Journal of Machine Learning Research 21 (2020) 1-44 Submitted 12/19; Revised 10/20; Published 10/20

Dynamic Assortment Optimization with Changing Contextual Information*

Xi Chen XC13@STERN.NYU.EDU
Leonard N. Stern School of Business
New York University
New York, NY 10012, USA

Yining Wang YINING.WANG@WARRINGTON.UFL.EDU
Warrington College of Business
University of Florida
Gainesville, FL 32611, USA

Yuan Zhou YUANZ@ILLINOIS.EDU
Department of Industrial and Enterprise Systems Engineering
University of Illinois at Urbana-Champaign
Urbana-Champaign, IL 61801, USA

Editor: Ambuj Tewari

Abstract

In this paper, we study the dynamic assortment optimization problem over a finite selling season of length T. At each time period, the seller offers an arriving customer an assortment of substitutable products under a cardinality constraint, and the customer makes the purchase among offered products according to a discrete choice model. Most existing work associates each product with a real-valued fixed mean utility and assumes a multinomial logit choice (MNL) model. In many practical applications, feature/contextual information of products is readily available. In this paper, we incorporate the feature information by assuming a linear relationship between the mean utility and the feature. In addition, we allow the feature information of products to change over time so that the underlying choice model can also be non-stationary. To solve the dynamic assortment optimization under this changing contextual MNL model, we need to simultaneously learn the underlying unknown coefficient and make the decision on the assortment. To this end, we develop an upper confidence bound (UCB) based policy and establish the regret bound on the order of $\widetilde{O}(d\sqrt{T})$, where d is the dimension of the feature and $\widetilde{O}$ suppresses logarithmic dependence. We further establish a lower bound $\Omega(d\sqrt{T}/K)$, where K is the cardinality constraint of an offered assortment, which is usually small. When K is a constant, our policy is optimal up to logarithmic factors. In the exploitation phase of the UCB algorithm, we need to solve a combinatorial optimization problem for assortment optimization based on the learned information. We further develop an approximation algorithm and an efficient greedy heuristic. The effectiveness of the proposed policy is further demonstrated by our numerical studies.

Keywords: Dynamic assortment optimization, regret analysis, contextual information, bandit learning, upper confidence bounds

*. Author names listed in alphabetical order.

1. Introduction

In operations, an important research problem facing a retailer is the selection of products/advertisements for display. For example, due to the limited shelf space, stocking restrictions, or available slots on a website, the retailer needs to carefully choose an assortment from the set of N substitutable products. In an assortment optimization problem, choice model plays an important role since it characterizes a customer's choice behavior. However, in many scenarios, customers' choice behavior (e.g., mean utilities of products) is not given as *a priori* and cannot be easily estimated due to the insufficiency of historical data. This motivates the research of dynamic assortment optimization, which has attracted a lot of attentions from the revenue management community in recent years. A typical dynamic assortment optimization problem assumes a finite selling horizon of length T with a large T. At each time period, the seller offers an assortment of products (with the size upper bounded by K) to an arriving customer. The seller observes the customer's purchase decision, which further provides useful information for learning utility parameters of the underlying choice model. The multinomial logit model (MNL) has been widely used in dynamic assortment optimization literature, see, e.g., Caro and Gallien (2007); Rusmevichientong et al. (2010); Saure and Zeevi (2013); Agrawal et al. (2017); Chen and Wang (2018); Wang et al. (2018); Agrawal et al. (2019); Chen et al. (2019, 2020b).

In the age of e-commerce, side information of products is widely available (e.g., brand, color, size, texture, popularity, historical selling information), which is important in characterizing customers' preferences for products. Moreover, some features are not static and could change over time (e.g., popularity score or ratings). The feature/contextual information of products will facilitate accurate assortment decisions that are tailored to customers' preferences. In particular, we assume at each time $t = 1, \ldots, T$, each product j is associated with a d-dimensional feature vector $v_{tj} \in \mathbb{R}^d$, where products' feature vectors can change over time. To incorporate the feature information, following the classical conditional logit model (McFadden, 1973), we assume that the mean utility of product j at time t (denoted by u_{tj}) obeys a linear model

