frac{v}{\|v\|_2}$. As there are many rotation matrices that rotate $(1,0,\ldots,0)$ to v, for any v, we just pick any one. Therefore, U is a well defined map that maps a vector to a unitary matrix.

Therefore,

$$\begin{aligned}
&\mathbb{E}(\boldsymbol{a}_{t+1}\mid\{\boldsymbol{x}_{i,\ell}\}_{j\le t+1,\ell\le L},\{\boldsymbol{a}_i\}_{i\le t},\{y_{i,\ell}\}_{i\le t,\ell\le L})\\
=&\mathbb{E}(\mathcal{P}_{\mathcal{A}_t}(\frac{\hat{\boldsymbol{\Theta}}_t\boldsymbol{b}_{t+1}}{\|\hat{\boldsymbol{\Theta}}_t\boldsymbol{b}_{t+1}\|_2}+\boldsymbol{\delta}_{t+1})\mid\{\boldsymbol{x}_{i,\ell}\}_{j\le t+1,\ell\le L},\{\boldsymbol{a}_i\}_{i\le t},\{y_{i,\ell}\}_{i\le t,\ell\le L})
\end{aligned}$$

$$
\begin{aligned}
=&\mathbb{E}\left(\mathcal{P}_{\mathcal{A}_t}\Big(U(\frac{\hat{\boldsymbol{\Theta}}_t\boldsymbol{b}_{t+1}}{\|\hat{\boldsymbol{\Theta}}_t\boldsymbol{b}_{t+1}\|_2})U(\frac{\hat{\boldsymbol{\Theta}}_t\boldsymbol{b}_{t+1}}{\|\hat{\boldsymbol{\Theta}}_t\boldsymbol{b}_{t+1}\|_2})^\top(\frac{\hat{\boldsymbol{\Theta}}_t\boldsymbol{b}_{t+1}}{\|\hat{\boldsymbol{\Theta}}_t\boldsymbol{b}_{t+1}\|_2}+\boldsymbol{\delta}_{t+1})\Big)\middle|\{\boldsymbol{x}_{i,l}\}_{j\le t+1,l\le L},\{\boldsymbol{a}_i\}_{i\le t},\{y_{i,l}\}_{i\le t,l\le L}\right)\\
=&U(\frac{\hat{\boldsymbol{\Theta}}_t\boldsymbol{b}_{t+1}}{\|\hat{\boldsymbol{\Theta}}_t\boldsymbol{b}_{t+1}\|_2})\mathbb{E}\left(\mathcal{P}_{\mathcal{A}_t}\Big((1,0,\ldots,0)+U(\frac{\hat{\boldsymbol{\Theta}}_t\boldsymbol{b}_{t+1}}{\|\hat{\boldsymbol{\Theta}}_t\boldsymbol{b}_{t+1}\|_2})^\top\boldsymbol{\delta}_{t+1}\Big)\middle|\{\boldsymbol{x}_{i,l}\}_{j\le t+1,l\le L},\{\boldsymbol{a}_i\}_{i\le t},\{y_{i,l}\}_{i\le t,l\le L}\right)\\
=&U(\frac{\hat{\boldsymbol{\Theta}}_t\boldsymbol{b}_{t+1}}{\|\hat{\boldsymbol{\Theta}}_t\boldsymbol{b}_{t+1}\|_2})\underbrace{\mathbb{E}\left(\mathcal{P}_{\mathcal{A}_t}\Big((1,0,\ldots,0)+h(\varepsilon_1,\ldots,\varepsilon_{d_a})^\top\Big)\middle|\{\boldsymbol{x}_{i,l}\}_{j\le t+1,l\le L},\{\boldsymbol{a}_i\}_{i\le t},\{y_{i,l}\}_{i\le t,l\le L}\right)}_{\nu},
\end{aligned}
\tag{EC.14}
$$

where $\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_{d_a}\overset{i.i.d}{\sim}N(0,1)$, and $\{\varepsilon_i\}_{1\le i\le d_a}$ are independent from $\Big\{\{\boldsymbol{x}_{i,l}\}_{j\le t+1,l\le L},\{\boldsymbol{a}_i\}_{i\le t},\{y_{i,l}\}_{i\le t,l\le L}\Big\}$. Due to the symmetricity of $\{\varepsilon_i\}_{i\ge 2}$, we have

$$
\nu=\mathbb{E}\left(\mathcal{P}_{\mathcal{A}_t}\Big((1,0,\ldots,0)+h(\varepsilon_1,\ldots,\varepsilon_{d_a})^\top\Big)\right)=\mathbb{E}\left(\Big(\frac{1+h\varepsilon_1}{\sqrt{(1+h\varepsilon_1)^2+h^2\sum_{i=2}^{d_a}\varepsilon_i^2}},0,\ldots,0\Big)\right).
$$

Going back to Equation (EC.14), we have

$$
\mathbb{E}(\boldsymbol{a}_{t+1}\mid\{\boldsymbol{x}_{i,\ell}\}_{j\le t+1,\ell\le L},\{\boldsymbol{a}_i\}_{i\le t},\{y_{i,\ell}\}_{i\le t,\ell\le L})=\frac{\hat{\boldsymbol{\Theta}}_t\boldsymbol{b}_{t+1}}{\|\hat{\boldsymbol{\Theta}}_t\boldsymbol{b}_{t+1}\|_2}\mathbb{E}\Big(\frac{1+h\varepsilon_1}{\sqrt{(1+h\varepsilon_1)^2+h^2\sum_{i=2}^{d_a}\varepsilon_i^2}}\Big).
\tag{EC.15}
$$

This gives Equation (EC.9) and Equation (EC.10).

Now we move to proving the statement about α_h in Inequality (EC.11).

First note that $\alpha_h\le 1$ follows directly from

$$
\Big|\frac{1+h\varepsilon_1}{(1+h\varepsilon_1)^2+h^2\sum_{i=2}^{d_a}\varepsilon_i^2}\Big|\le 1.
$$

Next, we simplify α_h into an expectation of terms that are easier to analyze. Taking conditional expectation conditioned on $(|\varepsilon_1|,\varepsilon_2,\ldots,\varepsilon_{d_a})$ and using the symmetricity of ε_1 gives

$$
\begin{aligned}
\alpha_h=&\mathbb{E}\left(\mathbb{E}(\frac{1+h\varepsilon_1}{1\vee\sqrt{(1+h\varepsilon_1)^2+h^2\sum_{i=2}^{d_a}\varepsilon_i^2}}\middle||\varepsilon_1|,\varepsilon_2,\ldots,\varepsilon_{d_a})\right)\\
=&\frac{1}{2}\mathbb{E}\left(\mathbb{E}(\frac{1+h|\varepsilon_1|}{1\vee\sqrt{(1+h|\varepsilon_1|)^2+h^2\sum_{i=2}^{d_a}\varepsilon_i^2}}+\frac{1-h|\varepsilon_1|}{1\vee\sqrt{(1-h|\varepsilon_1|)^2+h^2\sum_{i=2}^{d_a}\varepsilon_i^2}}\middle||\varepsilon_1|,\varepsilon_2,\ldots,\varepsilon_{d_a})\right).
\end{aligned}
$$

Clearly, $\sqrt{(1+h|\varepsilon_1|)^2+\sum_{i=2}^{d_a}h^2\varepsilon_i^2}\ge 1$, so the denominator in the first term is always 1. We separate the probability space into two disjoint sets based on whether the denominator of the second term is smaller than 1:

$$
\mathcal{H}_1=\{(1-h|\varepsilon_1|)^2+h^2\sum_{i=2}^{d_a}\varepsilon_i^2<1\},\qquad\mathcal{H}_2=\{(1-h|\varepsilon_1|)^2+h^2\sum_{i=2}^{d_a}\varepsilon_i^2\ge 1\}.
$$

Therefore, we use the pair of events $\mathcal{H}_1$ and $\mathcal{H}_2$ to decompose $1-\alpha$:

$$
\begin{aligned}
&1-\alpha_h=\\
&\underbrace{\mathbb{E}\left(\mathbb{E}\left(1-\frac{1+h|\varepsilon_1|}{2\sqrt{(1+h|\varepsilon_1|)^2+h^2\sum_{i=2}^{d_a}\varepsilon_i^2}}-\frac{1-h|\varepsilon_1|}{2}\middle||\varepsilon_1|,\varepsilon_2,\ldots,\varepsilon_{d_a}\right)\mathbb{1}\{\mathcal{H}_1\}\right)}_{\beta_h^1}+
\end{aligned}
$$

$$\underbrace{\mathbb{E}\left(\mathbb{E}\left(1-\frac{1+h|\varepsilon_1|}{2\sqrt{(1+h|\varepsilon_1|)^2+h^2\sum_{i=2}^{d_a}\varepsilon_i^2}}-\frac{1-h|\varepsilon_1|}{2\sqrt{(1-h|\varepsilon_1|)^2+h^2\sum_{i=2}^{d_a}\varepsilon_i^2}}\middle||\varepsilon_1|,\varepsilon_2,\ldots,\varepsilon_{d_a}\right)\mathbb{1}\{\mathcal{H}_2\}\right)}_{\beta_h^2}.$$

Next, we will bound β_h^1 and β_h^2 in turn.

We start with β_h^1. Note that event $\mathcal{H}_1$ implies $h^2\sum_{i=1}^{d_a}\varepsilon_i^2<2h|\varepsilon_1|$ and $|\varepsilon_1|h<2$. Using these two relationships after simplifying β_h^1 gives

$$\begin{aligned}\beta_h^1=&\mathbb{E}\left(\frac{1}{2}(1+|\varepsilon_1|h)(1-\frac{1}{\sqrt{(1+|\varepsilon_1|)^2+h^2\sum_{i=2}^{d_a}\varepsilon_i^2}})\mathbb{1}\{\mathcal{H}_1\}\right)\\ \overset{(a)}{<}&\mathbb{E}\left(\frac{1}{2}(1+|\varepsilon_1|h)\left(1-\frac{1}{\sqrt{1+4|\varepsilon_1|h}}\right)\mathbb{1}\{\mathcal{H}_1\}\right)\\ =&\mathbb{E}\left(\frac{1}{2}(1+|\varepsilon_1|h)\frac{4|\varepsilon_1|h}{\sqrt{1+4|\varepsilon_1|h}(1+\sqrt{1+4|\varepsilon_1|h})}\mathbb{1}\{\mathcal{H}_1\}\right)\\ \overset{(b)}{<}&\mathbb{E}\left((|\varepsilon_1|h\wedge 1)\mathbb{1}\{\mathcal{H}_1\}\right).\end{aligned}\tag{EC.16}$$

Step (a) follows from $h^2\sum_{i=1}^{d_a}\varepsilon_i^2<2h|\varepsilon_1|$. Step (b) follows from $1+|\varepsilon_1|h<\sqrt{1+4|\varepsilon_1|h}$, $|\varepsilon_1|h<2$, and that $\frac{2x}{1+\sqrt{1+4x}}$ is an increasing function.

Now we turn to bounding β_h^2. For notation simplicity, let $r^2=h^2\sum_{i=2}^{d_a}\varepsilon_i^2$. We have

$$\begin{aligned}\beta_h^2=&\mathbb{E}\left(\left(1-\frac{1}{2\sqrt{(1+|\varepsilon_1|h)^2+r^2}}-\frac{1}{2\sqrt{(1-|\varepsilon_1|h)^2+r^2}}\right)\mathbb{1}\{\mathcal{H}_2\}\right)+\\ &\mathbb{E}\left(\frac{1}{2}|\varepsilon_1|h\frac{4|\varepsilon_1|h}{\sqrt{(1+|\varepsilon_1|h)^2+r^2}\sqrt{(1-|\varepsilon_1|h)^2+r^2}\left(\sqrt{(1+|\varepsilon_1|h)^2+r^2}+\sqrt{(1-|\varepsilon_1|h)^2+r^2}\right)}\mathbb{1}\{\mathcal{H}_2\}\right)\\ \overset{(c)}{\leq}&\mathbb{E}\left((1-\frac{1}{\sqrt{1+\varepsilon_1^2h+r^2}})\mathbb{1}\{\mathcal{H}_2\}\right)+\mathbb{E}\left(\frac{2|\varepsilon_1|h}{\sqrt{(1+|\varepsilon_1|h)^2+r^2}+\sqrt{(1-|\varepsilon_1|h)^2+r^2}}\mathbb{1}\{\mathcal{H}_2\}\right)\\ \leq&\mathbb{E}\left(\frac{\varepsilon_1^2h^2+r^2}{\sqrt{1+\varepsilon_1^2h^2+r^2}(1+\sqrt{1+\varepsilon_1^2h^2+r^2})}\mathbb{1}\{\mathcal{H}_2\}\right)+\mathbb{E}\left(|\varepsilon_1|h\mathbb{1}\{\mathcal{H}_2\}\right)\\ \overset{(d)}{\leq}&h(1+\frac{\sqrt{d_a}}{2})\mathbb{P}(\mathcal{H}_2)\end{aligned}\tag{EC.17}$$

where step (c) follows from the convexity of $\sqrt{1/x}$ and that $\mathcal{H}_2$ implies $(1-|\varepsilon_1|h)^2+r^2\geq 1$, step (d) follows from applying Cauchy-schwarz Inequality to both terms and that $\frac{\sqrt{\varepsilon_1^2h^2+r^2}}{\sqrt{1+\varepsilon_1^2h^2+r^2}(1+\sqrt{1+\varepsilon_1^2h^2+r^2})}\leq\frac{1}{2}$.

Next, we use another way to show that

$$\beta_h^2\leq\mathbb{P}(\mathcal{H}_2).\tag{EC.18}$$

This is equivalent to showing

$$\begin{aligned}&\frac{1}{2}|\varepsilon_1|h\frac{4|\varepsilon_1|h}{\sqrt{(1+|\varepsilon_1|h)^2+r^2}\sqrt{(1-|\varepsilon_1|h)^2+r^2}\left(\sqrt{(1+|\varepsilon_1|h)^2+r^2}+\sqrt{(1-|\varepsilon_1|h)^2+r^2}\right)}\\ &\leq\frac{1}{2\sqrt{(1+|\varepsilon_1|h)^2+r^2}}+\frac{1}{2\sqrt{(1-|\varepsilon_1|h)^2+r^2}},\end{aligned}\tag{EC.19}$$

which is further equivalent to showing

$$2|\varepsilon_1|h \le \sqrt{(1+|\varepsilon_1|h)^2+r^2}+\sqrt{(1-|\varepsilon_1|h)^2+r^2}, \tag{EC.20}$$

which holds trivially due to $2|\varepsilon_1|h \le \big|1+|\varepsilon_1|h\big| + \big||\varepsilon_1|h-1\big|$.

Combining inequalities Equation (EC.16), Equation (EC.17), and Equation (EC.18) gives

$$1-\alpha_h < h(1+\frac{\sqrt{d_a}}{2}) \wedge 1. \tag{EC.21}$$

This finish the proof of Lemma EC.1.

EC.2.2.2. Bounding B_T: By similar arguments, we can establish the bound

$$B_T = \mathbb{E}\left(\sum_{t=0}^{(B_{\text{init}}-1)\wedge(T-1)}\sum_{\ell=1}^{L}\mathbb{E}\left(\boldsymbol{a}_t^{*\top}\boldsymbol{\Theta}^*\boldsymbol{x}_{t,\ell}-\boldsymbol{a}_t^{\top}\boldsymbol{\Theta}^*\boldsymbol{x}_{t,\ell}\right)\right) \le B_{\text{init}}\left\{2\sqrt{L}|||\boldsymbol{\Theta}^*|||_{\text{F}}\right\}.$$

Therefore,

$$\begin{aligned}\mathcal{R}^{\pi}(T) = \frac{B_T+D_T}{T} &\le 2\sqrt{L}|||\boldsymbol{\Theta}^*|||_{\text{F}}B_{\text{init}}T^{-1}+16|||\boldsymbol{\Theta}^*|||_{\text{F}}\sqrt{L}\frac{\log T}{T}+\frac{(1\wedge 2h\sqrt{d_a})}{T^{1/3}}\sqrt{L}|||\boldsymbol{\Theta}^*|||_{\text{F}}\\ &+\frac{\frac{12}{5}\sqrt{L}C_f}{\frac{1}{d_a-1}\wedge h^2\wedge(d_a-1)h^4\mathbb{1}\{h\ge\frac{1}{4(d_a-1)}\}}T^{-1/6}((\sqrt{d_x\log TL}+2\log TL)\sigma+\lambda_0\sqrt{r}).\end{aligned} \tag{EC.22}$$

Let $c_1=2$, $c_2=16$, $c_3=\frac{12}{5}C_f$, gives the statement.

EC.2.3. Proof of Lemma 2

Let us prove the claim with t replaced by T, so as to allow ourselves to use t as an index of summation. Recall that the noisy reward takes the form $y_{t,\ell}=\boldsymbol{a}_t^T\boldsymbol{\Theta}^*\boldsymbol{x}_{t,\ell}+\sigma\varepsilon_{t,\ell}$, so that the loss gradient can be decomposed as

$$\nabla\mathcal{L}_T(\boldsymbol{\Theta}^*) = \frac{\sigma}{LT}\sum_{t=1}^{T}\sum_{\ell=1}^{L}-\varepsilon_{t,\ell}\boldsymbol{x}_{t,\ell}\boldsymbol{a}_t^{\top} \tag{EC.23}$$

$$= \underbrace{\frac{\sigma}{LT}\sum_{t=2}^{T}\sum_{\ell=1}^{L}-\varepsilon_{t,\ell}\boldsymbol{x}_{t,\ell}\hat{\boldsymbol{a}}_t^{\top}}_{\mathbf{S}_1} + \underbrace{\frac{\sigma}{LT}\sum_{t=2}^{T}\sum_{\ell=1}^{L}-\varepsilon_{t,\ell}\boldsymbol{x}_{t,\ell}(\boldsymbol{a}_t-\hat{\boldsymbol{a}}_t)^{\top}}_{\mathbf{S}_2}, \tag{EC.24}$$

where we let $\hat{\boldsymbol{a}}_1=\boldsymbol{a}_1$.

From this decomposition, we have the upper bound

$$\big|\langle\!\langle\nabla\mathcal{L}_T(\Theta^*),\,\boldsymbol{\Delta}\rangle\!\rangle\big| = \big|\langle\!\langle\mathbf{S}_1,\,\boldsymbol{\Delta}_T\rangle\!\rangle+\langle\!\langle\mathbf{S}_2,\,\boldsymbol{\Delta}\rangle\!\rangle\big| \le |||\mathbf{S}_1|||_{\text{F}}|||\boldsymbol{\Delta}|||_{\text{F}}+|||\mathbf{S}_2|||_{\text{op}}|||\boldsymbol{\Delta}|||_{\text{nuc}}. \tag{EC.25}$$

Consequently, in order to establish the claim of the lemma, it suffices to show that the inequalities

$$|||\mathbf{S}_1|||_{\text{F}} \le \underbrace{\frac{2\sigma}{\sqrt{T}}(\sqrt{d_x\log(TL)}+2\log(TL))}_{\phi_1(T)} \tag{EC.26a}$$

$$|||\mathbf{S}_2|||_{\text{op}} \le \underbrace{\frac{\sigma}{T^{2/3}}\frac{12\sqrt{2}}{\sqrt{L}}\sqrt{\log(TL)(d_x+3\log(LT))\big((d_a+3\log(T))h^2\wedge 2\big)\big(\log(d_a+d_x)+2\log T\big)}}_{\phi_2(T)} \tag{EC.26b}$$

both hold with the probability claimed in the lemma statement.