$$u_{tj} = v_{tj}^\top \theta_0. \tag{1}$$

Here, $\theta_0 \in \mathbb{R}^d$ is the unknown coefficient to be learned. Based on this linear structure of the mean utility, we adopt the MNL model as the underlying choice model (see Section 2 and Eq. (3) for more details), and we consider the common setting in which the length of the time horizon T and the total number of products N are much larger than d. As compared to the standard MNL, this *changing contextual MNL* model not only incorporates contextual information but also allows the utility to evolve over time. The changing utility is an attractive property as it captures the reality in many applications but also brings new technical challenges in learning and decision-making. For example, in existing works of (Agrawal et al., 2019) for plain MNL choice models, upper confidence bounds are constructed by providing the same assortment repetitively to incoming customers until a no-purchase activity is observed. Such an approach, however, can no longer be applied to MNL with changing contextual information as the utility parameters of products constantly evolve with time. To overcome such challenges, we propose a policy that performs optimization at every single time period, without repetitions of assortments in general.

Note that in our model in Eq. (1), the feature vectors $\{v_{tj}\}$ of a specific product j can evolve with time t. Indeed, the features of a certain product over the T selling periods can be roughly divided into two types: those that do not evolve with time, such as the color, size, texture of the product, and those that vary from time to time, such as popularity, ratings, and seasonality. The varying features of a product certainly have profound (time-varying) implications on potential customers' decisions. For example, a skirt would see much higher sales volumes during summer compared to winter, and a product available for online sales is also likely to see increased sales volume if certain important figure has recommended the product. Therefore, features such as the season or whether it is holiday reason and the current rating of a product are typical time-evolving features, and can be easily affect customers' purchase decision.

To facilitate our analysis, we assume that the contextual vectors $\{v_{tj}\}$ are i.i.d. distributed with respect to an unknown underlying distribution that is generally well behaved. We also assume that both $\{v_{tj}\}$ and the regression model θ_0 are bounded in ℓ_2 norm. Details and discussion of the assumptions and the possible relaxations are made in Sec. 3.1.

Our model also allows the revenue for each product j to change over time. In particular, we associate the revenue parameter r_{tj} for the product j at time t.

This model generalizes the widely adopted (generalized) linear contextual bandit from machine learning literature (see, e.g., Filippi et al. (2010); Chu et al. (2011); Abbasi-Yadkori et al. (2011); Agrawal and Goyal (2013); Li et al. (2017) and references therein) in a non-trivial way since the MNL cannot be written in a generalized linear model form (when an assortment contains more than one product, see Section 1.1 for more details).

Given this contextual MNL choice model, the key challenge is how to design a policy that simultaneously learns the unknown coefficient θ_0 and sequentially makes the decision on offered assortment. The performance of a dynamic policy is usually measured by the *regret*, which is defined as the gap between the expected revenue generated by the policy and the oracle expected revenue when θ_0 (and thus the mean utilities) is known as a priori.

The first contribution of the paper is the construction of an upper confidence bound (UCB) policy. Our UCB policy is based on the maximum likelihood estimator (MLE) and thus is named MLE-UCB. Although UCB has been a well-known technique for bandit problems, how to adopt this high-level idea to solve a problem with specific structures certainly requires technical innovations (e.g., how to build a confidence interval varies from one problem to another). In particular, we propose a local-MLE idea, which divides the entire time horizon into two stages. The first stage is a *pure exploration stage* in which assortments are offered uniformly at random and a *"pilot MLE"* is computed based on the observed purchase actions. As we will show in Lemma 3, this pilot estimator serves as a good initial estimator of θ_0. After the exploration phase, the MLE-UCB enters the *simultaneous learning and decision-making phase*. We carefully construct an upper confidence bound of the expected revenue when offering an assortment. The added interval is based on the Fisher's information matrix of the computed MLE from the previous step. Then we solve a combinatorial optimization problem to search the assortment that maximizes the upper confidence bound. By observing the customer's purchase action based on the offered assortment, the policy updates the estimated MLE. In this update, we propose to compute a *"local MLE"*, which requires the solution to be close enough to our pilot estimator. The local MLE plays an important role in the analysis of our MLE-UCB policy since it guarantees

that the obtained estimator at each time period is also close to the unknown true coefficient θ_0.