In order to prove these bounds, let us recall a basic concentration inequality (cf. Lemma 1 in Laurent and Massart (2000)): for a χ^2-variable U with k degrees of freedom, we have

$$\mathbb{P}\Big[U - k \geq 2\sqrt{k\nu} + 2\nu\Big] \leq \exp(-\nu) \qquad \text{for any } \nu > 0. \tag{EC.27}$$

Define the event

$$\mathcal{J}_T := \Big\{ \max_{\substack{t\in[T]\\ \ell\in[L]}} \|\boldsymbol{x}_{t,\ell}\|_2^2 \leq d_x + 2\sqrt{2d_x \log(TL)} + 4\log(TL)\}\Big\}.$$

Since the random variable $\|\boldsymbol{x}_{t,\ell}\|_2^2$ follows a χ^2-distribution with degree of freedom d_x, applying Inequality (EC.27) yields

$$\mathbb{P}(\mathcal{J}_T) \geq 1 - TL\,\frac{1}{T^2L^2} \;=\; 1 - \frac{1}{TL}.$$

EC.2.3.1. Proof of the bound (EC.26a): We introduce the convenient shorthand

$$M := d_x + 2\sqrt{2d_x \log(TL)} + 4\log(TL),$$

and define the sum $W(\ell;T) := \sum_{t=1}^{T} -\varepsilon_{t,\ell}\boldsymbol{x}_{t,\ell}\hat{\boldsymbol{a}}_t^\top$ for $T \geq 1$ and $W(\ell;0) := 0$. Denote the history up to and including time T by

$$H_T := \big\{\boldsymbol{x}_{t,\ell}, \boldsymbol{a}_{t,\ell}, y_{t,\ell} \mid t = 1,\ldots,T,\ \ell = 1,\ldots,L\big\} \tag{EC.28}$$

for $T \geq 1$ and $H_0 = \{1\}$. Note that the noise variables $\big\{\varepsilon_{T,\ell} \mid \ell = 1,\ldots,L\big\}$ at time T are independent of H_{T-1}. For $\lambda > 0$, elementary calculation shows that the quantity $Q := \mathbb{E}\left[\exp\left(\lambda|||W(\ell;T)|||_{\mathrm{F}}^2\right)\mathbb{1}\{\mathcal{J}_T\}\right]$ can be upper bounded as

$$\begin{aligned}
Q &\leq \mathbb{E}\left[\mathbb{E}\Big[\exp\left(\lambda|||W(\ell;T-1)|||_{\mathrm{F}}^2 - 2\lambda\hat{\boldsymbol{a}}_T^\top W(\ell;T-1)^\top \boldsymbol{x}_{T,\ell}\varepsilon_{T,\ell} + \lambda\|\boldsymbol{x}_{T,\ell}\|_2^2\varepsilon_{T,\ell}^2\|\hat{\boldsymbol{a}}_T\|_2^2\right)\mathbb{1}\{\mathcal{J}_T\} \mid H_{T-1}, \boldsymbol{x}_{T,\cdot}\Big]\right] \\
&= \mathbb{E}\left[\frac{1}{1-2\lambda\|\boldsymbol{x}_{T,\ell}\|_2^2\|\hat{\boldsymbol{a}}_T\|_2^2}\exp\left(\lambda|||W(\ell;T-1)|||_{\mathrm{F}}^2 + \frac{(-2\lambda\hat{\boldsymbol{a}}_T^\top W(\ell;T-1)^\top \boldsymbol{x}_{T,\ell})^2}{2(1-2\lambda\|\boldsymbol{x}_{T,\ell}\|_2^2\|\hat{\boldsymbol{a}}_T\|_2^2)}\right)\mathbb{1}\{\mathcal{J}_T\}\right] \\
&\leq \frac{1}{1-2\lambda M}\mathbb{E}\left[\exp\left(\lambda|||W(\ell;T-1)|||_{\mathrm{F}}^2 + 2\lambda^2\frac{|||W(\ell;T-1)|||_{\mathrm{op}}^2 M}{1-2\lambda M}\right)\mathbb{1}\{\mathcal{J}_T\}\right] \\
&\leq \frac{1}{1-2\lambda M}\mathbb{E}\left[\exp\left(\frac{\lambda}{1-2\lambda M}\|W(\ell;T-1)\|_F^2\right)\mathbb{1}\{\mathcal{J}_T\}\right].
\end{aligned}$$

Setting $\lambda = \frac{1}{4TM}$ and recursively applying the above arguments, we find that

$$\mathbb{E}\Big[\exp\big(\frac{1}{4TM}|||W(\ell;T)|||_{\mathrm{F}}^2\big)\mathbb{1}\{\mathcal{J}_T\}\Big] \;\leq\; \prod_{t=T+1}^{2T}\Big(\frac{1}{1-1/t}\Big) = 2. \tag{EC.29}$$

Therefore, for any $s > 0$, we have

$$\begin{aligned}
\mathbb{P}\Big[|||W(\ell;T)|||_{\mathrm{F}}^2\mathbb{1}\{\mathcal{J}_T\} \geq s^2\Big] &= \mathbb{P}\Big[\exp(\tfrac{1}{4TM}|||W(\ell;T)|||_{\mathrm{F}}^2\mathbb{1}\{\mathcal{J}_T\}) \geq \exp(\tfrac{1}{4TM}s^2)\Big] \\
&\leq 2\exp(-\frac{1}{4TM}s^2),
\end{aligned}$$

where the last step uses Markov's inequality, along with the bound (EC.29).

We now set $s := \sqrt{4TM\log(TL)} = \sqrt{4T(d_x + 2\sqrt{2d_x\log(TL)} + 4\log(TL))\log(TL)}$, and find that

$$\mathbb{P}\Big[\big\{|||W(\ell;T)|||_{\mathrm{F}} \leq s \text{ for all } \ell = 1,\ldots,L\big\}^c \cap \mathcal{J}_T\Big] \leq L\frac{2}{TL}.$$

Noting that $s \le \sqrt{T}(2\sqrt{d_x \log(TL)} + 4\log(TL))$, we have that

$$\mathbb{P}\Big[\big\{|\!|\!|\mathbf{S}_1|\!|\!|_{\mathrm{F}} > \frac{\sigma}{\sqrt{T}}(2\sqrt{d_x \log(TL)} + 4\log(TL))\big\} \cap \mathcal{J}_T\Big] \le \frac{2}{T} \tag{EC.30}$$

EC.2.3.2. Proof of the bound (EC.26b): Turning to the analysis of $\mathbf{S}_2$, consider the inequalities

$$\max_{\substack{t\in[T]\\ \ell\in[L]}} |\varepsilon_{t,\ell}| \le 3\sqrt{\log(TL)}, \quad \max_{\substack{t\in[T]\\ \ell\in[L]}} \|\boldsymbol{x}_{t,\ell}\|_2^2 \le 2d_x + 6\log(LT), \quad \text{and} \quad \max_{t\in[T]} \|\boldsymbol{\delta}_t/h\|_2^2 \le 2d_a + 6\log(T).$$

and let $\mathcal{G}$ be the event that all three hold simultaneously. An elementary calculation shows that

$$\mathbb{P}(\mathcal{G}^c) \le \frac{2}{T^3L^3} + \frac{1}{LT} + \frac{1}{T}, \tag{EC.31}$$

and moreover, we have the inclusion $\mathcal{G} \subset \mathcal{J}_T$.

Define the following truncated variables:

$$\tilde{\varepsilon}_{t,\ell} = \varepsilon_{t,\ell}\mathbb{1}\{|\varepsilon_{t,\ell}| \le 3\sqrt{\log(TL)}\}, \quad \tilde{\mathbf{x}}_{t,\ell} = \boldsymbol{x}_{t,\ell}\mathbb{1}\{\|\boldsymbol{x}_{t,\ell}\|_2^2 \le 2d_x + 6\log(LT)\},$$
$$\tilde{\boldsymbol{\delta}}_t = (\boldsymbol{a}_t - \hat{\boldsymbol{a}}_t)\mathbb{1}\{\|\boldsymbol{\delta}_t/h\|_2^2 \le 2d_a + 6\log(T)\}, \quad \text{and} \quad \tilde{\mathbf{S}}_2 = \frac{\sigma}{LT}\sum_{t=2}^{T}\sum_{\ell=1}^{L} -\tilde{\varepsilon}_{t,\ell}\tilde{\mathbf{x}}_{t,\ell}\tilde{\boldsymbol{\delta}}_t^\top.$$

Clearly, on the event $\mathcal{G}$, we have the equivalence $\mathbf{S}_2 = \tilde{\mathbf{S}}_2$. Therefore, for any $\alpha > 0$, we have that

$$\mathbb{P}(\{|\!|\!|\mathbf{S}_2|\!|\!|_{\mathrm{op}} \ge \alpha\} \cap \mathcal{G}) = \mathbb{P}(|\!|\!|\tilde{\mathbf{S}}_2|\!|\!|_{\mathrm{op}} \ge \alpha). \tag{EC.32}$$

Recall the definition of $\boldsymbol{a}_t - \hat{\boldsymbol{a}}_t$, either $\boldsymbol{a}_t = \hat{\boldsymbol{a}}_t$, or

$$\|\boldsymbol{a}_t - \hat{\boldsymbol{a}}_t\|_2^2 = \|\mathcal{P}_{\mathcal{A}_{t-1}}(\hat{\boldsymbol{a}}_t + \boldsymbol{\delta}_t) - \hat{\boldsymbol{a}}\|_2^2 \le \min\{\|\boldsymbol{\delta}_t\|_2^2, 4\}, \tag{EC.33}$$

which gives $\|\tilde{\boldsymbol{\delta}}_t\|_2^2 \le \min\{(2d_a + 6\log T)h^2, 4\}$.

Define a shorthand:

$$\sigma_{\tilde{\mathbf{S}}} := 3\sqrt{2}\sqrt{\log(TL)(d_x + 3\log(LT))}\min\{2, h\sqrt{2(d_a + 3\log(T))}\}. \tag{EC.34}$$

Define two series of matrices

$$\mathbf{B}_{t,\ell} := -\tilde{\varepsilon}_{t,\ell}\tilde{\mathbf{x}}_{t,\ell}\tilde{\boldsymbol{\delta}}_t^\top \tag{EC.35}$$

$$\mathbf{A}_{t,\ell} := \begin{cases} \mathbf{0}_{(d_x+d_a)\times(d_x+d_a)} & \text{for } t-1 \notin \{\ \lfloor w^{3/2}\rfloor : w \in \mathbb{Z}_+\} \\ \sigma_{\tilde{\mathbf{S}}}\mathbf{I}_{(d_x+d_a)\times(d_x+d_a)} & \text{for } t-1 \in \{\ \lfloor w^{3/2}\rfloor : w \in \mathbb{Z}_+\} \end{cases}. \tag{EC.36}$$

Clearly, we have

$$\begin{pmatrix} \mathbf{B}_{t,\ell}\mathbf{B}_{t,\ell}^\top & \mathbf{0}_{d_x\times d_x} \\ \mathbf{0}_{d_a\times d_a} & \mathbf{B}_{t,\ell}^\top\mathbf{B}_{t,\ell} \end{pmatrix} \preceq \mathbf{A}_{t,\ell}^2, \quad |\!|\!|\sum_{t=2}^{T}\sum_{l=1}^{L}\mathbf{A}_{t,\ell}^2|\!|\!|_{\mathrm{op}} \le LT^{2/3}\sigma_{\tilde{\mathbf{S}}}^2, \tag{EC.37}$$

and the martingale property

$$\mathbb{E}\Big(\sum_{\substack{Lt'+\ell' \le Lt+\ell\\ \ell' \le L}} \mathbf{B}_{t',\ell'}\Big|\big\{\mathbf{B}_{t'',\ell''} : Lt''+\ell'' < Lt+\ell, \ell'' \le L\big\}\Big) = \sum_{\substack{Lt''+\ell'' < Lt+\ell\\ \ell'' \le L}} \mathbf{B}_{t'',\ell''} \tag{EC.38}$$

By the martingale version of matrix Hoeffding bound for rectangular matrix (Theorem 7.1 and Remark 7.3 of Tropp (2012)) we have for $\alpha > 0$,

$$\mathbb{P}\left(|\!|\!|\sum_{t=2}^{T}\sum_{l=1}^{L}\mathbf{B}_{t,\ell}|\!|\!|_{\mathrm{op}} \ge \alpha\right) \le (d_a + d_x)\exp\left(-\frac{\alpha^2}{8LT^{2/3}\sigma_{\tilde{\mathbf{S}}}^2}\right). \tag{EC.39}$$

Recall that $\tilde{\mathbf{S}}_2 = \frac{\sigma}{LT}\sum_{t=2}^{T}\sum_{l=1}^{L}\mathbf{B}_{t,\ell}$ and set $\alpha = \sqrt{8LT^{1/3}}\sigma_{\tilde{\mathbf{S}}}\sqrt{\log(d_a+d_x)+2\log T}$, we arrive at

$$\mathbb{P}\left(|||\frac{LT}{\sigma}\tilde{\mathbf{S}}_2|||_{\mathrm{op}} \geq \alpha\right) \leq \frac{1}{T^2}. \tag{EC.40}$$

The Inequality (EC.40) combined the relation (EC.32) gives

$$\mathbb{P}\Big[\Big\{|||\mathbf{S}_2|||_{\mathrm{op}} > \phi_2(T)\Big\} \cap \mathcal{G}\Big] \leq \frac{1}{T^2}, \tag{EC.41}$$

as claimed.

Therefore, Inequality (EC.30), Inequality (EC.41), Inequality (EC.31), and the fact $\mathcal{G} \subset \mathcal{J}_T$ combined together gives

$$\begin{aligned}\mathbb{P}\Big(|||\mathbf{S}_1|||_{\mathrm{F}} > \phi_1(T) \text{ or } |||\mathbf{S}_2|||_{\mathrm{op}} > \phi_2(T)\Big) &\leq \mathbb{P}\Big[\Big\{|||\mathbf{S}_1|||_{\mathrm{F}} > \phi_1(T)\Big\} \cap \mathcal{G}\Big] + \mathbb{P}\Big[\Big\{|||\mathbf{S}_2|||_{\mathrm{op}} > \phi_2(T) \cap \mathcal{G}\Big\}\Big] + \mathbb{P}(\mathcal{G}^c) \\ &\leq \frac{2}{L^3T^3} + \frac{1}{LT} + \frac{3}{T} + \frac{1}{T^2}. \end{aligned} \tag{EC.42}$$

EC.2.4. Proof of Lemma 1

For notational simplicity, we replace t in the statement of the lemma in the main paper with T in the proof. Basically, we prove that for any $T \geq 4$, we have the lower bound

$$\begin{aligned} e_T(\mathbf{\Delta}) \geq & \frac{\lfloor T^{2/3}\rfloor \tilde{c}_6\left(\frac{1}{d_a-1} \wedge h^2 \wedge (d_a-1)h^4 \mathbb{1}\{h \geq \frac{1}{4(d_a-1)}\}\right)}{2T}|||\mathbf{\Delta}|||_{\mathrm{F}}^2 \\ & - \frac{c}{T^{2/3}}(d_x + 3\log(TL) + d_x\log\log(TL))\sqrt{d_a d_x \log(T)}\left(1 \wedge h\sqrt{d_a + 3\log(T) + \frac{d_a}{2}\log\log T}\right)|||\mathbf{\Delta}|||_{\mathrm{F}}^2, \end{aligned}$$

with probability at least $1 - \frac{1}{LT} - \frac{2}{T} - \frac{1}{T^2}$, for some absolute positive constants $c > 0$ and $\tilde{c}_6 > 0$.

Denote the conditional expectation of $\boldsymbol{a}_{t+1}$ given available observations to be

$$\bar{\boldsymbol{a}}_{t+1} = \mathbb{E}(\boldsymbol{a}_{t+1} \mid \{\boldsymbol{x}_{i,\ell}\}_{j\leq t+1,\ell\leq L}, \{\boldsymbol{a}_i\}_{i\leq t}, \{y_{i,\ell}\}_{i\leq t,\ell\leq L}), \tag{EC.43}$$

and the difference between $\boldsymbol{a}_{t+1}$ and $\bar{\boldsymbol{a}}_{t+1}$ to be

$$\bar{\boldsymbol{\delta}}_{t+1} = \boldsymbol{a}_{t+1} - \bar{\boldsymbol{a}}_{t+1}, \tag{EC.44}$$

for $t \geq 0$. And let

$$\bar{\boldsymbol{a}}_1 = \boldsymbol{a}_1, \quad \bar{\boldsymbol{\delta}}_1 = 0. \tag{EC.45}$$

To prove the lemma, we take a detour to introduce some more basic properties about $\bar{\boldsymbol{a}}_t$ and $\bar{\boldsymbol{\delta}}_t$.

By Lemma EC.1, we have

$$\bar{\boldsymbol{a}}_{t+1} = \begin{cases} \frac{\hat{\mathbf{\Theta}}_t \boldsymbol{b}_{t+1}}{\|\hat{\mathbf{\Theta}}_t \boldsymbol{b}_{t+1}\|_2}, & \text{for } t \notin \{\lfloor w^{\frac{3}{2}}\rfloor : w \in \mathbb{Z}_+\} \\ \alpha_h \frac{\hat{\mathbf{\Theta}}_t \boldsymbol{b}_{t+1}}{\|\hat{\mathbf{\Theta}}_t \boldsymbol{b}_{t+1}\|_2}, & \text{for } t \in \{\lfloor w^{\frac{3}{2}}\rfloor : w \in \mathbb{Z}_+\} \end{cases} \tag{EC.46}$$

and

$$0 \leq 1 - \alpha_h \leq h(1 + \frac{\sqrt{d_a}}{2}) \wedge 1, \tag{EC.47}$$

where $\boldsymbol{b}_{t+1} = \sum_{\ell=1}^{L}\boldsymbol{x}_{t+1,\ell}$.

The definition of $\bar{\boldsymbol{\delta}}_{t+1}$ given in Equation (EC.44) directly gives

$$\mathbb{E}(\bar{\boldsymbol{\delta}}_{t+1} \mid \{\boldsymbol{x}_{i,\ell}\}_{j\leq t+1,\ell\leq L}, \{\boldsymbol{a}_i\}_{i\leq t}, \{y_{i,\ell}\}_{i\leq t,\ell\leq L}) = \mathbf{0}. \tag{EC.48}$$

The definition of $\bar{\boldsymbol{\delta}}_t$ given in Equation (EC.44) and Equation (EC.45), together with the expression of $\bar{\boldsymbol{a}}_{t+1}$ in Equation (EC.46) also gives

$$\bar{\boldsymbol{\delta}}_{t+1} = \begin{cases} 0, & \text{for } t \notin \{\lfloor w^{\frac{3}{2}} \rfloor : w \in \mathbb{Z}_+\} \\ \left[\mathcal{P}_{\mathcal{A}_t}\left(\frac{\hat{\boldsymbol{\Theta}}_t \boldsymbol{b}_{t+1}}{\|\hat{\boldsymbol{\Theta}}_t \boldsymbol{b}_{t+1}\|_2} + \boldsymbol{\delta}_{t+1}\right) - \frac{\hat{\boldsymbol{\Theta}}_t \boldsymbol{b}_{t+1}}{\|\hat{\boldsymbol{\Theta}}_t \boldsymbol{b}_{t+1}\|_2}\right] + (1-\alpha_h)\frac{\hat{\boldsymbol{\Theta}}_t \boldsymbol{b}_{t+1}}{\|\hat{\boldsymbol{\Theta}}_t \boldsymbol{b}_{t+1}\|_2}, & \text{for } t \in \{\lfloor w^{\frac{3}{2}} \rfloor : w \in \mathbb{Z}_+\} \end{cases}, \tag{EC.49}$$

where $\mathcal{A}_t$ in our case is a unit ball in $\mathbb{R}^{d_a}$.

Note that projection is a contraction mapping, and we have bounds for α_h in Equation (EC.47). Hence we have

$$\|\bar{\boldsymbol{\delta}}_{t+1}\|_2 \le \begin{cases} 0, & \text{for } t \notin \{\lfloor w^{\frac{3}{2}} \rfloor : w \in \mathbb{Z}_+\} \\ \min\{2, \|\boldsymbol{\delta}_{t+1}\| + h(1+\frac{\sqrt{d_a}}{2})\} & \text{for } t \in \{\lfloor w^{\frac{3}{2}} \rfloor : w \in \mathbb{Z}_+\} \end{cases}. \tag{EC.50}$$

For notational simplicity, let $\boldsymbol{\delta}_t = \mathbf{0}$ for exploitation rounds, and let

$$\alpha_{t,h} = \begin{cases} 1 & \text{for } t-1 \notin \{\lfloor w^{\frac{3}{2}} \rfloor : w \in \mathbb{Z}_+\} \\ \alpha_h & \text{for } t-1 \in \{\lfloor w^{\frac{3}{2}} \rfloor : w \in \mathbb{Z}_+\} \end{cases}. \tag{EC.51}$$

Then we have the decomposition

$$e_T(\boldsymbol{\Delta}) = \frac{1}{2LT}\sum_{t=1}^{T}\sum_{\ell=1}^{L}(\boldsymbol{a}_t^\top \boldsymbol{\Delta}\boldsymbol{x}_{t,\ell})^2 \;=\; \frac{1}{2LT}\sum_{t=1}^{T}\sum_{\ell=1}^{L}\left\{\left(\alpha_{t,h}\frac{\boldsymbol{b}_t^\top \widehat{\boldsymbol{\Theta}}_{t-1}^\top}{\|\boldsymbol{b}_t^\top \widehat{\boldsymbol{\Theta}}_{t-1}^\top\|_2} + \bar{\boldsymbol{\delta}}_t^\top\right)\boldsymbol{\Delta}\boldsymbol{x}_{t,\ell}\right\}^2,$$

where $\widehat{\boldsymbol{\Theta}}_0$ is interpreted as randomly selected non-zero $d_a \times d_x$ matrix.

This decomposition is well-defined and valid with probability one as we will show that $\widehat{\boldsymbol{\Theta}}_t$ is non-zero with probability one. By definition of $\widehat{\boldsymbol{\Theta}}_t$, we have that $\mathbf{0}_{d_a \times d_x} \in \nabla \mathcal{L}_t(\widehat{\boldsymbol{\Theta}}_t) + \partial\|\!|\cdot|\!\|_{\text{nuc}}(\widehat{\boldsymbol{\Theta}}_t)$. If $\widehat{\boldsymbol{\Theta}}_t = \mathbf{0}_{d_a \times d_x}$, the following equation needs to hold

$$\frac{1}{tL}\sum_{i=1}^{t}\sum_{\ell=1}^{L}\varepsilon_{i,\ell}\boldsymbol{a}_i\boldsymbol{x}_{i,\ell}^\top \in \lambda_t \partial\|\!|\cdot|\!\|_{\text{nuc}}(\boldsymbol{\Theta}^*).$$

Consider the singular value decomposition $\boldsymbol{\Theta}^* = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$, where $\boldsymbol{S}$ is an $r \times r$ diganal matrix. Then by the expression of subgradient of nuclear norm, the following equation needs to hold

$$\boldsymbol{U}^\top \frac{1}{tL}\sum_{i=1}^{t}\sum_{\ell=1}^{L}\varepsilon_{i,\ell}\boldsymbol{a}_i\boldsymbol{x}_{i,\ell}^\top \boldsymbol{V} = \mathbf{I}_r.$$

However, it only holds with probability 0.

Introduce the shorthand notation

$$\rho_T(\boldsymbol{\Delta}) := \frac{1}{2LT}\sum_{t=1}^{T}\sum_{\ell=1}^{L}\left((\alpha_{t,h}\frac{\boldsymbol{b}_t^\top \widehat{\boldsymbol{\Theta}}_{t-1}^\top}{\|\boldsymbol{b}_t^\top \widehat{\boldsymbol{\Theta}}_{t-1}^\top\|_2}\boldsymbol{\Delta}\boldsymbol{x}_{t,\ell})^2 + (\bar{\boldsymbol{\delta}}_t^\top \boldsymbol{\Delta}\boldsymbol{x}_{t,\ell})^2\right) \tag{EC.52}$$

$$\rho_{1,T}(\boldsymbol{\Delta}) := \frac{1}{2LT}\sum_{t=1}^{T}\sum_{\ell=1}^{L}(\alpha_{t,h}\frac{\boldsymbol{b}_t^\top \widehat{\boldsymbol{\Theta}}_{t-1}^\top}{\|\boldsymbol{b}_t^\top \widehat{\boldsymbol{\Theta}}_{t-1}^\top\|_2}\boldsymbol{\Delta}\boldsymbol{x}_{t,\ell})^2, \quad \text{and} \tag{EC.53}$$

$$\rho_{2,T}(\boldsymbol{\Delta}) := \frac{1}{2LT}\sum_{t=1}^{T}\sum_{\ell=1}^{L}(\bar{\boldsymbol{\delta}}_t^\top \boldsymbol{\Delta}\boldsymbol{x}_{t,\ell})^2 \tag{EC.54}$$

$$\mathbb{M}(\rho_{2,T}(\boldsymbol{\Delta})) := \frac{1}{2LT}\sum_{t=1}^{T}\sum_{\ell=1}^{L}\mathbb{E}\left((\bar{\boldsymbol{\delta}}_t^\top \boldsymbol{\Delta}\boldsymbol{x}_{t,\ell})^2 \middle| \{\boldsymbol{a}_{\tau,j}, \boldsymbol{x}_{\tau,j}, y_{\tau,j}\}_{\tau \le t-1, j \le L}\right). \tag{EC.55}$$

With these definitions, we have the relations

$$e_T(\boldsymbol{\Delta}) - \rho_T(\boldsymbol{\Delta}) = \frac{1}{LT}\sum_{t=1}^{T}\sum_{\ell=1}^{L}(\alpha_{t,h}\frac{\boldsymbol{b}_t^\top \widehat{\boldsymbol{\Theta}}_{t-1}^\top}{\|\boldsymbol{b}_t^\top \widehat{\boldsymbol{\Theta}}_{t-1}^\top\|_2}\boldsymbol{\Delta}\boldsymbol{x}_{t,\ell})(\bar{\boldsymbol{\delta}}_t^\top \boldsymbol{\Delta}\boldsymbol{x}_{t,\ell}).$$

Next, we introduce a lemma that we will prove later.

LEMMA EC.2. *For any $d_a \times d_x$ matrix $\mathbf{\Delta}$,*

$$\mathbb{M}(\rho_{2,T}(\mathbf{\Delta})) \geq \frac{\lfloor T^{2/3} \rfloor \tilde{c}_6 \left(\frac{1}{d_a - 1} \wedge h^2 \wedge (d_a - 1) h^4 \mathbb{1}\{h \geq \frac{1}{4(d_a-1)}\} \right)}{2T} |||\mathbf{\Delta}|||_F^2, \tag{EC.56}$$

where $\tilde{c}_6 > 0$ is an absolute constant.

Clearly, we have

$$\begin{aligned} e(\mathbf{\Delta}) \geq & \frac{\lfloor T^{2/3} \rfloor \tilde{c}_6 \left(\frac{1}{d_a - 1} \wedge h^2 \wedge (d_a - 1) h^4 \mathbb{1}\{h \geq \frac{1}{4(d_a-1)}\} \right)}{2T} |||\mathbf{\Delta}|||_{\mathrm{F}}^2 \\ & + \frac{e_T(\mathbf{\Delta}) - \rho_T(\mathbf{\Delta})}{|||\mathbf{\Delta}|||_{\mathrm{F}}^2} |||\mathbf{\Delta}|||_{\mathrm{F}}^2 + \frac{\rho_{2,T}(\mathbf{\Delta}) - \mathbb{M}(\rho_{2,T}(\mathbf{\Delta}))}{|||\mathbf{\Delta}|||_{\mathrm{F}}^2} |||\mathbf{\Delta}|||_{\mathrm{F}}^2. \end{aligned} \tag{EC.57}$$

Next, we prove that the following two bounds hold with high probability:

$$\begin{aligned} & \inf_{|||\mathbf{\Delta}|||_{\mathrm{F}} > 0} \frac{e_T(\mathbf{\Delta}) - \rho_T(\mathbf{\Delta})}{|||\mathbf{\Delta}|||_{\mathrm{F}}^2} \geq \\ & \quad - \frac{6}{T^{2/3}} \alpha_h (d_x + 3\log(TL)) \left(2 \wedge \frac{3h}{\sqrt{2}} \sqrt{d_a + 3\log(T)} \right) \sqrt{\log T + d_a d_x (\frac{1}{3}\log(T) + \log 3)} \end{aligned} \tag{EC.58a}$$

$$\begin{aligned} & \inf_{|||\mathbf{\Delta}|||_{\mathrm{F}} > 0} \frac{\rho_{2,T}(\mathbf{\Delta}) - \mathbb{M}(\rho_{2,T}(\mathbf{\Delta}))}{|||\mathbf{\Delta}|||_{\mathrm{F}}^2} \geq \\ & \quad - \frac{1}{T^{2/3}} (2d_x + 6\log(TL) + d_x \log\log(TL))(2 + \sqrt{3 d_a d_x \log(T)}) \left(4 \wedge \frac{9}{2} h^2 (d_a + 3\log(T) + \frac{1}{2} d_a \log\log T) \right). \end{aligned} \tag{EC.58b}$$

Before proceeding to proving Equation (EC.58a) and Equation (EC.58b), we define an event and introduce some truncated random variables.

Note that the random variable $\|\boldsymbol{x}_{t,\ell}\|_2^2$ follows a $\chi^2_{d_x}$-distribution, whereas $\|\boldsymbol{\delta}_t/h\|_2^2$ follows a $\chi^2_{d_a}$-distribution. Therefore, by combining standard χ^2-tail bounds (Lemma 1 in Laurent and Massart (2000)) with the union bound, for any choice of $\epsilon_1, \epsilon_2 > 0$, we have

$$\max_{\substack{t \in [T] \\ \ell \in [L]}} \|\boldsymbol{x}_{t,\ell}\|_2^2 \leq d_x + 2\epsilon_1 + 2\sqrt{\epsilon_1 d_x}, \quad \text{and} \quad \max_{t \in [T]} \|\boldsymbol{\delta}_t/h\|_2^2 \leq d_a + 2\epsilon_2 + 2\sqrt{\epsilon_2 d_a} \tag{EC.59}$$

with probability at least $1 - (LT\exp(-\epsilon_1) + T\exp(-\epsilon_2))$.

Now we set $\epsilon_1 = 2\log(LT)$ $\epsilon_2 = 2\log(T)$, and we introduce the shorthand

$$\begin{aligned} & U_1 := d_x + 2\epsilon_1 + 2\sqrt{\epsilon_1 d_x} \quad \text{and} \quad U_2 := d_a + 2\epsilon_2 + 2\sqrt{\epsilon_2 d_a}, \quad \text{and the event} \\ & \mathcal{U} := \Big\{ \max_{\substack{t \in [T] \\ \ell \in [L]}} \|\boldsymbol{x}_{t,\ell}\|_2^2 \leq U_1 \quad \text{and} \quad \max_{t \in [T]} \|\boldsymbol{\delta}_t/h\|_2^2 \leq U_2 \Big\}. \end{aligned}$$

Clearly, $\mathbb{P}(\mathcal{U}) \geq 1 - \frac{1}{LT} - \frac{1}{T}$.

Now we introduce the truncated variables

$$\tilde{\mathbf{x}}_{t,\ell} = \boldsymbol{x}_{t,\ell} \mathbb{1}\{\boldsymbol{x}_{t,\ell} \leq U_1\}, \quad \tilde{\boldsymbol{\delta}}_t = \bar{\boldsymbol{\delta}}_t \mathbb{1}\{\|\boldsymbol{\delta}_t/h\|_2^2 \leq U_2\}.$$

Using the truncated variables, we define the truncated version of the quantities defined above.

$$\tilde{e}_T(\mathbf{\Delta}) := \frac{1}{2LT} \sum_{t=1}^{T} \sum_{\ell=1}^{L} \left\{ \left(\alpha_{t,h} \frac{\boldsymbol{b}_t^\top \widehat{\mathbf{\Theta}}_{t-1}^\top}{\|\boldsymbol{b}_t^\top \widehat{\mathbf{\Theta}}_{t-1}^\top\|_2} + \tilde{\boldsymbol{\delta}}_t^\top \right) \mathbf{\Delta} \tilde{\mathbf{x}}_{t,\ell} \right\}^2, \quad \text{and}$$

$$\tilde{\rho}_T(\mathbf{\Delta}) := \frac{1}{2LT}\sum_{t=1}^{T}\sum_{\ell=1}^{L}\left((\alpha_{t,h}\frac{\boldsymbol{b}_t^\top\widehat{\mathbf{\Theta}}_{t-1}^\top}{\|\boldsymbol{b}_t^\top\widehat{\mathbf{\Theta}}_{t-1}^\top\|_2}\mathbf{\Delta}\tilde{\mathbf{x}}_{t,\ell})^2+(\tilde{\boldsymbol{\delta}}_t^\top\mathbf{\Delta}\tilde{\mathbf{x}}_{t,\ell})^2\right).$$

Clearly, on event $\mathcal{U}$, we have that

$$\tilde{e}_T(\mathbf{\Delta}) = e_T(\mathbf{\Delta}), \qquad \tilde{\rho}_T(\mathbf{\Delta}) = \rho_T(\mathbf{\Delta}).$$

Introduce the shorthand:

$$\bar{U}_2 = \left(\min\{2, h(1+\frac{\sqrt{d_a}}{2})+h\sqrt{U_2}\}\right)^2. \tag{EC.60}$$

Through Inequality (EC.50), we have that

$$\|\tilde{\boldsymbol{\delta}}_t\| \le \sqrt{\bar{U}_2}. \tag{EC.61}$$

EC.2.4.1. Proof of the bound (EC.58a): We introduce the shorthand notation

$$\mathfrak{D}_1(T;\mathbf{\Delta}) := \frac{e_T(\mathbf{\Delta})-\rho_T(\mathbf{\Delta})}{|||\mathbf{\Delta}|||_{\mathrm{F}}}, \quad \text{and} \quad \tilde{\mathfrak{D}}_1(T;\mathbf{\Delta}) := \frac{\tilde{e}_T(\mathbf{\Delta})-\tilde{\rho}_T(\mathbf{\Delta})}{|||\mathbf{\Delta}|||_{\mathrm{F}}}$$

Then conditioned on the event $\mathcal{U}$, we have

$$\inf_{|||\mathbf{\Delta}|||_{\mathrm{F}}>0}\mathfrak{D}_1(T;\mathbf{\Delta}) = \inf_{|||\mathbf{\Delta}|||_{\mathrm{F}}>0}\tilde{\mathfrak{D}}_1(T;\mathbf{\Delta}).$$

Next we focus on bounding $\inf_{|||\mathbf{\Delta}|||_{\mathrm{F}}>0}\tilde{\mathfrak{D}}_1(T;\mathbf{\Delta})$. Elementary calculation gives

$$\tilde{\mathfrak{D}}_1(T;\mathbf{\Delta}) = \frac{1}{T}\sum_{t=1}^{T}\left\{\frac{1}{L}\sum_{\ell=1}^{L}(\alpha_{t,h}\frac{\boldsymbol{b}_t^\top\widehat{\mathbf{\Theta}}_{t-1}^\top}{\|\boldsymbol{b}_t^\top\widehat{\mathbf{\Theta}}_{t-1}^\top\|_2}\frac{\mathbf{\Delta}}{|||\mathbf{\Delta}|||_{\mathrm{F}}}\tilde{\mathbf{x}}_{t,\ell})(\tilde{\boldsymbol{\delta}}_t^\top\frac{\mathbf{\Delta}}{|||\mathbf{\Delta}|||_{\mathrm{F}}}\tilde{\mathbf{x}}_{t,\ell})\right\}.$$

Clearly, we have the boundedness of each term

$$\left|\frac{1}{L}\sum_{\ell=1}^{L}(\alpha_{t,h}\frac{\boldsymbol{b}_t^\top\widehat{\mathbf{\Theta}}_{t-1}^\top}{\|\boldsymbol{b}_t^\top\widehat{\mathbf{\Theta}}_{t-1}^\top\|_2}\frac{\mathbf{\Delta}}{|||\mathbf{\Delta}|||_{\mathrm{F}}}\tilde{\mathbf{x}}_{t,\ell})(\tilde{\boldsymbol{\delta}}_t^\top\frac{\mathbf{\Delta}}{|||\mathbf{\Delta}|||_{\mathrm{F}}}\tilde{\mathbf{x}}_{t,\ell})\right| \le \begin{cases}0, & \text{if } t-1\notin\{\lfloor w^{\frac{3}{2}}\rfloor : w\in\mathbb{Z}_+\}\\ \alpha_{t,h}U_1\sqrt{\bar{U}_2}, & \text{if } t-1\in\{\lfloor w^{\frac{3}{2}}\rfloor : w\in\mathbb{Z}_+\}\end{cases}, \tag{EC.62}$$

and the martingale property of the series of the sum

$$\mathbb{E}\left(\frac{1}{L}\sum_{\ell=1}^{L}(\alpha_{t,h}\frac{\boldsymbol{b}_t^\top\widehat{\mathbf{\Theta}}_{t-1}^\top}{\|\boldsymbol{b}_t^\top\widehat{\mathbf{\Theta}}_{t-1}^\top\|_2}\frac{\mathbf{\Delta}}{|||\mathbf{\Delta}|||_{\mathrm{F}}}\tilde{\mathbf{x}}_{t,\ell})(\tilde{\boldsymbol{\delta}}_t^\top\frac{\mathbf{\Delta}}{|||\mathbf{\Delta}|||_{\mathrm{F}}}\tilde{\mathbf{x}}_{t,\ell})\middle|\{\boldsymbol{x}_{\tau,j}\}_{\tau\le t-1,j\le L},\{\boldsymbol{a}_{\tau,j}\}_{\tau\le t-1,j\le L}\{y_{\tau,j}\}_{\tau\le t-1,j\le L}\right)=0$$

$$\mathbb{E}\left(\tilde{\mathfrak{D}}_1(t-1;\mathbf{\Delta})\middle|\{\boldsymbol{x}_{\tau,j}\}_{\tau\le t-1,j\le L},\{\boldsymbol{a}_{\tau,j}\}_{\tau\le t-1,j\le L}\{y_{\tau,j}\}_{\tau\le t-1,j\le L}\right)=\tilde{\mathfrak{D}}_1(t-1;\mathbf{\Delta}). \tag{EC.63}$$

Using Azuma-Hoeffding's inequality, we have for $\nu>0$,

$$\mathbb{P}\left(\tilde{\mathfrak{D}}_1(T;\mathbf{\Delta})\le-\nu\right)\le\exp\{-\frac{\nu^2}{2\lfloor T^{\frac{2}{3}}\rfloor\alpha_{t,h}^2U_1^2\bar{U}_2/T^2}\}. \tag{EC.64}$$

Let $\nu=\sqrt{2}T^{-\frac{2}{3}}\alpha_hU_1\sqrt{\bar{U}_2}\sqrt{\log T+d_ad_x(\frac{1}{3}\log(T)+\log 3)}$, we have that

$$\mathbb{P}\left(\tilde{\mathfrak{D}}_1(T;\mathbf{\Delta})\le-\nu\right)\le\frac{1}{T}\left(\frac{1}{3T^{1/3}}\right)^{d_ad_x}. \tag{EC.65}$$

Next we take an η-covering of the set $\mathbb{B}:=\{\mathbf{\Delta}\mid|||\mathbf{\Delta}|||_{\mathrm{F}}=1\}$: $\mathbb{F}=\{\tilde{\mathbf{\Delta}}_1,\tilde{\mathbf{\Delta}}_2,\ldots,\tilde{\mathbf{\Delta}}_{\mathbf{N}_\eta}\}$. By a standard covering number calculation (e.g. see Example 5.8 in Wainwright (2019)), $\log\mathbf{N}_\eta\le d_xd_a\log(1+\frac{2}{\eta})$). We have

$$\inf_{|||\mathbf{\Delta}|||_{\mathrm{F}}=1}\tilde{\mathfrak{D}}_1(T;\mathbf{\Delta})\ge\underbrace{\inf_{\mathbf{\Delta}\in\mathbb{F}}\tilde{\mathfrak{D}}_1(T;\mathbf{\Delta})}_{W_1}-\underbrace{\sup_{\substack{|||\mathbf{\Delta}_1-\mathbf{\Delta}_2|||_{\mathrm{F}}\le\eta,\\ |||\mathbf{\Delta}_1|||_{\mathrm{F}}=1,\,|||\mathbf{\Delta}_2|||_{\mathrm{F}}=1}}|\tilde{\mathfrak{D}}_1(T;\mathbf{\Delta}_1)-\tilde{\mathfrak{D}}_1(T;\mathbf{\Delta}_2)|}_{W_2}. \tag{EC.66}$$

Next we analyze the quantities W_2 and W_1 separately with $\eta=T^{-1/3}$.