Under some mild assumptions on features and coefficients, we are able to establish a regret bound $\widetilde{O}(d\sqrt{T})$, where the $\widetilde{O}$ notation suppresses logarithmic dependence on T, K (cardinality constraint), and some other problem dependent parameters[1]. One remarkable aspect of our regret bound is that our regret has no dependence on the total number of products N (not even in a logarithmic factor). This makes the result attractive to online applications where N is large (e.g., online advertisement). Moreover, it is also worthwhile noting the dependence of K is only through a logarithmic term.

Our second contribution is to establish the lower bound result $\Omega(d\sqrt{T}/K)$. When the maximum size of an assortment K is small (which usually holds in practice), this result shows that our policy is almost optimal.

Moreover, at each time period in the exploitation phase, our UCB policy needs to solve a combinatorial optimization problem, which searches for the best assortment (under the cardinality constraint) that minimizes the upper confidence bound of the expected revenue. Given the complicated structure of the upper confidence bound, there is no simple solution for this combinatorial problem. When K is small and N is not too large, one can directly search over all the possible sets with the size less than or equal to K. In addition to the solution of solving the combinatorial optimization exactly, the third contribution of the work is to provide an approximation algorithm based on dynamic programming that runs in polynomial time with respect to N, K, T. Although the proposed approximation algorithm has a theoretical guarantee, it is still not efficient for dealing with large-scale applications. To this end, we further describe a computationally efficient greedy heuristic for solving this combinatorial optimization problem. The heuristic algorithm is based on the idea of local search by greedy swapping, with more details described in Sec. 5.2.

1.1 Related work

Due to the popularity of data-driven revenue management, dynamic assortment optimization, which adaptively learns unknown customers' choice behavior, has received increasing attention in the past few years. Motivated by fast-fashion retailing, the work by Caro and Gallien (2007) first studied dynamic assortment optimization problem, but it makes a strong assumption that the demands for different product are independent. Recent works by Rusmevichientong et al. (2010); Saure and Zeevi (2013); Agrawal et al. (2017); Chen and Wang (2018); Wang et al. (2018); Agrawal et al. (2019) incorporated MNL models into dynamic assortment optimization and formulated the problem into a online regret minimization problem. In particular, for capacitated MNL, Agrawal et al. (2019) and Agrawal et al. (2017) proposed UCB and Thompson sampling techniques and established the regret bound $\widetilde{O}(\sqrt{NT})$ (when $T \gg N^2$). Chen and Wang (2018) further established a matching lower bound of $\Omega(\sqrt{NT})$. While our proposed algorithm is also based on the UCB framework, there are significant differences between the *contextual* MNL-Bandit model and the

1. For the ease of presentation in the introduction, we only present the dominating term under the common scenario that the selling horizon T is larger than the dimensionality d and the cardinality constraint K. Please refer to Theorem 1 for a more explicit expression of the obtained regret.

plain one studied in Agrawal et al. (2019, 2017). We will explain the technical details of the differences at the end of this section.

It is also interesting to compare our regret to the bound for the standard MNL case. In practice, the number of features d extracted for a product is usually much smaller than the total number of products N. When N is much larger than d, by incorporating the contextual information, the regret reduces from $\widetilde{O}(\sqrt{NT})$ to $\widetilde{O}(d\sqrt{T})$. The latter one only depends on d and is completely independent of the total number of products N, which also demonstrates the usefulness of the contextual information.

We also note that to highlight our key idea and focus on the balance between learning of θ_0 and revenue maximization, we study the stylized dynamic assortment optimization problems following the existing literature (Rusmevichientong et al., 2010; Saure and Zeevi, 2013; Agrawal et al., 2017; Chen and Wang, 2018; Agrawal et al., 2019), which ignore operations considerations such as price decisions and inventory replenishment.