Analysis of W_2:

$$
\begin{aligned}
&W_2\\
&\le \sup_{\substack{|||\boldsymbol{\Delta}_1|||_{\mathrm{F}}=1,|||\boldsymbol{\Delta}|||_{\mathrm{F}}\le\eta\\ |||\boldsymbol{\Delta}_1+\boldsymbol{\Delta}|||_{\mathrm{F}}=1}} |\tilde{\mathfrak{D}}_1(T;\boldsymbol{\Delta}_1+\boldsymbol{\Delta})-\tilde{\mathfrak{D}}_1(T;\boldsymbol{\Delta}_1)|\\
&\le \sup_{\substack{|||\boldsymbol{\Delta}_1|||_{\mathrm{F}}=1,|||\boldsymbol{\Delta}|||_{\mathrm{F}}\le\eta\\ |||\boldsymbol{\Delta}_1+\boldsymbol{\Delta}|||_{\mathrm{F}}=1}} \left|\frac{1}{LT}\sum_{t=1}^{T}\sum_{\ell=1}^{L}\left\{\left(\alpha_{t,h}\tfrac{\boldsymbol{b}_t^\top\widehat{\boldsymbol{\Theta}}_{t-1}^\top}{\|\boldsymbol{b}_t^\top\widehat{\boldsymbol{\Theta}}_{t-1}^\top\|_2}\boldsymbol{\Delta}\tilde{\mathbf{x}}_{t,\ell}\right)\left(\tilde{\boldsymbol{\delta}}_t^\top\boldsymbol{\Delta}_1\tilde{\mathbf{x}}_{t,\ell}\right)+\left(\alpha_{t,h}\tfrac{\boldsymbol{b}_t^\top\widehat{\boldsymbol{\Theta}}_{t-1}^\top}{\|\boldsymbol{b}_t^\top\widehat{\boldsymbol{\Theta}}_{t-1}^\top\|_2}(\boldsymbol{\Delta}+\boldsymbol{\Delta}_1)\tilde{\mathbf{x}}_{t,\ell}\right)\left(\tilde{\boldsymbol{\delta}}_t^\top\boldsymbol{\Delta}\tilde{\mathbf{x}}_{t,\ell}\right)\right\}\right|\\
&\le \frac{\eta}{T^{1/3}}\cdot 2\alpha_h U_1\sqrt{\bar{U}_2}=\frac{2}{T^{2/3}}\alpha_h U_1\sqrt{\bar{U}_2}.
\end{aligned}
$$

Analysis of W_1: By Inequality (EC.65), we have

$$\mathbb{P}(W_1\le-\nu)\le\frac{1}{T}\exp\left(\mathbf{N}_\eta-d_ad_x(\frac{1}{3}\log(T)+\log 3)\right)\le\frac{1}{T}. \tag{EC.67}$$

Combining the analysis of W_1 and W_2, we see that, with probability at most $\frac{1}{T}$,

$$
\begin{aligned}
\inf_{|||\boldsymbol{\Delta}|||_{\mathrm{F}}=1}\tilde{\mathfrak{D}}_1(T;\boldsymbol{\Delta}) &\le -\nu-\frac{2}{T^{2/3}}\alpha_h U_1\sqrt{\bar{U}_2}\\
&\le -\sqrt{2}T^{-\frac{2}{3}}\alpha_h U_1\sqrt{\bar{U}_2}\sqrt{\log T+d_ad_x(\frac{1}{3}\log(T)+\log 3)}-\frac{2}{T^{2/3}}\alpha_h U_1\sqrt{\bar{U}_2}\\
&\le \underbrace{-3T^{-\frac{2}{3}}\alpha_h U_1\sqrt{\bar{U}_2}\sqrt{\log T+d_ad_x(\frac{1}{3}\log(T)+\log 3)}}_{\xi}.
\end{aligned}
$$

Note that we have

$$U_1\le 2d_x+6\log T+6\log L, \tag{EC.68}$$

$$\sqrt{\bar{U}_2}=\min\{2,h(1+\frac{\sqrt{d_a}}{2})+h\sqrt{U_2}\}\le\min\{2,h(1+\frac{\sqrt{d_a}}{2}+\sqrt{2d_a+6\log T})\} \tag{EC.69}$$

$$\le\min\{2,h\frac{3}{\sqrt{2}}\sqrt{d_a+3\log T}\}\text{ for }T\ge 4. \tag{EC.70}$$

Therefore,

$$\xi\ge\underbrace{-\frac{6}{T^{2/3}}\alpha_h(d_x+3\log(TL))\left(2\wedge\frac{3h}{\sqrt{2}}\sqrt{d_a+3\log(T)}\right)\sqrt{\log T+d_ad_x(\frac{1}{3}\log(T)+\log 3)}}_{\kappa_1}. \tag{EC.71}$$

Hence for $T\ge 4$ we have

$$\mathbb{P}\left(\left\{\inf_{|||\boldsymbol{\Delta}|||_{\mathrm{F}}>0}\mathfrak{D}_1(T;\boldsymbol{\Delta})<\kappa_1\right\}\cap\mathcal{U}\right)<\frac{1}{T}. \tag{EC.72}$$

EC.2.4.2. Proof of the bound (EC.58b): We first introduce a slightly different form of truncation. Define the truncation thresholds

$$\check{U}_1:=2d_x+6\log(TL)+d_x\log\log(TL)\quad\text{and}\quad\check{U}_2:=2d_a+6\log(T)+d_a\log\log(T). \tag{EC.73}$$

Using these truncation levels, we define the "truncation event" as

$$\check{\mathcal{U}}:=\Big\{\max_{\substack{t\in[T]\\ \ell\in[L]}}\|\boldsymbol{x}_{t,\ell}\|_2^2\le\check{U}_1\quad\text{and}\quad\max_{t\in[T]}\|\boldsymbol{\delta}_t/h\|_2^2\le\check{U}_2\Big\}.$$

Clearly, this newly defined "truncation" event $\check{\mathcal{U}}\subset\mathcal{U}$.

We introduce the shorthand for truncated variables associated with $\check{\mathcal{U}}$:

$$\check{\mathbf{x}}_{t,\ell} = \boldsymbol{x}_{t,\ell}\mathbb{1}\{\|\check{\mathbf{x}}_{t,\ell}\|_2^2 \le \check{U}_1\}, \qquad \check{\boldsymbol{\delta}}_t = \bar{\boldsymbol{\delta}}_t\mathbb{1}\{\|\boldsymbol{\delta}_t/h\|_2^2 \le \check{U}_2\}.$$

With the truncated variables, we introduce the shorthand for the quantity to bound and its truncated version.

$$\check{\rho}_{2,T}(\boldsymbol{\Delta}) = \frac{1}{2LT}\sum_{t=1}^{T}\sum_{\ell=1}^{L}\big(\check{\boldsymbol{\delta}}_t^\top\boldsymbol{\Delta}\check{\mathbf{x}}_{t,\ell}\big)^2, \tag{EC.74a}$$

$$\mathbb{M}(\check{\rho}_{2,T}(\boldsymbol{\Delta})) = \frac{1}{2LT}\sum_{t=1}^{T}\sum_{\ell=1}^{L}\big(\check{\boldsymbol{\delta}}_t^\top\boldsymbol{\Delta}\check{\mathbf{x}}_{t,\ell}\Big|\{\boldsymbol{x}_{\tau,j},\boldsymbol{a}_\tau,y_{\tau,j}\}_{\tau\le t-1,j\le L}\big)^2, \tag{EC.74b}$$

$$\mathfrak{D}_2(T;\boldsymbol{\Delta}) = \frac{\rho_{2,T}(\boldsymbol{\Delta}) - \mathbb{M}(\rho_{2,T}(\boldsymbol{\Delta}))}{\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2}, \qquad \check{\mathfrak{D}}_2(T;\boldsymbol{\Delta}) = \frac{\check{\rho}_{2,T}(\boldsymbol{\Delta}) - \mathbb{M}(\check{\rho}_{2,T}(\boldsymbol{\Delta}))}{\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2}. \tag{EC.74c}$$

Elementary calculation shows that

$$\begin{aligned}\inf_{\|\boldsymbol{\Delta}\|_{\mathrm{F}}>0}\mathfrak{D}_2(T;\boldsymbol{\Delta}) &= \inf_{\|\boldsymbol{\Delta}\|_{\mathrm{F}}=1}\mathfrak{D}_2(T;\boldsymbol{\Delta})\\ &\ge \underbrace{\inf_{\|\boldsymbol{\Delta}\|_{\mathrm{F}}=1}\check{\mathfrak{D}}_2(T;\boldsymbol{\Delta})}_{Z_1(T)} - \underbrace{\sup_{\|\boldsymbol{\Delta}\|_{\mathrm{F}}=1}\left|\mathfrak{D}_2(T;\boldsymbol{\Delta}) - \check{\mathfrak{D}}_2(T;\boldsymbol{\Delta})\right|}_{Z_2(T)}.\end{aligned} \tag{EC.75}$$

Therefore, we only need to bound $Z_2(T)$ and $Z_1(T)$.

Before proceeding, introduce the shorthand:

$$\tilde{U}_2 := \left(\min\{2, h(1+\frac{\sqrt{d_a}}{2}) + h\sqrt{\check{U}_2}\}\right)^2. \tag{EC.76}$$

Through Inequality (EC.50), we have that

$$\|\check{\boldsymbol{\delta}}_t\| \le \begin{cases}\sqrt{\tilde{U}_2}, & \text{for } t-1 \in \{\lfloor w^{\frac{3}{2}}\rfloor : w\in\mathbb{Z}_+\}\\ 0, & \text{for } t-1 \notin \{\lfloor w^{\frac{3}{2}}\rfloor : w\in\mathbb{Z}_+\}\end{cases}. \tag{EC.77}$$

We denote the set of exploration rounds as $\mathfrak{S} := \{\lfloor w^{3/2}\rfloor + 1 : w\in\mathbb{Z}_+\}$.

Analysis of $Z_1(T)$: Rewrite $Z_1(T)$ as

$$Z_1(T) = \inf_{\|\boldsymbol{\Delta}\|_{\mathrm{F}}=1}\frac{1}{2LT}\sum_{t\in\mathfrak{S}\cap[T]}\sum_{\ell=1}^{L}\left\{\big(\check{\boldsymbol{\delta}}_t^\top\boldsymbol{\Delta}\check{\mathbf{x}}_{t,\ell}\big)^2 - \mathbb{E}\Big(\big(\check{\boldsymbol{\delta}}_t^\top\boldsymbol{\Delta}\check{\mathbf{x}}_{t,\ell}\big)^2\Big|\{\boldsymbol{x}_{\tau,j},\boldsymbol{a}_\tau,y_{\tau,j}\}_{\tau\le t-1,j\le L}\Big)\right\}.$$

We introduce the following shorthand for $\boldsymbol{\Delta}$ with $\|\boldsymbol{\Delta}\|_{\mathrm{F}} = 1$ and $T\ge 3$:

$$S(T;\boldsymbol{\Delta}) := \frac{1}{2LT}\sum_{t\in\mathfrak{S}\cap[T]}\sum_{\ell=1}^{L}\left\{\big(\check{\boldsymbol{\delta}}_t^\top\boldsymbol{\Delta}\check{\mathbf{x}}_{t,\ell}\big)^2 - \mathbb{E}\Big(\big(\check{\boldsymbol{\delta}}_t^\top\boldsymbol{\Delta}\check{\mathbf{x}}_{t,\ell}\big)^2\Big|\{\boldsymbol{x}_{\tau,j},\boldsymbol{a}_\tau,y_{\tau,j}\}_{\tau\le t-1,j\le L}\Big)\right\}. \tag{EC.78}$$

Again, we take an η-covering of the set $\mathbb{B} := \{\boldsymbol{\Delta} \mid \|\boldsymbol{\Delta}\|_{\mathrm{F}} = 1\}$: $\mathbb{F} = \{\tilde{\boldsymbol{\Delta}}_1, \tilde{\boldsymbol{\Delta}}_2, \ldots, \tilde{\boldsymbol{\Delta}}_{\mathbf{N}_\eta}\}$. By a standard covering number calculation (e.g. see Example 5.8 in Wainwright (2019)), $\log\mathbf{N}_\eta \le d_x d_a\log(1+\frac{2}{\eta})$). We have

$$Z_1(T) = \inf_{\|\boldsymbol{\Delta}\|_{\mathrm{F}}=1}S(T;\boldsymbol{\Delta}) \ge \underbrace{\inf_{\boldsymbol{\Delta}\in\mathbb{F}}S(T;\boldsymbol{\Delta})}_{\tilde{W}_1} - \underbrace{\sup_{\substack{\|\boldsymbol{\Delta}_1-\boldsymbol{\Delta}_2\|_{\mathrm{F}}\le\eta,\\ \|\boldsymbol{\Delta}_1\|_{\mathrm{F}}=1,\|\boldsymbol{\Delta}_2\|_{\mathrm{F}}=1}}\left|S(T;\boldsymbol{\Delta}_1) - S(T;\boldsymbol{\Delta}_2)\right|}_{\tilde{W}_2}. \tag{EC.79}$$

Next, we analyze $\tilde{W}_2$ and $\tilde{W}_1$ in turn with $\eta = T^{-\frac{1}{3}}$.

We start with $\tilde{W}_2$.

Note that for $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_2$ satisfying $|||\boldsymbol{\Delta}_1 - \boldsymbol{\Delta}_2|||_{\mathrm{F}} \le \eta$, $|||\boldsymbol{\Delta}_1|||_{\mathrm{F}} = 1$, and $|||\boldsymbol{\Delta}_2|||_{\mathrm{F}} = 1$, we have

$$\begin{aligned}|S(T;\boldsymbol{\Delta}_1) - S(T;\boldsymbol{\Delta}_2)| &\le \frac{1}{2LT}\sum_{t\in\mathfrak{S}\cap[T]}\sum_{\ell=1}^{L}\Big\{\Big|\big(\check{\boldsymbol{\delta}}_t^\top(\boldsymbol{\Delta}_1-\boldsymbol{\Delta}_2)\check{\mathbf{x}}_{t,\ell}\big)\big(\check{\boldsymbol{\delta}}_t^\top(\boldsymbol{\Delta}_1+\boldsymbol{\Delta}_2)\check{\mathbf{x}}_{t,\ell}\big)\\ &\quad - \mathbb{E}\Big(\big(\check{\boldsymbol{\delta}}_t^\top(\boldsymbol{\Delta}_1-\boldsymbol{\Delta}_2)\check{\mathbf{x}}_{t,\ell}\big)\big(\check{\boldsymbol{\delta}}_t^\top(\boldsymbol{\Delta}_1+\boldsymbol{\Delta}_2)\check{\mathbf{x}}_{t,\ell}\big)\Big|\{\boldsymbol{x}_{\tau,j},\boldsymbol{a}_\tau,y_{\tau,j}\}_{\tau\le t-1,j\le L}\Big)\Big|\Big\}\\ &\le \frac{\lfloor T^{\frac{2}{3}}\rfloor}{2LT}\times 4L\eta\check{U}_1\tilde{U}_2 \le \frac{2\check{U}_1\tilde{U}_2}{T^{\frac{2}{3}}}.\end{aligned} \tag{EC.80}$$

Therefore, we have $\tilde{W}_2 \le \frac{2\check{U}_1\tilde{U}_2}{T^{\frac{2}{3}}}$.

Now we turn to $\tilde{W}_1$. The key observation is that $S(T;\boldsymbol{\Delta})$ is associated with a martingale difference sequence, which we will check below.

Clearly, for any $\boldsymbol{\Delta}\in\mathbb{F}$, by the definition of truncated variables $\check{\boldsymbol{\delta}}_t$ and $\check{\mathbf{x}}_{t,\ell}$, we have

$$\Big|\big(\check{\boldsymbol{\delta}}_t^\top\boldsymbol{\Delta}\check{\mathbf{x}}_{t,\ell}\big)^2 - \mathbb{E}\Big(\big(\check{\boldsymbol{\delta}}_t^\top\boldsymbol{\Delta}\check{\mathbf{x}}_{t,\ell}\big)^2\Big|\{\boldsymbol{x}_{\tau,j},\boldsymbol{a}_\tau,y_{\tau,j}\}_{\tau\le t-1,j\le L}\Big)\Big| \le \check{U}_1\tilde{U}_2.$$

Consider the filtration $\{H_t\}_{t\ge 0}$ where H_t is generated by $\{\boldsymbol{x}_{\tau,j},\boldsymbol{a}_\tau,y_{\tau,j}\}_{\tau\le t,j\le L}$, clearly, we have

$$\mathbb{E}\left(S(t;\boldsymbol{\Delta})\mid H_{t-1}\right) = S(t-1;\boldsymbol{\Delta}). \tag{EC.81}$$

By Azuma-Hoeffding Inequality, we have for any $\alpha>0$,

$$\mathbb{P}\left(S(T;\boldsymbol{\Delta})\le -\alpha\right) \le \exp\big(-\frac{\alpha^2\times 4T^2}{2\lfloor T^{2/3}\rfloor\check{U}_1^2\tilde{U}_2^2}\big). \tag{EC.82}$$

Set $\alpha = \frac{\check{U}_1\tilde{U}_2}{2\sqrt{2}T^{\frac{2}{3}}}\sqrt{\log(T)+d_ad_x(\log 3+\frac{1}{3}\log T)}$, then we have

$$\mathbb{P}(\tilde{W}_1 < -\alpha) \le \frac{1}{T^2}. \tag{EC.83}$$

Note that for $T\ge 3$,

$$\frac{\check{U}_1\tilde{U}_2}{\sqrt{2}T^{\frac{2}{3}}}\sqrt{2\log(T)+d_ad_x(\log 3+\frac{1}{3}\log T)} + \frac{2\check{U}_1\tilde{U}_2}{T^{\frac{2}{3}}} < \frac{\check{U}_1\tilde{U}_2}{T^{\frac{2}{3}}}(2+\sqrt{2d_ad_x\log(T)}). \tag{EC.84}$$

Therefore,

$$\mathbb{P}\left(Z_1(T)\le -\frac{\check{U}_1\tilde{U}_2}{T^{2/3}}(2+\sqrt{2d_ad_x\log(T)})\right)\le\frac{1}{T^2}. \tag{EC.85}$$

Analysis of $Z_2(T)$: By the definition of $Z_2(T)$, we have the following decomposition

$$Z_2(T) \le \sup_{|||\boldsymbol{\Delta}|||_{\mathrm{F}}=1}\Big|\mathbb{M}\Big(\rho_{2,T}(\boldsymbol{\Delta})\Big) - \mathbb{M}\Big(\check{\rho}_{2,T}(\boldsymbol{\Delta})\Big)\Big| + \mathbb{1}\{\check{\mathcal{U}}^c\}\cdot\sup_{|||\boldsymbol{\Delta}|||_{\mathrm{F}}=1}\big|\rho_{2,T}(\boldsymbol{\Delta})-\check{\rho}_{2,T}(\boldsymbol{\Delta})\big|,$$

where $\rho_{2,T}(T;\boldsymbol{\Delta})$ and $\check{\rho}_{2,T}(T;\boldsymbol{\Delta})$ are defined in Equation (EC.54) and Equation (EC.74a). The second term equals to zero with high probability, since the event $\check{\mathcal{U}}$ happens with high probability.

Turning to the first term, we have

$$\begin{aligned}&\sup_{|||\boldsymbol{\Delta}|||_{\mathrm{F}}=1}\Big|\mathbb{M}\Big(\rho_{2,T}(\boldsymbol{\Delta})\Big) - \mathbb{M}\Big(\check{\rho}_{2,T}(\boldsymbol{\Delta})\Big)\Big|\\ &= \sup_{|||\boldsymbol{\Delta}|||_{\mathrm{F}}=1}\frac{1}{2LT}\sum_{t\in[T]\cap\mathfrak{S}}\sum_{\ell=1}^{L}\mathbb{E}\Big(\Big(\bar{\boldsymbol{\delta}}_t^\top\boldsymbol{\Delta}\boldsymbol{x}_{t,\ell}\Big)^2\mathbb{1}\Big\{\|\boldsymbol{\delta}_t/h\|_2^2>\check{U}_2 \text{ or } \|\boldsymbol{x}_{t,\ell}\|_2^2>\check{U}_1\Big\}\Big|\{\boldsymbol{x}_{\tau,j},\boldsymbol{a}_\tau,y_{\tau,j}\}_{\tau\le t-1,j\le L}\Big)\\ &\le \underbrace{\sup_{|||\boldsymbol{\Delta}|||_{\mathrm{F}}=1}\frac{1}{2LT}\sum_{t\in[T]\cap\mathfrak{S}}\sum_{\ell=1}^{L}\mathbb{E}\Big(\big(\bar{\boldsymbol{\delta}}_t^\top\boldsymbol{\Delta}\boldsymbol{x}_{t,\ell}\big)^2\mathbb{1}\big\{\|\boldsymbol{\delta}_t/h\|_2^2>\check{U}_2\big\}\Big|\{\boldsymbol{x}_{\tau,j},\boldsymbol{a}_\tau,y_{\tau,j}\}_{\tau\le t-1,j\le L}\Big)}_{\varsigma_2}\end{aligned}$$

$$+\underbrace{\sup_{\|\boldsymbol{\Delta}\|_{\mathrm{F}}=1}\frac{1}{2LT}\sum_{t\in[T]\cap\mathfrak{S}}\sum_{\ell=1}^{L}\mathbb{E}\bigg(\Big(\bar{\boldsymbol{\delta}}_t^\top\boldsymbol{\Delta}\boldsymbol{x}_{t,\ell}\Big)^2\mathbb{1}\{\|\boldsymbol{\delta}_t/h\|_2^2\le\check{U}_2 \text{ and } \|\boldsymbol{x}_{t,\ell}\|_2^2>\check{U}_1\}\bigg|\{\boldsymbol{x}_{\tau,j},\boldsymbol{a}_\tau,y_{\tau,j}\}_{\tau\le t-1,j\le L}\bigg)}_{\varsigma_1}.$$

Next we bound ς_2 and ς_1 separately.

For notation simplicity, we denote $H_t=\{\{\boldsymbol{x}_{\tau,j},\boldsymbol{a}_\tau,y_{\tau,j}\}_{\tau\le t,j\le L}\}$. For $T\ge 4$, we have

$$\begin{aligned}
\varsigma_2 &= \sup_{\|\boldsymbol{\Delta}\|_{\mathrm{F}}=1}\frac{1}{2T}\sum_{t\in[T]\cap\mathfrak{S}}\sum_{\ell=1}^{L}\frac{1}{L}\mathbb{E}\Big(\big(\bar{\boldsymbol{\delta}}_t^\top\boldsymbol{\Delta}\boldsymbol{x}_{t,\ell}\big)^2\mathbb{1}\big\{\|\boldsymbol{\delta}_t/h\|_2^2>\check{U}_2\big\}\Big|H_{t-1}\Big)\\
&\le \sup_{\|\boldsymbol{\Delta}\|_{\mathrm{F}}=1}\frac{1}{2T}\sum_{t\in[T]\cap\mathfrak{S}}\sum_{\ell=1}^{L}\frac{1}{L}\mathbb{E}\Big(\|\bar{\boldsymbol{\delta}}_t\|_2^2\|\boldsymbol{\Delta}\boldsymbol{x}_{t,\ell}\|_2^2\mathbb{1}\big\{\|\boldsymbol{\delta}_t/h\|_2^2>\check{U}_2\big\}\Big|H_{t-1}\Big)\\
&\le \sup_{\|\boldsymbol{\Delta}\|_{\mathrm{F}}=1}\frac{1}{2T}\sum_{t\in[T]\cap\mathfrak{S}}\sum_{\ell=1}^{L}\frac{1}{L}\mathbb{E}\Big(\big(\min\{2,\|\boldsymbol{\delta}_t\|_2+h(1+\frac{\sqrt{d_a}}{2})\}\big)^2\|\boldsymbol{\Delta}\boldsymbol{x}_{t,\ell}\|_2^2\mathbb{1}\big\{\|\boldsymbol{\delta}_t/h\|_2^2>\check{U}_2\big\}\Big|H_{t-1}\Big)\\
&= \frac{1}{2T}\sum_{t\in[T]\cap\mathfrak{S}}\sum_{\ell=1}^{L}\frac{1}{L}\mathbb{E}\Big(\big(\min\{2,\|\boldsymbol{\delta}_t\|_2+h(1+\frac{\sqrt{d_a}}{2})\}\big)^2\mathbb{1}\big\{\|\boldsymbol{\delta}_t/h\|_2^2>\check{U}_2\big\}\Big)\\
&\le \frac{\lfloor T^{\frac{2}{3}}\rfloor}{2T}\min\Big\{4\mathbb{P}(\|\boldsymbol{\delta}\|_2^2>\check{U}_2),\mathbb{E}(\frac{9h^2}{4}\|\boldsymbol{\delta}\|_2^2\mathbb{1}\{\|\boldsymbol{\delta}\|_2^2>\check{U}_2\})\Big\}
\end{aligned}$$

where $\boldsymbol{\delta}\sim N(\mathbf{0},\mathbf{I}_{d_a})$. Standard Chi-square tail bound shows that $\mathbb{P}(\|\boldsymbol{\delta}\|_2^2\ge\check{U}_2)\le\frac{1}{T^2}$. Next, we will calculate $\mathbb{E}(\|\boldsymbol{\delta}\|_2^2\mathbb{1}\{\|\boldsymbol{\delta}\|_2>\check{U}_2\})$. We use the spherical coordinates. and let $V(d)$ denote the volume of the unit ball in $\mathbb{R}^d$. Then, by dividing the integral of the normal distribution density and canceling the same terms, we have

$$\mathbb{E}(|\boldsymbol{\delta}\|_2^2\mathbb{1}\{\|\boldsymbol{\delta}\|_2>\check{U}_2\})=\frac{\int_{\sqrt{\check{U}_2}}^{\infty}\exp(-r^2/2)r^2\cdot r^{d_a-1}\mathrm{d}r}{\int_0^{\infty}\exp(-r^2/2)\cdot r^{d_a-1}\mathrm{d}r}.$$

Elementary calculation shows that

$$\begin{aligned}
&\int_{\sqrt{\tau}}^{\infty}\exp(-r^2/2)\cdot r^d\mathrm{d}r\\
&\quad=\begin{cases}\exp(-\frac{\tau}{2})2^{k-\frac{1}{2}}\Big(\frac{(k-1/2)!}{(1/2)!}\int_{\frac{\tau}{2}}^{\infty}\sqrt{t}\exp(\frac{\tau}{2}-t)\mathrm{d}t+\sum_{i=0}^{k-2}(\frac{\tau}{2})^{k-\frac{1}{2}-i}\frac{(k-1/2)!}{(k-(1/2)-i)!}\Big), & \text{even } d=2k\\ \exp(-\frac{\tau}{2})2^k\sum_{i=0}^{k}(\frac{\tau}{2})^{k-i}\frac{k!}{(k-i)!}, & \text{odd } d=2k+1\end{cases}.
\end{aligned}$$

Note that $\check{U}_2>d_a$. Therefore, when d_a is even, we have

$$\begin{aligned}
\mathbb{E}(\|\boldsymbol{\delta}\|_2^2\mathbb{1}\{\|\boldsymbol{\delta}\|_2>\check{U}_2\}) &= 2\frac{\exp(-\check{U}_2/2)\sum_{i=0}^{d_a/2}(\frac{\check{U}_2}{2})^{(d_a/2)-i}\frac{(d_a/2)!}{((d_a/2)-i)!}}{((d_a/2)-1)!}\\
&=\exp(-\check{U}_2/2)d_a\sum_{i=0}^{d_a/2}(\frac{\check{U}_2}{2})^{(d_a/2)-i}\frac{1}{((d_a/2)-i)!}\\
&<\exp(-\check{U}_2/2)(\frac{\check{U}_2}{2})^{(d_a/2)}\frac{1}{(d_a/2)!}\frac{d_a}{1-(d_a/\check{U}_2)},
\end{aligned}$$

whereas when d_a is odd, we have

$$\begin{aligned}
\mathbb{E}(\|\boldsymbol{\delta}\|_2^2\mathbb{1}\{\|\boldsymbol{\delta}\|_2>\check{U}_2\}) &= \frac{\exp(-\frac{\check{U}_2}{2})2^{\frac{d_a}{2}}\Big(\frac{(\frac{d_a}{2})!}{(1/2)!}\int_{\frac{\check{U}_2}{2}}^{\infty}\sqrt{t}\exp(\frac{\check{U}_2}{2}-t)\mathrm{d}t+\sum_{i=0}^{(d_a-3)/2}(\frac{\check{U}_2}{2})^{\frac{d_a}{2}-i}\frac{(\frac{d_a}{2})!}{(\frac{d_a}{2}-i)!}\Big)}{2^{\frac{d_a-2}{2}}\sqrt{\pi}\prod_{i=0}^{(d_a-5)/2}(d_a/2-1-i)}\\
&=\exp(-\frac{\check{U}_2}{2})\frac{d_a}{\sqrt{\pi}}\Big(\int_{\frac{\check{U}_2}{2}}^{\infty}\frac{1}{2\sqrt{t}}\exp(\frac{\check{U}_2}{2}-t)\mathrm{d}t+\sum_{i=0}^{(d_a-1)/2}(\frac{\check{U}_2}{2})^{\frac{d_a}{2}-i}\frac{(1/2)!}{(\frac{d_a}{2}-i)!}\Big)
\end{aligned}$$

$$< \exp(-\frac{\check{U}_2}{2})(\frac{\check{U}_2}{2})^{(d_a/2)}\frac{d_a}{\sqrt{\pi}}\Big(\frac{1}{1-(d_a/\check{U}_2)}\frac{(1/2)!}{(d_a/2)!}+\sqrt{\pi}(\frac{\check{U}_2}{2})^{(-d_a/2)}\Big).$$

Plugging in $\check{U}_2 \geq 2d_a + 6\log T + d_a \log\log T$ and using Stirling formula, we have that for $d_a \geq 1, T \geq 2$,

$$\mathbb{E}(\|\boldsymbol{\delta}\|_2^2 \mathbb{1}\{\|\boldsymbol{\delta}\|_2 > \check{U}_2\}) < \frac{2}{T^{3/2}}.$$

Therefore, for $T \geq 4$, we have the bound

$$\varsigma_2 < \min\{T^{-(11/6)}, \frac{9h^2}{4}T^{-(11/6)}\}.$$

Similarly, for ς_1, as $\check{U}_1 \geq 2d_x + 6\log LT + \log\log(LT)$, we have that for $T \geq 2$,

$$\varsigma_1 = \sup_{\|\boldsymbol{\Delta}\|_{\mathrm{F}}=1}\frac{1}{2T}\sum_{t\in[T]\cap\mathfrak{S}}\sum_{\ell=1}^{L}\frac{1}{L}\mathbb{E}\Big(\big(\bar{\boldsymbol{\delta}}_t^\top \boldsymbol{\Delta}\boldsymbol{x}_{t,\ell}\big)^2 \mathbb{1}\big\{\|\boldsymbol{\delta}_t/h\|_2^2 \leq \check{U}_2, \|\boldsymbol{x}_{t,\ell}\|_2^2 > \check{U}_1\big\}\Big| H_{t-1}\Big) \tag{EC.86}$$

$$\leq \frac{1}{2T}\sum_{t\in[T]\cap\mathfrak{S}}\sum_{\ell=1}^{L}\frac{1}{L}\mathbb{E}\Big(\|\bar{\boldsymbol{\delta}}_t\|_2^2\|\boldsymbol{x}_{t,\ell}\|_2^2 \mathbb{1}\big\{\|\boldsymbol{\delta}_t/h\|_2^2 \leq \check{U}_2, \|\boldsymbol{x}_{t,\ell}\|_2^2 > \check{U}_1\big\}\Big| H_{t-1}\Big) \tag{EC.87}$$

$$\leq \frac{1}{2T}\tilde{U}_2\sum_{t\in[T]\cap\mathfrak{S}}\sum_{\ell=1}^{L}\frac{1}{L}\mathbb{E}\Big(\|\boldsymbol{x}_{t,\ell}\|_2^2 \mathbb{1}\big\{\|\boldsymbol{x}_{t,\ell}\|_2^2 > \check{U}_1\big\}\Big| H_{t-1}\Big) \tag{EC.88}$$

$$\leq \frac{\lfloor T^{\frac{2}{3}}\rfloor}{2T}\tilde{U}_2\frac{2}{LT^{3/2}} \leq \frac{\tilde{U}_2}{T^{11/6}}, \tag{EC.89}$$

Therefore, we have shown that

$$Z_2(T) \leq \frac{\tilde{U}_2}{T^{11/6}} + \min\{T^{-(11/6)}, \frac{9h^2}{4}T^{-(11/6)}\} + \mathbb{1}\{\check{\mathcal{U}}^c\}\cdot\sup_{\|\boldsymbol{\Delta}\|_{\mathrm{F}}=1}|\rho_{2,T}(\boldsymbol{\Delta}) - \check{\rho}_{2,T}(\boldsymbol{\Delta})|.$$

Combining the analysis of $Z_1(T)$ and $Z_2(T)$ with the decomposition (EC.75) yields

$$\mathbb{P}\Big(\Big\{\inf_{\|\boldsymbol{\Delta}\|_{\mathrm{F}}>0}\mathfrak{D}_2(T;\boldsymbol{\Delta}) < -\frac{\check{U}_1\tilde{U}_2}{T^{2/3}}(2+\sqrt{2d_a d_x\log(T)}) - T^{-11/6}(\tilde{U}_2 + (1\wedge\frac{9h^2}{4}))\Big\}\cap\check{\mathcal{U}}\Big) \leq \frac{1}{T^2}$$

Plugging in $\check{U}_1$, $\check{U}_2$, and $\tilde{U}_2$ from Equation (EC.73) and Equation (EC.76), and setting

$$\kappa_2 := -\frac{1}{T^{2/3}}\Big(2d_x + 6\log(TL) + d_x\log\log(TL))(2+\sqrt{3d_a d_x\log(T)}\Big)\Big(4\wedge\frac{9}{2}h^2(d_a + 3\log(T) + \frac{1}{2}d_a\log\log T)\Big), \tag{EC.90}$$

we have that for $T \geq 4$,

$$\mathbb{P}\Big(\big\{\inf_{\|\boldsymbol{\Delta}\|_{\mathrm{F}}>0}\mathfrak{D}_2(T;\boldsymbol{\Delta}) < \kappa_2\big\}\cap\check{\mathcal{U}}\Big) < \frac{1}{T^2}. \tag{EC.91}$$

Next we combine the high-probability bounds for $\inf_{\|\boldsymbol{\Delta}\|_{\mathrm{F}}>0}\mathfrak{D}_2(T;\boldsymbol{\Delta})$ Inequality (EC.91) and $\inf_{\|\boldsymbol{\Delta}\|_{\mathrm{F}}>0}\mathfrak{D}_1(T;\boldsymbol{\Delta})$ Inequality (EC.72) to prove lower bound for $e(\boldsymbol{\Delta})$.

Recalling the inclusion $\check{\mathcal{U}} \subset \mathcal{U}$, we have

$$\mathbb{P}\left(\inf_{\|\boldsymbol{\Delta}\|_{\mathrm{F}}>0}\mathfrak{D}_1(T;\boldsymbol{\Delta}) < \kappa_1 \text{ or } \inf_{\|\boldsymbol{\Delta}\|_{\mathrm{F}}>0}\mathfrak{D}_2(T;\boldsymbol{\Delta}) < \kappa_2\right) < \frac{1}{T} + \frac{1}{T^2} + \mathbb{P}(\mathcal{U}^c) \leq \frac{2}{T} + \frac{1}{LT} + \frac{1}{T^2},$$

where κ_1 and κ_2 are defined in Equation (EC.71) and Equation (EC.90), respectively. Therefore, going back to Inequality (EC.57), we have for $T \geq 4$,

$$e(\boldsymbol{\Delta}) \geq \frac{\lfloor T^{2/3}\rfloor\tilde{c}_6\left(\frac{1}{d_a-1}\wedge h^2\wedge(d_a-1)h^4\mathbb{1}\{h\geq\frac{1}{4(d_a-1)}\}\right)}{2T}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 + \kappa_1\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 + \kappa_2\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2$$

$$
\begin{aligned}
&\overset{\text{plug in } \kappa_1,\kappa_2}{=} \frac{\lfloor T^{2/3}\rfloor \tilde{c}_6\left(\frac{1}{d_a-1}\wedge h^2\wedge (d_a-1)h^4\mathbb{1}\{h\geq \frac{1}{4(d_a-1)}\}\right)}{2T}\|\!|\boldsymbol{\Delta}|\!\|_{\mathrm{F}}^2\\
&-\frac{6}{T^{2/3}}\alpha_h(d_x+3\log(TL))\left(2\wedge\frac{3h}{\sqrt{2}}\sqrt{d_a+3\log(T)}\right)\sqrt{\log T+d_ad_x(\frac{1}{3}\log(T)+\log 3)}\|\!|\boldsymbol{\Delta}|\!\|_{\mathrm{F}}^2\\
&-\frac{1}{T^{2/3}}(2d_x+6\log(TL)+d_x\log\log(TL))(2+\sqrt{3d_ad_x\log(T)})\left(4\wedge\frac{9}{2}h^2(d_a+3\log(T)+\frac{1}{2}d_a\log\log T)\right)\|\!|\boldsymbol{\Delta}|\!\|_{\mathrm{F}}^2\\
&\overset{\text{simplify terms}}{\geq} \frac{\lfloor T^{2/3}\rfloor \tilde{c}_6\left(\frac{1}{d_a-1}\wedge h^2\wedge (d_a-1)h^4\mathbb{1}\{h\geq \frac{1}{4(d_a-1)}\}\right)}{2T}\|\!|\boldsymbol{\Delta}|\!\|_{\mathrm{F}}^2\\
&\quad-\frac{c}{T^{2/3}}(d_x+3\log(TL)+d_x\log\log(TL))\sqrt{d_ad_x\log(T)}\left(1\wedge h\sqrt{d_a+3\log(T)+\frac{d_a}{2}\log\log T}\right)\|\!|\boldsymbol{\Delta}|\!\|_{\mathrm{F}}^2,
\end{aligned}
$$

for some absolute positive constant $c>0$, with probability at least $1-\frac{2}{T}-\frac{1}{T^2}-\frac{1}{LT}$. This concludes the proof of the lemma.

EC.2.4.3. Proof of Lemma EC.2 For $t-1\notin\{\lfloor w^{\frac{3}{2}}\rfloor : w\in\mathbb{Z}_+\}$,

$$
\mathbb{E}((\bar{\boldsymbol{\delta}}_t^\top\boldsymbol{\Delta}\boldsymbol{x}_{t,\ell})^2\Big|\{\boldsymbol{a}_\tau,\boldsymbol{x}_{\tau,j},y_{\tau,j}\}_{\tau\leq t-1,j\leq L})=0. \tag{EC.92}
$$

Now we turn to $t-1\in\{\lfloor w^{\frac{3}{2}}\rfloor : w\in\mathbb{Z}_+\}$. Let $U(v)$ be a rotation matrix that rotates $(1,0,\ldots,0)^\top$ to $\frac{v}{\|v\|_2}$. As there are many rotation matrices that rotate $(1,0,\ldots,0)$ to v, for any v, we just pick anyone. Therefore, U is a well-defined map that maps a vector to a unitary matrix. Then

$$
\begin{aligned}
\bar{\boldsymbol{\delta}}_t&=U(\frac{\hat{\boldsymbol{\Theta}}_{t-1}\boldsymbol{b}_t}{\|\hat{\boldsymbol{\Theta}}_{t-1}\boldsymbol{b}_t\|_2})U(\frac{\hat{\boldsymbol{\Theta}}_{t-1}\boldsymbol{b}_t}{\|\hat{\boldsymbol{\Theta}}_{t-1}\boldsymbol{b}_t\|_2})^\top\left(\mathcal{P}_{\mathcal{A}_{t-1}}(\frac{\hat{\boldsymbol{\Theta}}_{t-1}\boldsymbol{b}_t}{\|\hat{\boldsymbol{\Theta}}_{t-1}\boldsymbol{b}_t\|_2}+\boldsymbol{\delta}_t)-\alpha_h\frac{\hat{\boldsymbol{\Theta}}_{t-1}\boldsymbol{b}_t}{\|\hat{\boldsymbol{\Theta}}_{t-1}\boldsymbol{b}_t\|_2}\right) &\text{(EC.93)}\\
&=U(\frac{\hat{\boldsymbol{\Theta}}_{t-1}\boldsymbol{b}_t}{\|\hat{\boldsymbol{\Theta}}_{t-1}\boldsymbol{b}_t\|_2})\left(\mathcal{P}_{\mathcal{A}_{t-1}}((1,0,\ldots,0)^\top+U(\frac{\hat{\boldsymbol{\Theta}}_{t-1}\boldsymbol{b}_t}{\|\hat{\boldsymbol{\Theta}}_{t-1}\boldsymbol{b}_t\|_2})^\top\boldsymbol{\delta}_t)-\alpha_h(1,0,\ldots,0)^\top)\right). &\text{(EC.94)}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\bar{\boldsymbol{\delta}}_t\mid\{\boldsymbol{x}_{i,\ell}\}_{i\leq t,\ell\leq L},\{\boldsymbol{a}_i\}_{i\leq t-1},\{y_{i,\ell}\}_{i\leq t-1,\ell\leq L} &\text{(EC.95)}\\
&=U(\frac{\hat{\boldsymbol{\Theta}}_{t-1}\boldsymbol{b}_t}{\|\hat{\boldsymbol{\Theta}}_{t-1}\boldsymbol{b}_t\|_2})\left(\mathcal{P}_{\mathcal{A}_{t-1}}((1,0,\ldots,0)^\top+h(\varepsilon_1,\ldots,\varepsilon_{d_a})^\top)-\alpha_h(1,0,\ldots,0)^\top)\right), &\text{(EC.96)}
\end{aligned}
$$

where $\varepsilon_1,\ldots,\varepsilon_{d_a}\overset{i.i.d.}{\sim}N(0,1)$ and also independent of $\{\boldsymbol{x}_{i,\ell}\}_{j\leq t+1,\ell\leq L},\{\boldsymbol{a}_i\}_{i\leq t},\{y_{i,\ell}\}_{i\leq t,\ell\leq L}$. Therefore,

$$
\mathbb{E}\left(\bar{\boldsymbol{\delta}}_t\bar{\boldsymbol{\delta}}_t^\top\mid\{\boldsymbol{x}_{\tau,\ell}\}_{\tau\leq t,\ell\leq L}\{\boldsymbol{a}_i\}_{i\leq t-1},\{y_{i,\ell}\}_{i\leq t-1,\ell\leq L}\right)=U(\frac{\hat{\boldsymbol{\Theta}}_{t-1}\boldsymbol{b}_t}{\|\hat{\boldsymbol{\Theta}}_{t-1}\boldsymbol{b}_t\|_2})\begin{pmatrix}V_1&0&0&\cdots&0\\0&V_2&0&\cdots&0\\0&0&V_3&\cdots&0\\\vdots&\vdots&\vdots&\ddots&\vdots\\0&0&0&0&V_{d_a}\end{pmatrix}U(\frac{\hat{\boldsymbol{\Theta}}_{t-1}\boldsymbol{b}_t}{\|\hat{\boldsymbol{\Theta}}_{t-1}\boldsymbol{b}_t\|_2})^\top, \tag{EC.97}
$$

where

$$
V_j=\begin{cases}Var(\frac{1+h\varepsilon_1}{1\vee\sqrt{(1+h\varepsilon_1)^2+\sum_{i=2}^{d_a}h^2\varepsilon_i^2}}) & \text{for } j=1\\ Var(\frac{h\varepsilon_j}{1\vee\sqrt{(1+h\varepsilon_1)^2+\sum_{i=2}^{d_a}h^2\varepsilon_i^2}}) & \text{for } j\geq 2\end{cases}. \tag{EC.98}
$$

We will show later that

$$\begin{pmatrix} V_1 & 0 & 0 & \cdots & 0 \\ 0 & V_2 & 0 & \cdots & 0 \\ 0 & 0 & V_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & V_{d_a} \end{pmatrix} \succeq \tilde{c}_6 \left(\frac{1}{d_a - 1} \wedge h^2 \wedge (d_a - 1) h^4 \mathbb{1}\{h \geq \frac{1}{4(d_a-1)}\} \right) \mathbf{I}_{d_a} \tag{EC.99}$$

for some absolute positive constant $\tilde{c}_6$.