There is another line of recent research on investigating personalized assortment optimization. By incorporating the feature information of each arriving customer, both the static and dynamic assortment optimization problems are studied in Chen et al. (2020a) and Cheung and Simchi-Levi (2017), respectively. It is also worth noting that our model is different from the personalized MNL model proposed by Cheung and Simchi-Levi (2017), in which each product j is associated with a fixed but unknown coefficient $\theta(j)$ and each arriving customer at time t with an observable feature vector x_t. In contrast, our model considers feature vectors on products rather than on customers since products' features are easier to extract and involve less privacy issues in some applications. In addition, the developed techniques in our work and Cheung and Simchi-Levi (2017) are different, Our policy is based on UCB, while the policy in Cheung and Simchi-Levi (2017) is based on Thompson sampling.

Furthermore, other research studies personalized assortment optimization in an adversarial setting rather than stochastic setting. For example, Golrezaei et al. (2014); Chen et al. (2016) assumed that each customer's choice behavior is known, but that the customers' arriving sequence (or customers' types) can be adversarially chosen and took the inventory level into consideration. Since the arriving sequence can be arbitrary, there is no learning component in the problem and both Golrezaei et al. (2014) and Chen et al. (2016) adopted the competitive ratio as the performance evaluation metric.

Another field of related research is the *contextual bandit* literature, in which the linear contextual bandit has been widely studied as a special case (see, e.g., Dani et al. (2008); Rusmevichientong and Tsitsiklis (2010); Chu et al. (2011); Abbasi-Yadkori et al. (2011); Agrawal and Goyal (2013) and references therein). Some recent work extends the linear contextual bandit to *generalized linear bandit* (Filippi et al., 2010; Li et al., 2017), which assumes a generalized linear reward structure. In particular, the reward r of pulling an arm given the observed feature vector of this arm x is modeled by

$$\mathbb{E}[r|x] = \sigma(x^\top \theta_0), \tag{2}$$

for an unknown linear model θ_0 and a known link function $\sigma : \mathbb{R} \to \mathbb{R}$. For example, for a linear contextual bandit, σ is the identity mapping, i.e., $\mathbb{E}[r|x] = \sigma(x^\top \theta_0)$. For the logistic contextual bandit, we have $r \in \{0, 1\}$ and $\Pr(r = 1|x) = \frac{\exp(x^\top \theta_0)}{1+\exp(x^\top \theta_0)}$. In a standard

generalized linear bandit problem (see, e.g., Li et al. (2017)) with N arms, it is assumed that a context vector v_{tj} is revealed at time t for each arm $j \in [N]$. Given a selected arm $i_t \in [N]$ at time t, the expected reward follows Eq. (2), i.e., $\mathbb{E}[r_t|v_{t,i_t}] = \sigma(v_{t,i_t}^\top \theta_0)$. At first glance, our contextual MNL model is a natural extension of the generalized linear bandit to the MNL choice model. However, when the size of an assortment $K \geq 2$, the contextual MNL *cannot* be written in the form of Eq. (2) and the denominator in the choice probability (see Eq. (3) in the next section) has a more complicated structure. Therefore, our problem is technically *not* a generalized linear model and is therefore more challenging. Moreover, in contextual bandit problems, only one arm is selected by the decision-maker at each time period. In contrast, each action in an assortment optimization problem involves a set of items, which makes the action space more complicated.

We would also like to briefly remark on the technical novelty of our paper, comparing to the previous works on plain MNL-Bandit models. While it is true that UCB-type policies have already been applied to dynamic assortment planning in the literature, our policy utilizing UCB is considerably more sophisticated than the one introduced in Agrawal et al. (2019) for plain MNL models. This is because, in a contextual MNL-Bandit model, the estimation/prediction errors of revenues are heterogeneous and have to be very carefully computed, in our case through careful approximation of the Fisher's information matrix and a pilot estimator.

To be more specific, in a plain MNL-Bandit model, the lengths of confidence intervals of a specific assortment $S \subseteq [N]$ only depend on the number of times the products in S are offered in prior time periods. However, in a contextual MNL-Bandit model, the lengths of confidence levels depend on the contextual information v_{ti} at time t, which in turn involve all realized purchases from previous selling periods through the sample Fisher's information term. Therefore, this makes our UCB-type algorithm much more sophisticated than the UCB algorithms for plain MNL-Bandit models.