Therefore,

$$\begin{aligned}
&\mathbb{E}\Big((\bar{\boldsymbol{\delta}}_t^\top \boldsymbol{\Delta} \boldsymbol{x}_{t,\ell})^2 \Big| \{\boldsymbol{a}_\tau, \boldsymbol{x}_{\tau,j}, y_{\tau,j}\}_{\tau \leq t-1, j \leq L} \Big) \\
&= \mathbb{E}\left(\mathbb{E}\left(\mathbb{E}\Big((\bar{\boldsymbol{\delta}}_t^\top \boldsymbol{\Delta} \boldsymbol{x}_{t,\ell})^2 \mid \{\boldsymbol{x}_{\tau,\ell}\}_{\tau \leq t, \ell \leq L} \{\boldsymbol{a}_i\}_{i \leq t-1}, \{y_{i,\ell}\}_{i \leq t-1, \ell \leq L} \Big) \right) \Big| \{\boldsymbol{a}_\tau, \boldsymbol{x}_{\tau,j}, y_{\tau,j}\}_{\tau \leq t-1, j \leq L} \right) \\
&\geq \tilde{c}_6 \left(\frac{1}{d_a - 1} \wedge h^2 \wedge (d_a - 1) h^4 \mathbb{1}\{h \geq \frac{1}{4(d_a-1)}\} \right) \times \\
&\qquad \mathbb{E}\left(\boldsymbol{x}_{t,l}^\top \boldsymbol{\Delta}^\top U(\frac{\hat{\boldsymbol{\Theta}}_{t-1} \boldsymbol{b}_t}{\|\hat{\boldsymbol{\Theta}}_{t-1} \boldsymbol{b}_t\|_2}) U(\frac{\hat{\boldsymbol{\Theta}}_{t-1} \boldsymbol{b}_t}{\|\hat{\boldsymbol{\Theta}}_{t-1} \boldsymbol{b}_t\|_2})^\top \boldsymbol{\Delta} \boldsymbol{x}_{t,l} \Big| \{\boldsymbol{a}_\tau, \boldsymbol{x}_{\tau,j}, y_{\tau,j}\}_{\tau \leq t-1, j \leq L} \right) \\
&= \tilde{c}_6 \left(\frac{1}{d_a - 1} \wedge h^2 \wedge (d_a - 1) h^4 \mathbb{1}\{h \geq \frac{1}{4(d_a-1)}\} \right) |\!|\!| \boldsymbol{\Delta} |\!|\!|_{\mathrm{F}}^2.
\end{aligned} \tag{EC.100}$$

Therefore,

$$\begin{aligned}
\mathbb{M}(\rho_{2,T}(\boldsymbol{\Delta})) &= \frac{1}{2LT} \sum_{t=1}^{T} \sum_{\ell=1}^{L} \mathbb{E}\Big((\bar{\boldsymbol{\delta}}_t^\top \boldsymbol{\Delta} \boldsymbol{x}_{t,\ell})^2 \Big| \{\boldsymbol{a}_\tau, \boldsymbol{x}_{\tau,j}, y_{\tau,j}\}_{\tau \leq t-1, j \leq L} \Big) \\
&\geq \frac{L \lfloor T^{\frac{2}{3}} \rfloor}{2LT} \tilde{c}_6 \left(\frac{1}{d_a - 1} \wedge h^2 \wedge (d_a - 1) h^4 \mathbb{1}\{h \geq \frac{1}{4(d_a-1)}\} \right) |\!|\!| \boldsymbol{\Delta} |\!|\!|_{\mathrm{F}}^2,
\end{aligned} \tag{EC.101}$$

which gives the statement of Lemma EC.2.

Now we only need to show Inequality (EC.99).

For $j \geq 2$, elementary calculation show that

$$\begin{aligned}
V_j &= \mathbb{E}\left(\frac{h^2 \varepsilon_j^2}{1 \vee \left((1 + h\varepsilon_1)^2 + \sum_{i=2}^{d_a} h^2 \varepsilon_i^2 \right)}\right) && \text{(EC.102)} \\
&= \frac{1}{d_a - 1} \mathbb{E}\left(\frac{\sum_{i=2}^{d_a} h^2 \varepsilon_i^2}{1 \vee \left((1 + h\varepsilon_1)^2 + \sum_{i=2}^{d_a} h^2 \varepsilon_i^2 \right)}\right) && \text{(EC.103)} \\
&\overset{(a)}{\geq} \frac{1}{d_a - 1} \mathbb{E}\left(\frac{\sum_{i=2}^{d_a} h^2 \varepsilon_i^2}{1 \vee \left(\mathbb{E}\big((1 + h\varepsilon_1)^2\big) + \sum_{i=2}^{d_a} h^2 \varepsilon_i^2 \right)}\right) && \text{(EC.104)} \\
&= \frac{1}{d_a - 1} \mathbb{E}\left(\frac{\sum_{i=2}^{d_a} h^2 \varepsilon_i^2}{1 + h^2 + \sum_{i=2}^{d_a} h^2 \varepsilon_i^2}\right), && \text{(EC.105)}
\end{aligned}$$

where step (a) follows from taking conditional expectation conditioned on $\varepsilon_2, \ldots, \varepsilon_{d_a}$ and the convexity of the function (of x)

$$\frac{\sum_{i=2}^{d_a} h^2 \varepsilon_i^2}{1 \vee (x + \sum_{i=2}^{d_a} h^2 \varepsilon_i^2)}.$$

Note that for $H < d_a - 1$,

$$\mathbb{P}\left(\sum_{i=2}^{d_a} \varepsilon_i^2 \leq H\right) \leq \inf_{\lambda > 0} \mathbb{E}\left(\exp\left(-\lambda\left(\sum_{i=2}^{d_a} \varepsilon_i^2\right) + \lambda H\right)\right)$$

$$= \inf_{\lambda>0} \exp\left(\lambda H - \frac{d_a-1}{2}\log(1+2\lambda)\right)$$
$$= \exp\left(\frac{d_a-1-H}{2} - \frac{d_a-1}{2}\log\left(\frac{d_a-1}{H}\right)\right). \tag{EC.106}$$

Letting $H = \frac{d_a-1}{3}$ gives

$$\mathbb{P}\left(\sum_{i=2}^{d_a}\varepsilon_i^2 \le \frac{d_a-1}{3}\right) \le \left(\frac{\exp(1)}{3\sqrt{3}}\right)^{\frac{d_a-1}{3}} \le \frac{\exp(1/3)}{\sqrt{3}}. \tag{EC.107}$$

Going back to Equation (EC.105) gives

$$V_j \ge \frac{1}{d_a-1}\left(1-\frac{\exp(1)}{3\sqrt{3}}\right)\frac{h^2(d_a-1)/3}{1+h^2(d_a+2)/3}$$
$$\ge \tilde{c}_0 \frac{h^2}{1+h^2(d_a+2)/3}, \tag{EC.108}$$

where $\tilde{c}_0 = \frac{1}{3}\left(1-\frac{\exp(1)}{3\sqrt{3}}\right)$.

Now we turn to calculating V_1, for notational simplicity, we let $r^2 = h^2\sum_{i=2}^{d_a}\varepsilon_i^2$. We bound V_1 in three settings: $h \ge 1/\sqrt{d_a-1}$, $h < \frac{1}{2(d_a-1)}$, and $h \in [\frac{1}{d_a-1}, 1/\sqrt{d_a-1})$. We will show that

$$V_1 \ge \begin{cases} \tilde{c}_3\frac{1}{d_a-1}, & \text{for } h \ge \frac{1}{\sqrt{d_a-1}} \\ \tilde{c}_5(d_a-1)h^4, & \text{for } h \in [\frac{1}{4(d_a-1)}, 1/\sqrt{d_a-1}] \\ \tilde{c}_4 h^2, & \text{for } h \le \frac{1}{4(d_a-1)} \end{cases}, \tag{EC.109}$$

where $\tilde{c}_3, \tilde{c}_4, \tilde{c}_5$ are positive constants.

We start with $h \ge 1/\sqrt{d_a-1}$. Therefore,

$$\begin{aligned} V_1 =& Var\left(\frac{1+h\varepsilon_1}{1\vee\sqrt{(1+h\varepsilon_1)^2+\sum_{i=2}^{d_a}h^2\varepsilon_i^2}}\right) \\ \ge& \frac{1}{4}\mathbb{E}\left(\left(\frac{1+h|\varepsilon_1|}{1\vee\sqrt{(1+h|\varepsilon_1|)^2+\sum_{i=2}^{d_a}h^2\varepsilon_i^2}} - \frac{1-h|\varepsilon_1|}{1\vee\sqrt{(1-h|\varepsilon_1|)^2+\sum_{i=2}^{d_a}h^2\varepsilon_i^2}}\right)^2\right) \\ =& \underbrace{\frac{1}{4}\mathbb{E}\left(\left(\frac{1+h|\varepsilon_1|}{1\vee\sqrt{(1+h|\varepsilon_1|)^2+\sum_{i=2}^{d_a}h^2\varepsilon_i^2}} - \frac{1-h|\varepsilon_1|}{1\vee\sqrt{(1-h|\varepsilon_1|)^2+\sum_{i=2}^{d_a}h^2\varepsilon_i^2}}\right)^2\mathbb{1}\{|\varepsilon_1|h\le 1\}\right)}_{\xi_1} \\ &+ \underbrace{\frac{1}{4}\mathbb{E}\left(\left(\frac{1+h|\varepsilon_1|}{1\vee\sqrt{(1+h|\varepsilon_1|)^2+\sum_{i=2}^{d_a}h^2\varepsilon_i^2}} - \frac{1-h|\varepsilon_1|}{1\vee\sqrt{(1-h|\varepsilon_1|)^2+\sum_{i=2}^{d_a}h^2\varepsilon_i^2}}\right)^2\mathbb{1}\{|\varepsilon_1|h> 1\}\right)}_{\xi_2} \end{aligned} \tag{EC.110}$$

Now we bound ξ_1 and ξ_2 separately and start with ξ_1.

Consider the function $f: x \mapsto \frac{x}{\sqrt{x^2+r^2}}$ and note that the first order and second order derivatives of it are

$$f'(x) = \frac{r^2}{(x^2+r^2)^{\frac{3}{2}}}, \qquad f''(x) = -3\frac{xr^2}{(x^2+r^2)^{\frac{5}{2}}}. \tag{EC.111}$$

Therefore, $f(x)$ is concave for $x \ge 0$. So $\min\{f(x), x\}$ is also concave for $x \ge 0$. Therefore, we have

$$f(1+h|\varepsilon_1|) - f(1-h|\varepsilon_1|) \ge f(1) - f(1-h|\varepsilon_1|) \ge f'(1)h|\varepsilon_1| = \frac{r^2}{(1+r^2)^{\frac{3}{2}}}h|\varepsilon_1| \tag{EC.112}$$

Plugging in this relationship into ξ_1 gives

$$\begin{aligned}\xi_1 \geq & \mathbb{E}\left(\left(\frac{r^2}{(1+r^2)^{\frac{3}{2}}}h|\varepsilon_1|\right)^2 \mathbb{1}\{|\varepsilon_1|h \leq 1\}\right) \\ = & \mathbb{E}\left(\frac{r^4}{(1+r^2)^3}\right)\mathbb{E}\left(h^2\varepsilon_1^2\mathbb{1}\{|\varepsilon_1|h \leq 1\}\right) \\ \geq & \mathbb{1}\{h \leq 1\}h^2\mathbb{E}\left(\frac{r^4}{(1+r^2)^3}\right)\mathbb{E}\left(\varepsilon_1^2\mathbb{1}\{|\varepsilon_1| \leq 1\}\right). \end{aligned} \tag{EC.113}$$

Note that by Inequality (EC.27) and Inequality (EC.107) we have that

$$\mathbb{P}\left(r^2/h^2 \in \left[\frac{d_a-1}{3}, (d_a-1)(1+2\sqrt{2}+4)\right]\right) \geq \underbrace{1-\exp\left(-2(d_a-1)\right)-\left(\frac{\exp\left(1/3\right)}{\sqrt{3}}\right)^{d_a-1}}_{\tilde{c}_1(d_a-1)}. \tag{EC.114}$$

Clearly, $\tilde{c}_1(d_a-1)$ increases with d_a, and $\tilde{c}_1(1) > 0$, $\lim_{n\to\infty}\tilde{c}_1(n) = 1$. Going back to Inequality (EC.113) gives

$$\xi_1 \geq \mathbb{1}\{h \leq 1\}\frac{1}{16}\frac{1}{(d_a-1)(6+2\sqrt{2})}\tilde{c}_1(1)\mathbb{E}\left(\varepsilon_1^2\mathbb{1}\{|\varepsilon_1| \leq 1\}\right) = \mathbb{1}\{h \leq 1\}\tilde{c}_2\frac{1}{d_a-1}. \tag{EC.115}$$

Now we turn to bounding ξ_2. Repeatedly using $h|\varepsilon_1| > 1$ and convexity gives

$$\begin{aligned}\xi_2 \geq & \mathbb{E}\left(\left(\frac{1+h|\varepsilon_1|}{1\vee\sqrt{(1+h|\varepsilon_1|)^2+\sum_{i=2}^{d_a}h^2\varepsilon_i^2}} - \frac{1-h|\varepsilon_1|}{1\vee\sqrt{(1+h|\varepsilon_1|)^2+\sum_{i=2}^{d_a}h^2\varepsilon_i^2}}\right)^2\mathbb{1}\{|\varepsilon_1|h > 1\}\right) \\ = & \mathbb{E}\left(\frac{4h^2\varepsilon_1^2}{(1+h|\varepsilon_1|)^2+r^2}\mathbb{1}\{|\varepsilon_1|h > 1\}\right) \\ \overset{(ii)}{\geq} & \mathbb{E}\left(\frac{4h^2\varepsilon_1^2}{(1+h|\varepsilon_1|)^2+(d_a-1)h^2}\mathbb{1}\{|\varepsilon_1|h > 1\}\right) \\ \geq & \mathbb{E}\left(\frac{4h^2\varepsilon_1^2}{4h^2\varepsilon_1^2+3(d_a-1)h^2}\mathbb{1}\{|\varepsilon_1|h > 1\}\right) \\ \geq & \mathbb{1}\{h > 1\}2\Phi(-1)\frac{4}{4+3(d_a-1)}, \end{aligned} \tag{EC.116}$$

where step (i) follows from convexity of the function $g : x \mapsto \frac{4h^2\varepsilon_1^2}{(1+h|\varepsilon_1|)^2+x}$.

Combining Inequality (EC.115) and Inequality (EC.116) back to Equation (EC.110) gives

$$V_1 \geq \tilde{c}_3\frac{1}{d_a-1} \tag{EC.117}$$

for some absolute constant $\tilde{c}_3 > 0$.

Now we turn to the case $h < \frac{1}{4(d_a-1)}$.

$$\begin{aligned}V_1 = & Var(\frac{1+h\varepsilon_1}{1\vee\sqrt{(1+h\varepsilon_1)^2+\sum_{i=2}^{d_a}h^2\varepsilon_i^2}}) \\ \geq & \frac{1}{4}\mathbb{E}\left(\left(\frac{1+h|\varepsilon_1|}{1\vee\sqrt{(1+h|\varepsilon_1|)^2+\sum_{i=2}^{d_a}h^2\varepsilon_i^2}} - \frac{1-h|\varepsilon_1|}{1\vee\sqrt{(1-h|\varepsilon_1|)^2+\sum_{i=2}^{d_a}h^2\varepsilon_i^2}}\right)^2\right) \\ \geq & \frac{1}{4}\mathbb{E}\left(\left(\frac{1}{\sqrt{1+r^2}} - (1-h|\varepsilon_1|)\right)^2\mathbb{1}\{(1-h|\varepsilon_1|)^2+r^2 \leq 1\}\right). \end{aligned} \tag{EC.118}$$

By Inequality (EC.27), we have

$$\mathbb{P}(r^2/h^2 \leq d_a-1+2\sqrt{d_a-1}+1) \geq 1-\exp\left(-1\right). \tag{EC.119}$$

Note that under the event $\{r^2/h^2 \le d_a - 1 + 2\sqrt{d_a-1} + 1, |\varepsilon_1| \in [1,4]\}$, we have $\mathbb{1}\{|\varepsilon_1|h \le 1, (1-h|\varepsilon_1|)^2 + r^2 \le 1\} = 1$, and

$$\frac{1}{\sqrt{1+r^2}} - (1-h|\varepsilon_1|) \tag{EC.120}$$

$$\ge \frac{1}{\sqrt{1+h^2(d_a-1+2\sqrt{d_a-1}+1)}} - 1 + h|\varepsilon_1| \tag{EC.121}$$

$$= h|\varepsilon_1| - \frac{h^2(d_a-1+2\sqrt{d_a-1}+1)}{\sqrt{1+h^2(d_a-1+2\sqrt{d_a-1}+1)}(1+\sqrt{1+h^2(d_a-1+2\sqrt{d_a-1}+1)})} \tag{EC.122}$$

$$\overset{(i)}{\ge} h|\varepsilon_1| - \frac{1}{4(d_a-1)}h(d_a-1+2\sqrt{d_a-1}+1) \ge h(|\varepsilon_1|-1), \tag{EC.123}$$

where step (i) follows from the fact that the long denominator in Equation (EC.122) is no smaller than 2 and $h < \frac{d_a-1}{4}$.

Therefore,

$$\begin{aligned} V_1 \ge& \frac{1}{4}h^2\mathbb{E}\left((|\varepsilon_1|-1)^2\mathbb{1}\{r^2/h^2 \le d_a - 1 + 2\sqrt{d_a-1}+1, |\varepsilon_1| \in [1,4]\}\right) \\ \ge& h^2\frac{\mathbb{E}\left((|\varepsilon_1|-1)^2\mathbb{1}\{|\varepsilon_1| \in [1,4]\}\right)(1-\exp(-1))}{4} \\ =& \tilde{c}_4 h^2, \end{aligned} \tag{EC.124}$$

where $\tilde{c}_4 = \mathbb{E}\left((|\varepsilon_1|-1)^2\mathbb{1}\{|\varepsilon_1| \in [1,4]\}\right)\frac{(1-\exp(-1))}{4}$.

Now we turn to the case $h \in [\frac{1}{4(d_a-1)}, \frac{1}{\sqrt{d_a-1}})$. We first take a detour on providing bounds on the probability of Chi-square distributions that we will prove later. Suppose χ^2_d follows Chi-square distribution with degree of freedom d, then we have for $d \ge 1$,

$$\mathbb{P}(\chi^2_d \ge d + \sqrt{d}) > 0.018, \qquad \mathbb{P}(\chi^2_d \le d) > 0.09. \tag{EC.125}$$

With Inequality (EC.125), we have

$$\begin{aligned} V_1 \ge& \mathbb{E}\left(0.018 \times \frac{1}{4}\left(\frac{1+h\varepsilon_1}{\sqrt{(1+h\varepsilon_1)^2+h^2(d_a-1)}} - \frac{1+h\varepsilon_1}{\sqrt{(1+h\varepsilon_1)^2+h^2(d_a-1+\sqrt{d_a-1})}}\right)^2 \mathbb{1}\{\varepsilon_1 \ge 0\}\right) \\ \ge& \mathbb{E}\left(0.018 \times \frac{1}{4}\left(\frac{1+h\varepsilon_1}{2((1+h\varepsilon_1)^2+h^2(d_a-1+\sqrt{d_a-1}))^{\frac{3}{2}}}\sqrt{d_a-1}h^2\right)^2 \mathbb{1}\{\varepsilon_1 \ge 0\}\right) \\ \overset{(i)}{\ge}& \mathbb{E}\left(\frac{0.018}{16}\left(\frac{1}{\sqrt{1+h^2(d_a-1+\sqrt{d_a-1})}}\frac{1}{(1+h)^2+h^2(d_a-1+\sqrt{d_a-1})}\sqrt{d_a-1}h^2\right)^2 \mathbb{1}\{1 \ge \varepsilon_1 \ge 0\}\right) \\ \overset{(ii)}{>}& \tilde{c}_5(d_a-1)h^4, \end{aligned} \tag{EC.126}$$

where $\tilde{c}_5 = 0.018 \times 2^{-6} \times 3^{-3}$. Step (i) follows from the facts that $\frac{1+h\varepsilon_1}{\sqrt{(1+h\varepsilon_1)^2+h^2(d_a-1+\sqrt{d_a-1})}}$ increases with ε_1 and that $\frac{1}{(1+h\varepsilon_1)^2+h^2(d_a-1+\sqrt{d_a-1})}$ decreases with ε_1 for $1 \ge \varepsilon_1 \ge 0$. Step (ii) follows from the fact that the denominator increases with h and $h < \frac{1}{\sqrt{d_a-1}}$.

Therefore, Inequality (EC.109) is proved. Combining Inequality (EC.109) and Inequality (EC.108) gives Inequality (EC.99).

Proof of Inequality (EC.125) To avoid tedious complications of dealing with small d, we first compute the probabilities for $d \le 700$ using R. Then we have $\mathbb{P}(\chi_d^2 \ge d + \sqrt{d}) > 0.15$ and $\mathbb{P}(\chi_d^2 \ge d) < 0.5$ for all $d \le 700$. Now we proceed with large d.

By definition of chi-square distribution, we have

$$\mathbb{P}(\chi_{d+1}^2 \ge \tau) = \frac{\int_{\sqrt{\tau}}^{\infty} \exp(-r^2/2) \cdot r^d \mathrm{d}r}{\int_0^{\infty} \exp(-r^2/2) \cdot r^d \mathrm{d}r}.$$

Elementary calculation shows that

$$\int_{\sqrt{\tau}}^{\infty} \exp(-r^2/2) \cdot r^d \mathrm{d}r = \begin{cases} \exp(-\frac{\tau}{2}) 2^{k-\frac{1}{2}} \left(\frac{(k-1/2)!}{(1/2)!} \int_{\frac{\tau}{2}}^{\infty} \sqrt{t} \exp(\frac{\tau}{2} - t) \mathrm{d}t + \sum_{i=0}^{k-2} (\frac{\tau}{2})^{k-\frac{1}{2}-i} \frac{(k-1/2)!}{(k-(1/2)-i)!} \right), & \text{even } d = 2k \\ \exp(-\frac{\tau}{2}) 2^k \sum_{i=0}^{k} (\frac{\tau}{2})^{k-i} \frac{k!}{(k-i)!}, & \text{odd } d = 2k+1 \end{cases}.$$

Now, we calculate the tail probabilities for even and odd degrees of freedom. Before proceeding, recall the Stirling approximation that we will use frequently:

$$\sqrt{2\pi n}(\frac{n}{e})^n < \sqrt{2\pi n}(\frac{n}{e})^n e^{(\frac{1}{12n} - \frac{1}{360n^3})} \le n! \le \sqrt{2\pi n}(\frac{n}{e})^n e^{\frac{1}{12n}} < 1.1 \times \sqrt{2\pi n}(\frac{n}{e})^n. \tag{EC.127}$$

For $d = 2k$, we have

$$\mathbb{P}(\chi_{d+1}^2 \ge \tau) = \exp(-\frac{\tau}{2}) \frac{\int_{\frac{\tau}{2}}^{\infty} \sqrt{t} \exp(\frac{\tau}{2} - t) \mathrm{d}t + \sum_{i=0}^{k-2} (\frac{\tau}{2})^{k-\frac{1}{2}-i} \frac{1/2}{(k-(1/2)-i)!}}{\int_0^{\infty} \sqrt{t} \exp(-t) \mathrm{d}t}. \tag{EC.128}$$

Let $\tau = 2k + 1 + \sqrt{2k+1}\beta$, then

$$\begin{aligned}
\mathbb{P}(\chi_{d+1}^2 \ge \tau) &= \frac{\int_{\frac{\tau}{2}}^{\infty} \sqrt{t} \exp(-t) \mathrm{d}t}{\sqrt{\pi}/2} + \frac{\exp(-\frac{\tau}{2}) \sum_{i=0}^{k-2} (\frac{\tau}{2})^{k-\frac{1}{2}-i} \frac{2^{2(k-i-1)}(k-i-1)!}{(2k-2i-1)!}}{\sqrt{\pi}/2} \\
&\overset{(a)}{\ge} \frac{20}{11\sqrt{\pi}} \exp(-\frac{\tau}{2}) \sum_{i=0}^{k-2} (\frac{\tau}{2})^{k-\frac{1}{2}-i} 2^{2(k-i-1)} \frac{((k-i-1)/e)^{k-i-1}\sqrt{k-i-1}}{((2k-2i-1)/e)^{2k-2i-1}\sqrt{2k-2i-1}} && \text{(EC.129)} \\
&\overset{(b)}{=} \frac{10}{11\sqrt{2\pi}} \exp(-\frac{\tau}{2}) \sum_{i=0}^{k-2} \exp(k-i) (\frac{\tau}{2})^{k-\frac{1}{2}-i} (\frac{2k-2i-1}{2})^{-(k-i)} \left(1 - \frac{1}{2k-2i-1}\right)^{\frac{2k-2i-1}{2}} && \text{(EC.130)} \\
&\overset{(c)}{\ge} \frac{10}{11\sqrt{2\pi}} (\frac{2}{3})^{\frac{3}{2}} \exp(-\frac{\tau}{2}) \sum_{i=0}^{k-2} \exp(k-i) (\frac{\tau}{2})^{k-\frac{1}{2}-i} (\frac{2k-2i-1}{2})^{-(k-i)} \\
&\ge \frac{10}{11\sqrt{2\pi}} (\frac{2}{3})^{\frac{3}{2}} \exp(-\frac{\tau}{2}) \sum_{i=0}^{k-2} \exp(k-i) (\frac{\tau}{2k-2i-1})^{k-\frac{1}{2}-i} \frac{1}{\sqrt{k}} \\
&= \frac{10}{11\sqrt{2\pi}} (\frac{2}{3})^{\frac{3}{2}} \exp(-k - 1/2 - \frac{\sqrt{2k+1}\beta}{2}) \sum_{i=0}^{k-2} \exp(k-i) (1 + \frac{i+1+\sqrt{2k+1}\beta/2}{k-i-1/2})^{k-\frac{1}{2}-i} \frac{1}{\sqrt{k}} \\
&\overset{(d)}{\ge} \frac{10}{11\sqrt{2\pi}} (\frac{2}{3})^{\frac{3}{2}} \exp(1/2) \sum_{i=0}^{k-2} \exp(-\frac{1}{2} \frac{(i+1+\sqrt{2k+1}\beta)^2}{k-i-1/2}) \frac{1}{\sqrt{k}}, && \text{(EC.131)}
\end{aligned}$$

where step (a) results from Stirling approximation, step (b) follows from elementary simplification, step (c) comes from the fact that $(1 - \frac{1}{x})^x$ is an increasing function, and step (d) comes from the fact that $log(1+x) \ge x - \frac{x^2}{2}$ for $x \ge 0$.

Let $\beta = 1$. Clearly, for $k \geq 100$,we have

$$\frac{1}{2}\frac{(\sqrt{k}/2+1+\sqrt{2k+1}\beta)^2}{k-\sqrt{k}/2-1/2} \leq \frac{(\sqrt{k}/2+1+\sqrt{2k+1}\beta)^2}{1.89k} < 2.16. \tag{EC.132}$$

Continue with Inequality (EC.131) gives

$$\mathbb{P}(\chi^2_{2k+1} \geq 2k+1+\sqrt{2k+1}) \geq \frac{10}{11\sqrt{2\pi}}(\frac{2}{3})^{\frac{3}{2}} \exp(1/2)\exp(-2.16)\frac{\sqrt{k}/2}{\sqrt{k}} > 0.018, \text{ for } k \geq 701. \tag{EC.133}$$

On the other hand,

$$\begin{aligned}
\mathbb{P}(\chi^2_{d+1} \geq \tau) &= \frac{\int_{\frac{\tau}{2}}^{\infty} \frac{1}{\sqrt{t}}\exp(-t)\mathrm{d}t}{\sqrt{\pi}/2} + \frac{\exp(-\frac{\tau}{2})\sum_{i=0}^{k-1}(\frac{\tau}{2})^{k-\frac{1}{2}-i}\frac{2^{2(k-i-1)}(k-i-1)!}{(2k-2i-1)!}}{\sqrt{\pi}/2} \\
&\leq \frac{\exp(-\tau/2)}{\sqrt{\pi\tau/2}} + \frac{\exp(-\frac{\tau}{2})\sum_{i=0}^{m-1}(\frac{\tau}{2})^{k-\frac{1}{2}-i}\frac{2^{2(k-i-1)}(k-i-1)!}{(2k-2i-1)!}}{\sqrt{\pi}/2} + \frac{\exp(-\frac{\tau}{2})\sum_{i=m}^{k-1}(\frac{\tau}{2})^{k-\frac{1}{2}-i}\frac{2^{2(k-i-1)}(k-i-1)!}{(2k-2i-1)!}}{\sqrt{\pi}/2} \\
&\leq \frac{\exp(-\tau/2)}{\sqrt{\pi\tau/2}} + m\frac{\exp(-\frac{\tau}{2})\ (\frac{\tau}{2})^{k-\frac{1}{2}}\frac{2^{2(k-1)}(k-1)!}{(2k-1)!}}{\sqrt{\pi}/2} \\
&\quad + \frac{\exp(-\frac{\tau}{2})(\frac{\tau}{2})^{k-\frac{1}{2}-m}\frac{2^{2(k-m-1)}(k-m-1)!}{(2k-2m-1)!}}{\sqrt{\pi}/2}\frac{1}{1-(2k-2(m+1)-1)/\tau} \\
&\overset{(i)}{\leq} \frac{\exp(-\tau/2)}{\sqrt{\pi\tau/2}} + m\frac{11}{10\sqrt{2\pi}}\exp(-\frac{\tau}{2})\exp(k)(\frac{\tau}{2})^{k-\frac{1}{2}}(k-\frac{1}{2})^{-k}\left(1-\frac{1}{2k-1}\right)^{k-\frac{1}{2}} + \\
&\quad \frac{11}{10\sqrt{2\pi}}\exp(-\frac{\tau}{2})\exp(k-m)(\frac{\tau}{2})^{k-\frac{1}{2}-m}(k-m-\frac{1}{2})^{-k+m}\left(1-\frac{1}{2k-2m-1}\right)^{k-m-\frac{1}{2}}\frac{\tau}{\tau-2k+2m+3} \\
&\leq \frac{\exp(-\tau/2)}{\sqrt{\pi\tau/2}} + m\frac{11}{10\sqrt{2\pi}}\exp(-\frac{\tau}{2})\exp(k-\frac{1}{2})(\frac{\tau}{2k-1})^{k-\frac{1}{2}}(k-\frac{1}{2})^{-1/2} + \\
&\quad \frac{11}{10\sqrt{2\pi}}\exp(-\frac{\tau}{2})\exp(k-m-\frac{1}{2})(\frac{\tau/2}{k-m-\frac{1}{2}})^{k-\frac{1}{2}-m}(k-m-\frac{1}{2})^{-1/2}\frac{\tau}{\tau-2k+2m+3} \\
&\leq \frac{\exp(-\tau/2)}{\sqrt{\pi\tau/2}} + m\frac{11}{10\sqrt{2\pi}}\exp(-\frac{\tau}{2}+k-\frac{1}{2}+\frac{\tau-2k+1}{2})(k-\frac{1}{2})^{-1/2} + \\
&\quad \frac{11}{10\sqrt{2\pi}}\exp(-\frac{\tau}{2}+k-m-\frac{1}{2}+\tau/2-k+m+1/2)(k-m-\frac{1}{2})^{-1/2}\frac{\tau}{\tau-2k+2m+3} \\
&\leq \frac{\exp(-\tau/2)}{\sqrt{\pi\tau/2}} + m\frac{11}{10\sqrt{2\pi}}(k-\frac{1}{2})^{-1/2} + \frac{11}{10\sqrt{2\pi}}(k-m-\frac{1}{2})^{-1/2}\frac{\tau}{\tau-2k+2m+3},
\end{aligned} \tag{EC.134}$$

for non-negative integer $m \leq k-1$. Step (i) follows from the Stirling approximation, and steps after step (i) repeatedly use the fact that $\log(1+x) \leq x$ for $x > -1$.

For $k \geq 100$, let $\tau = 2k+1$, $m = \lfloor\sqrt{k-\frac{1}{2}}\rfloor$. Then we have

$$\begin{aligned}
\mathbb{P}(\chi^2_{2k+1} \geq 2k+1) &\leq \exp(-100) + \frac{11}{10\sqrt{2\pi}} + \frac{11}{10\sqrt{2\pi}}\frac{2k+1}{2(\lfloor\sqrt{k-\frac{1}{2}}\rfloor+2)\sqrt{k-\frac{1}{2}-\lfloor\sqrt{k-\frac{1}{2}}\rfloor}} \\
&\leq \exp(-100) + \frac{22}{10\sqrt{2\pi}} < 0.88.
\end{aligned} \tag{EC.135}$$

For $d = 2k+1$, let $\tau = 2k+2+\sqrt{2k+2}\beta$ for some $\beta \geq 0$.

$$\mathbb{P}(\chi^2_{d+1} \geq \tau) = \exp(-\frac{\tau}{2})\sum_{i=0}^{k}(\frac{\tau}{2})^{k-i}\frac{1}{(k-i)!}$$

$$\geq \frac{10}{11}\exp(-\frac{\tau}{2})\sum_{i=0}^{k}(\frac{\tau}{2})^{k-i}\frac{1}{((k-i)/e)^{k-i}\sqrt{2\pi(k-i)}}$$
$$\geq \frac{10}{11}\exp(-\frac{\tau}{2})\sum_{i=0}^{k}\exp(k-i)(\frac{\tau}{2(k-i)})^{k-i}\frac{1}{\sqrt{2\pi k}}$$
$$\geq \frac{10}{11}\exp(-\frac{\tau}{2})\sum_{i=0}^{k}\exp(k-i)\exp\left(\frac{\tau-2(k-i)}{2(k-i)}(k-i)-\frac{1}{2}\left(\frac{\tau-2(k-i)}{2(k-i)}\right)^2(k-i)\right)\frac{1}{\sqrt{2\pi k}}$$
$$\geq \frac{10}{11}\sum_{i=0}^{k}\exp\left(-\frac{1}{2}\left(\frac{\tau-2(k-i)}{2(k-i)}\right)^2(k-i)\right)\frac{1}{\sqrt{2\pi k}}. \tag{EC.136}$$

Note that when $\beta=1$ and $k\geq 100$, for $i\leq\sqrt{k}$, we have

$$\frac{1}{2}\left(\frac{\tau-2(k-i)}{2(k-i)}\right)^2(k-i)\leq\frac{1}{2}\left(\frac{\tau-2(k-\sqrt{k})}{2(k-\sqrt{k})}\right)^2(k-\sqrt{k})=\frac{1}{2}\frac{\left(2+\sqrt{2k+2}\beta+2\sqrt{k}\right)^2}{4(k-\sqrt{k})}<1.9. \tag{EC.137}$$

Going back to Inequality (EC.136), we have

$$\mathbb{P}(\chi^2_{d+1}\geq\tau)\geq\frac{10}{11}\exp(-1.9)/\sqrt{2\pi}>0.05. \tag{EC.138}$$

On the other hand, for $\tau=2k+2$, we have

$$\mathbb{P}(\chi^2_{d+1}\geq\tau)=\exp(-\frac{\tau}{2})\sum_{i=0}^{m-1}(\frac{\tau}{2})^{k-i}\frac{1}{(k-i)!}+\exp(-\frac{\tau}{2})\sum_{i=m}^{k}(\frac{\tau}{2})^{k-i}\frac{1}{(k-i)!}$$
$$\leq m\exp(-\frac{\tau}{2})(\frac{\tau}{2})^k\frac{1}{k!}+\exp(-\frac{\tau}{2})(\frac{\tau}{2})^{k-m}\frac{1}{(k-m)!}\frac{1}{1-2(k-m)/\tau}$$
$$\leq \frac{11}{10}m\exp(-\frac{\tau}{2}+k)(\frac{\tau}{2k})^k\frac{1}{\sqrt{2\pi k}}+\frac{11}{10}\exp(-\frac{\tau}{2}+k-m)(\frac{\tau}{2(k-m)})^{k-m}\frac{1}{\sqrt{2\pi(k-m)}}\frac{1}{1-2(k-m)/\tau}$$
$$=\frac{11}{10}m\exp(-1)(\frac{k+1}{k})^k\frac{1}{\sqrt{2\pi k}}+\frac{11}{10}\exp(-1-m)(\frac{k+1}{k-m})^{k-m}\frac{1}{\sqrt{2\pi(k-m)}}\frac{k+1}{1+m}$$
$$\leq \frac{11}{10}m\frac{1}{\sqrt{2\pi k}}+\frac{11}{10}\frac{1}{\sqrt{2\pi(k-m)}}\frac{k+1}{m+1}, \tag{EC.139}$$

for non-negative integer $m\leq k$. Let $m=\lfloor k\rfloor$, for $k\geq 100$, we have

$$\mathbb{P}(\chi^2_{d+1}\geq 2k+2)\leq\frac{11}{10}\frac{1}{\sqrt{2\pi}}+\frac{11}{10}\frac{1}{\sqrt{2\pi(k-\sqrt{k})}}\frac{k+1}{\sqrt{k}}<0.91. \tag{EC.140}$$

EC.3. Details on Simulation and Real Case Studies

In this section, we provide more details on the tuning parameters of different algorithms in Simulation II in Section EC.3.1 and on the two case studies in Sections EC.3.2 and EC.3.3.

EC.3.1. Details on Simulation II

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 initialization steps as $t_{init}=100$; the initial penalization parameter $\lambda_0=\|\frac{1}{2t_{init}L}\sum_{i=1}^{t_{init}}\sum_{\ell=1}^{L}|\boldsymbol{a}_i^\top\widehat{\boldsymbol{\Theta}}_{t_{init}}\boldsymbol{x}_{i,\ell}-y_{i,\ell}|\boldsymbol{x}_{i,\ell}\boldsymbol{a}_i^\top\|_{\text{op}}$; and the exploration parameter $h=0.1$.

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

Lasso Bandit (Bastani and Bayati 2020). There are several 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.

G-ESTT (Kang et al. 2022). We implement the algorithm sketched in Appendix H of Kang et al. (2022) as a potential extension of their main algorithm to contextual setting. Their Theorem 4.3 suggests a choice of T_1 for which they documented good performance using their main algorithm, but not for the contextual bandits of interest here. To best implement their idea in our setting, we note that our moderate dimensions are already relatively large for their algorithm and our typical time scope $T = 1000$ is far too small for what is required in their algorithm. With this issue in mind, we set $T_1 = \sqrt{rT\log((d_1+d_2)/\delta)}/D_{rr}$ instead where $d_1 = d_x$, $d_2 = d_a$, $\delta = 0.01$ as in their setup, $D_{rr} = 0.5$ is the smallest non-zero singular value.

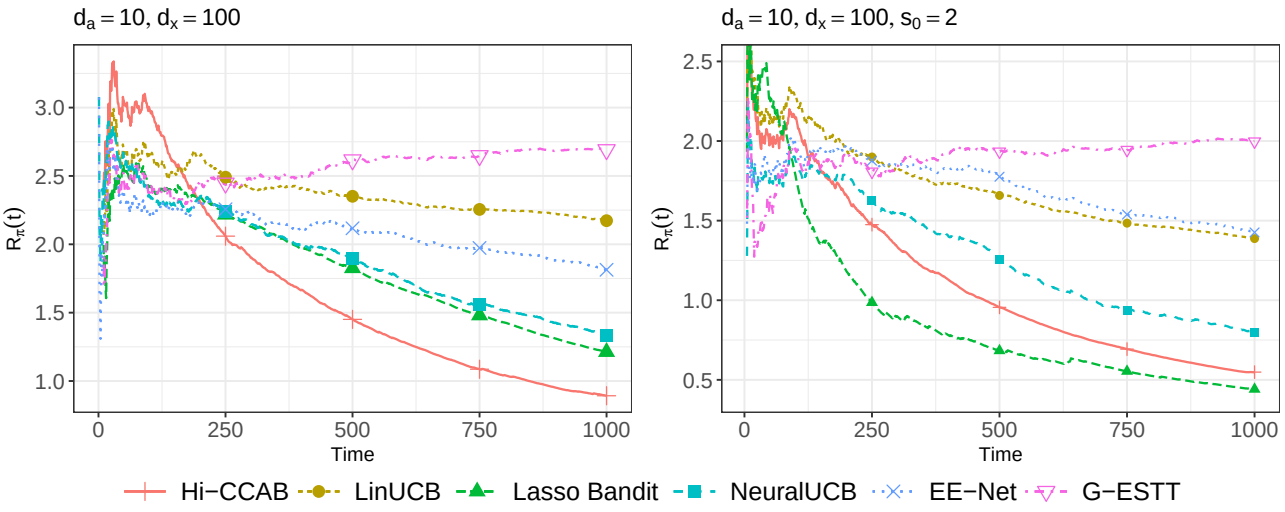


Figure EC.1 Expected average regret.

In Kang et al. (2022), the main algorithm with theoretical guarantees is designed for low-rank matrix bandit but not for contextual bandit. In Appendix H, the author sketched an extension to the contextual

setting, but their theory and numerical validation only apply to non-contextual bandits. We applied the modified G-ESTT to the contextual bandit setting where $d_a = 10, d_x = 100$ and the sparse setting with $d_a = 10, d_x = 100, s_0 = 2$ as in Section 5.2. Figure EC.1 shows that the modified G-ESTT does not perform well compared to other contextual bandit algorithms. One reason can be the following. They adopt the explore-then-commit algorithm and their initialization step is required to be of the order $\sqrt{d_1 d_2 r T}$. In our simulation setting, the dimension of covariate $d_2 = d_x$ is high and therefore their algorithm will require a large T_1 to perform well. Therefore, the modified G-ESTT does not perform well in high-dimensional settings (note that the dimensions in their simulations is of the order 10), especially when T is relatively small. In addition, the computational complexity of the modified G-ESTT is high compared to other methods. We tried to run the modified G-ESTT for other settings as in Section 5.2 with larger d_x but it would have taken too long and we do not expect a better performance of the modified G-ESTT in higher dimension settings given the above-mentioned claim.

EC.3.2. More details on Case Study I

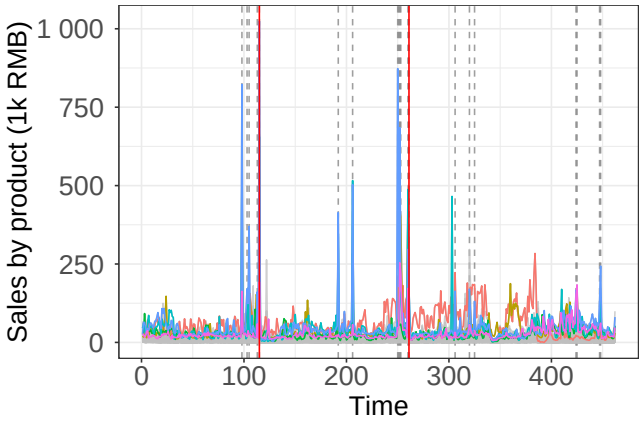
In this section, we provides more background information on Case Study I and additional numerical results.

Figure EC.2a 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 EC.2b 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 EC.2c 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 EC.2d 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.

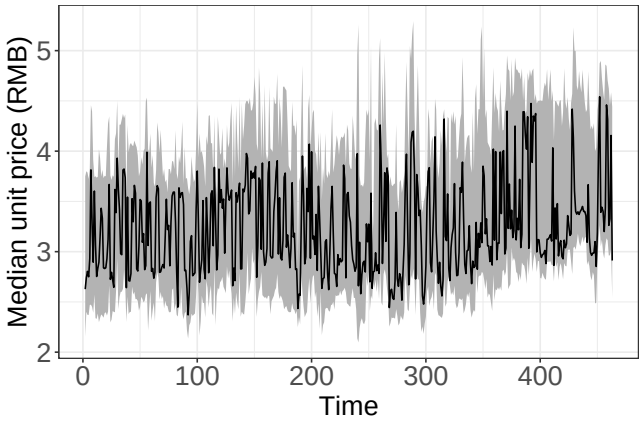
To check our model assumption (3) on the data, Figure EC.3 shows the hold-out-sample prediction of the sales versus the real sales. The predicted sale at each time point t is the predicted total sales across $L = 31$ locations based on $\widehat{\boldsymbol{\Theta}}_{-t}$ estimated from all the data except for data at time t, i.e.,

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

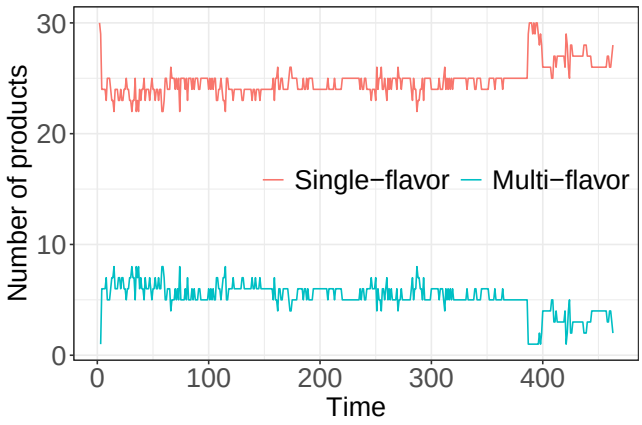
where $\widehat{\boldsymbol{\Theta}}_{-t} = \arg\min_{\boldsymbol{\Theta}} \sum_{i=1, i\neq t}^{T} \sum_{\ell=1}^{L} \left(\boldsymbol{a}_i^\top \boldsymbol{\Theta} \boldsymbol{x}_{i,\ell} - r_{i,\ell}\right)^2 + \lambda |||\boldsymbol{\Theta}|||_{\text{nuc}}$. As shown in Figure EC.3, the real sales and the out-of-sample predicted sales follow quite closely across time and the out-of-sample prediction error rate $\sum_{i=1}^{T} \sum_{\ell=1}^{L} (y_{t,\ell} - \hat{y}_{t,\ell})^2 / \sum_{i=1}^{T} \sum_{\ell=1}^{L} y_{t,\ell}^2 = 0.07$, which indicates that both our model and estimation are reasonable.



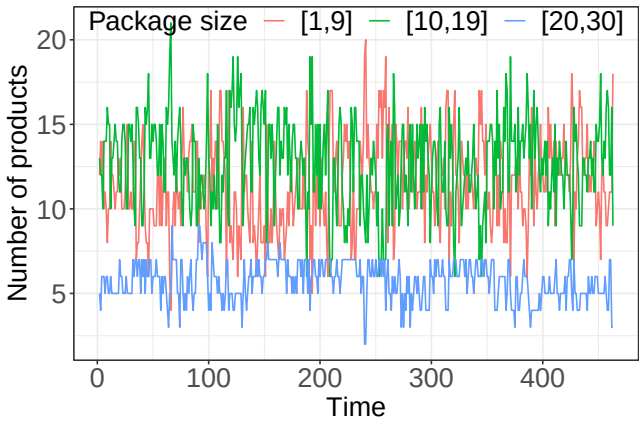
(a) Total sales by product in 1k RMB.



(b) Median price of products.



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



(d) Number of products with various package sizes.

Figure EC.2 Summary of the products.

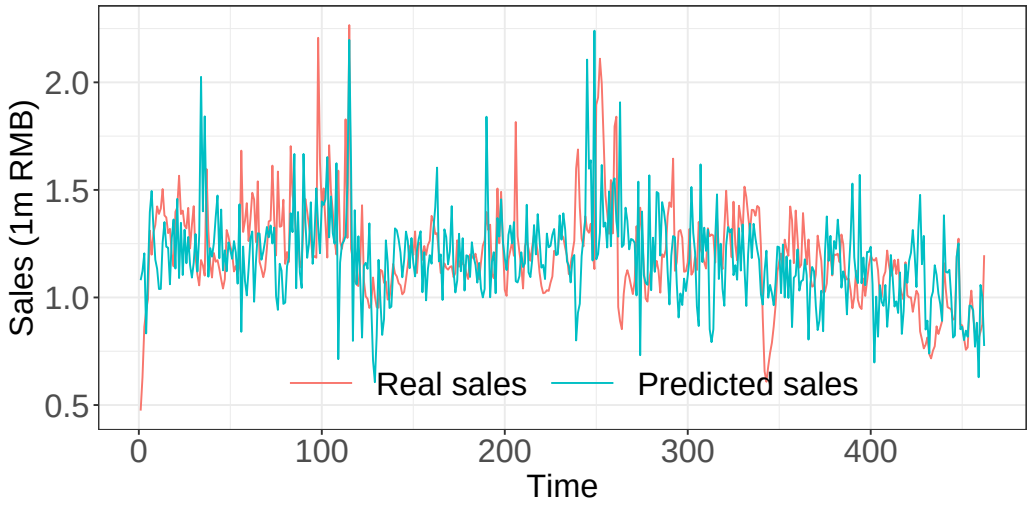
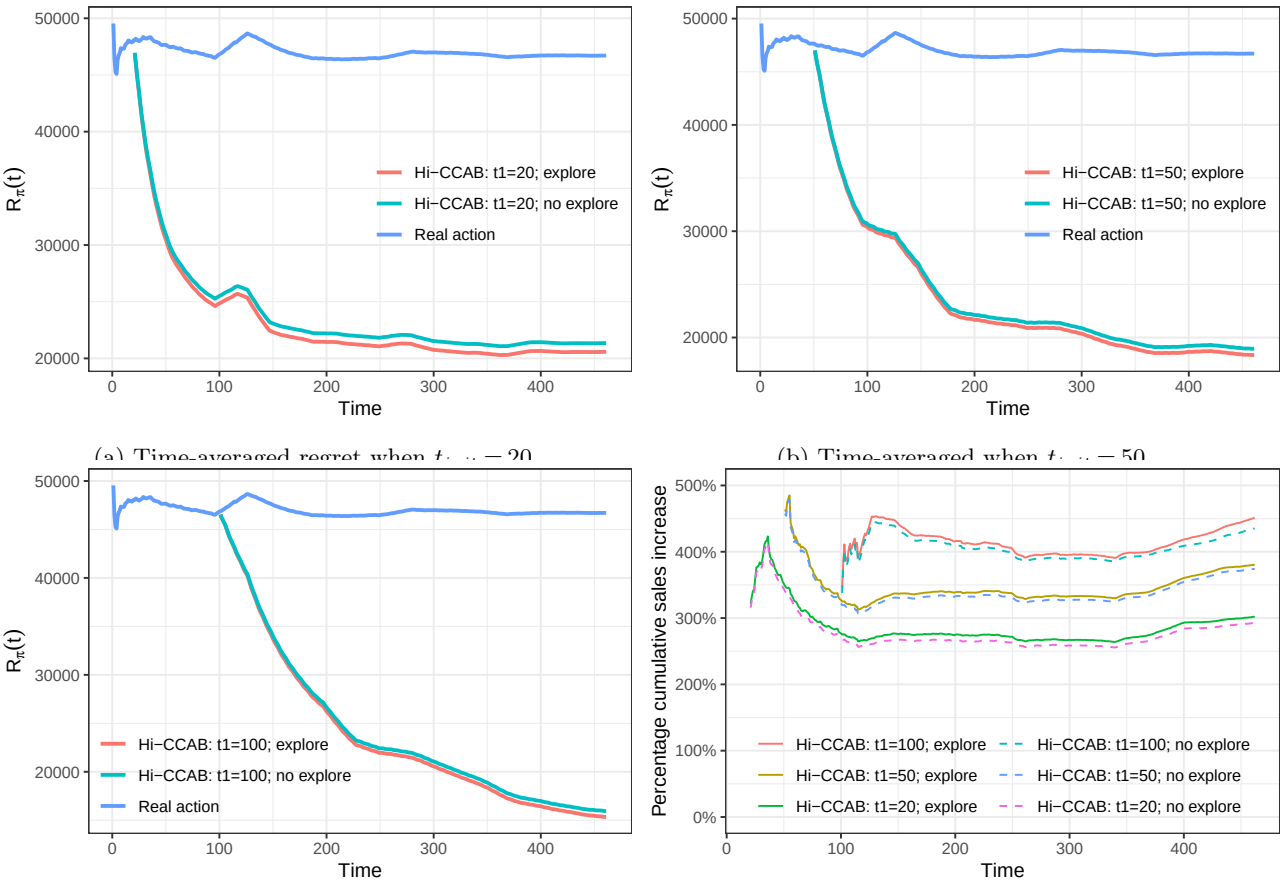


Figure EC.3 Real sales vs predicted sales.



(a) Time-averaged regret when $t_{init} = 20$.

(b) Time-averaged when $t_{init} = 50$.

(c) Time-averaged when $t_{init} = 100$.

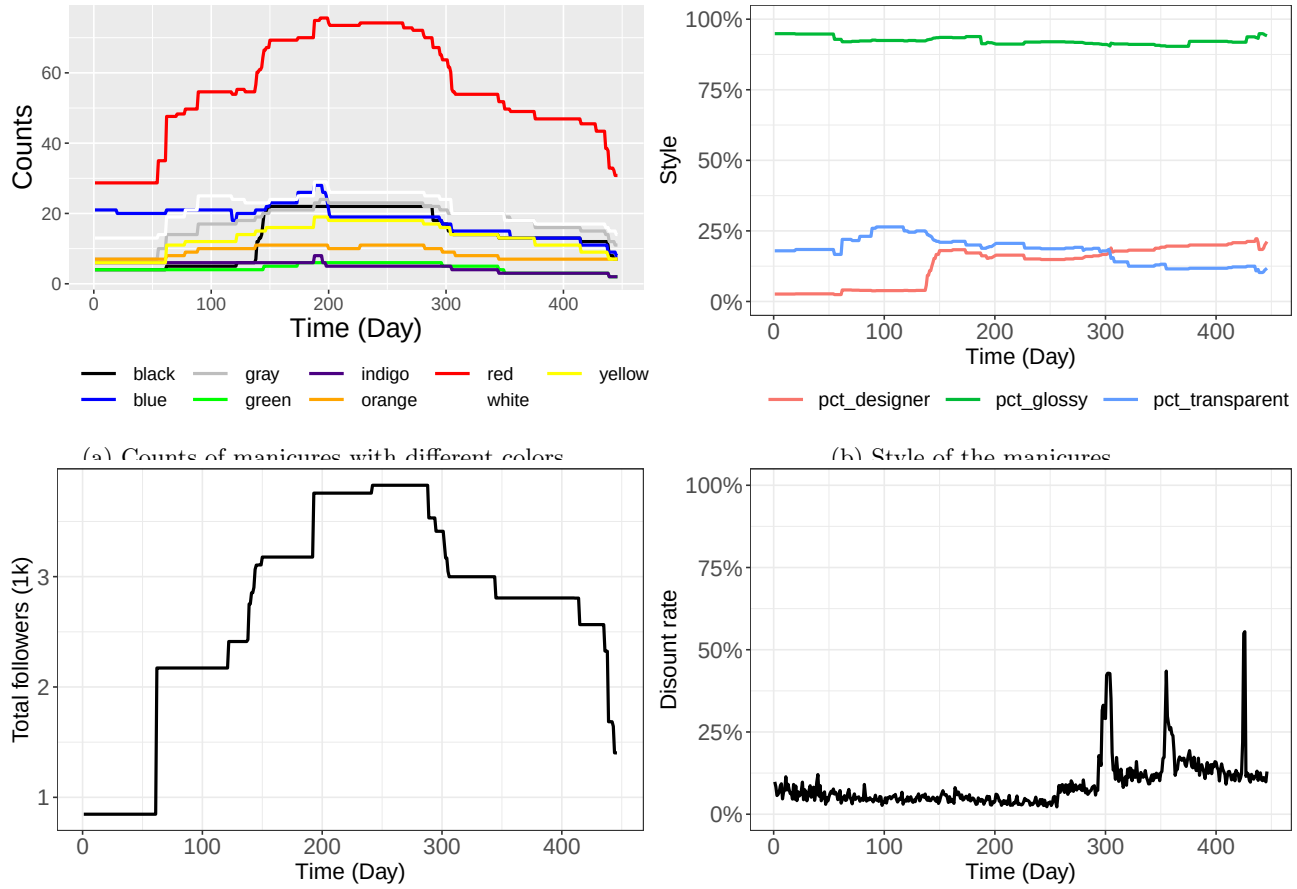
(d) Percentage cumulative sales gain.

Figure EC.4 Performance of `Hi-CCAB` with different initialization times t_{init} and with and without exploration.

Further simulation results. We first detail how we ran the simulation and then provide more simulation results.

We run 100 trials, in each of which we set $t_{init} = 100$ for the initialization step and λ_0 according to Algorithm 1 to estimate $\widehat{\boldsymbol{\Theta}}_{t_{init}}$; and then at each time $t = t_{init} + 1, \ldots, T$, we follow Algorithm 1 to make an assortment-pricing decide $\boldsymbol{a}_t$ given covariate $\boldsymbol{x}_t$. After determining $\boldsymbol{a}_t$, we generate the sales $y_{t,\ell}$ for $\ell = 1, \ldots, L = 31$ locations according to model (3) based on the pseudo-truth-model with $(\boldsymbol{\Theta}, \sigma)$. We further compare the performance of the assortment-pricing policy with exploration and without exploration and with different initialization time t_{init}. Each setup is simulated 100 times.

Figures EC.4a-EC.4b show the time-averaged regret and Figure EC.4d show percentage gain in cumulative sales when $t_{init} = 20, 50, 100$ 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 $\boldsymbol{\Theta}$ and thus helps with the performance in a short time windows. As time goes by, all of the time-averaged regrets converge to zero and the percentage gain in cumulative sales should converge.



(a) Counts of manicures with different colors.

(b) Style of the manicures.

(c) Total number of follows of designers.

(d) Discount rate.

Figure EC.5 Summary of the products.

EC.3.3. More details on Case Study II

In this section, we provides more background information on Case Study II and additional numerical results.

Figure EC.5a shows the total counts of manicures featuring various colors. Note that one manicure can potentially use multiple colors. Red was the most prevalent color, followed by white, gray, black, blue, yellow, and orange. The ranking, apart from red, was determined by sales volumes from previous periods. Figure EC.5b shows the style of the manicures, i.e., percentages of designer, glossy, and transparent manicures respectively. The count of designer manicures surged in around June 2020 after the total profits increased and then plateaued as shown in Figure EC.6. Figure EC.5c shows the total number of followers on Instagram of the designers and Figure EC.5d presents the discount rate, calculated as the percentage of total daily discount amounts. Notably, discount peaks are observed around Thanksgiving, New Year's, and April Fool's Day.

We check our model assumption (3) on the data similar to Case Study I as detailed in Section EC.3.2. Figure EC.3 shows the hold-out-sample prediction of the profits versus the real profits. The real sales and the out-of-sample predicted sales follow quite closely over time and the out-of-sample prediciton error rate is around 8%, which again indicates that both our model and estimation are reasonable.

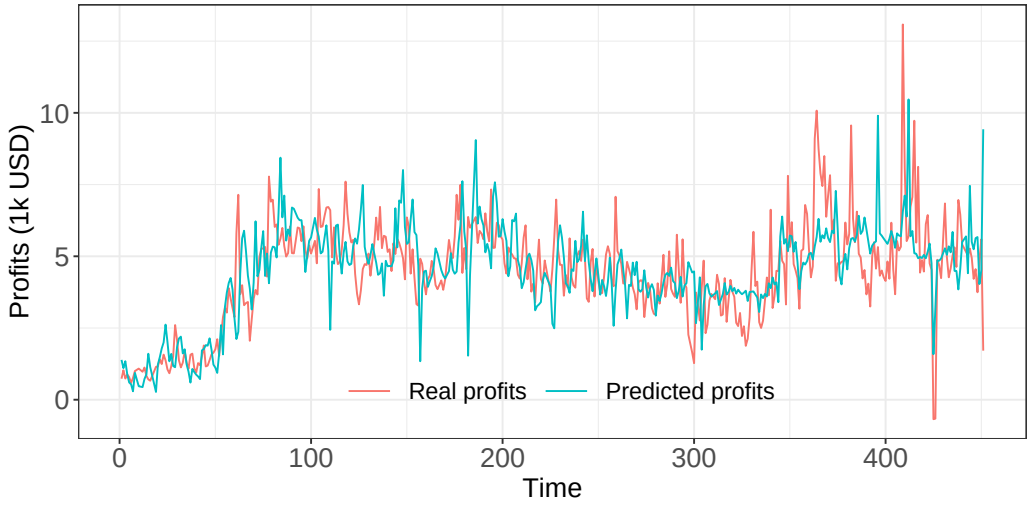


Figure EC.6 Real profits vs predicted profits.