We also note that the elliptical potential lemma is an important component of virtually all linear/generalized linear contextual bandit research. Therefore, we also use this important lemma to prove the upper regret bounds of our proposed policies.

1.2 Notations and paper organization

Throughout the paper, we adopt the standard asymptotic notations. In particular, we use $f(\cdot) \lesssim g(\cdot)$ to denote that $f(\cdot) = O(g(\cdot))$. Similarly, by $f(\cdot) \gtrsim g(\cdot)$, we denote $f(\cdot) = \Omega(g(\cdot))$. We also use $f(\cdot) \asymp g(\cdot)$ for $f(\cdot) = \Theta(g(\cdot))$. Throughout this paper, we will use $C_0, C_1, C_2, \ldots$ to denote universal constants. For a vector v and a matrix M, we will use $\|v\|_2$ and $\|M\|_{\text{op}}$ to denote the vector ℓ_2-norm and the matrix spectral norm (i.e., the maximum singular value), respectively. Moreover, for a real-valued symmetric matrix M, we denote the maximum eigenvalue and the minimum eigenvalue of M by $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$, respectively, and define $\|v\|_M^2 := v^T M v$ for any given vector v. For a given integer N, we denote the set $\{1, \ldots, N\}$ by $[N]$.

The rest of the paper is organized as follows. In Section 2, we introduce the mathematical formulation of our models and define the regret. In Section 3, we describe the proposed MLE-UCB policy and provide the regret analysis. The lower bound result is provided in Section 4. In Section 5, we investigate the combinatorial optimization problem in MLE-

UCB and propose the approximation algorithm and greedy heuristic. The multivariate case of the approximation algorithm is relegated to the appendix. In Section 6, we provide the numerical studies. The conclusion and future directions are discussed in Section 7. Some technical proofs are provided in the appendix.

2. The problem setup

There are N items, conveniently labeled as $1, 2, \cdots, N$. At each time t, a set of time-sensitive "feature vectors" $v_{t1}, v_{t2}, \cdots, v_{tN} \in \mathbb{R}^d$ and revenues $r_{t1}, \cdots, r_{tN} \in [0,1]$ are observed, reflecting time-varying changes of items' revenues and customers' preferences. A retailer, based on the features $\{v_{ti}\}_{i=1}^N$ and previous purchasing actions, picks an assortment $S_t \subseteq [N]$ under the cardinality constraint $|S_t| \leq K$ to present to an incoming customer; the retailer then observes a purchasing action $i_t \in S_t \cup \{0\}$ and collects the associated revenue r_{i_t} of the purchased item (if $i_t = 0$ then no item is purchased and zero revenue is collected).

We use an MNL model with features to characterize how a customer makes choices. Let $\theta_0 \in \mathbb{R}^d$ be an *unknown* time-invariant coefficient. For any $S \subseteq [N]$, the choice model $p_{\theta_0,t}(\cdot|S)$ is specified as (let $r_0 = 0$ and $v_{t0} = 0$)

$$p_{\theta_0,t}(j|S) = \frac{\exp\{v_{tj}^\top \theta_0\}}{1 + \sum_{k\in S} \exp\{v_{tk}^\top \theta_0\}} \qquad \forall j \in S \cup \{0\}. \tag{3}$$

For simplicity, in the rest of the paper we use $p_{\theta,t}(\cdot|S)$ to denote the law of the purchased item i_t conditioned on given assortment S at time t, parameterized by the coefficient $\theta \in \mathbb{R}^d$. The expected revenue $R_t(S)$ of assortment $S \subseteq [N]$ at time t is then given by

$$R_t(S) := \mathbb{E}_{\theta_0,t}[r_{tj}|S] = \frac{\sum_{j\in S} r_{tj} \exp\{v_{tj}^\top \theta_0\}}{1 + \sum_{j\in S} \exp\{v_{tj}^\top \theta_0\}}. \tag{4}$$

Note that throughout the paper, we use $\mathbb{E}_{\theta_0,t}[\cdot|S]$ to denote the expectation with respect to the choice probabilities $p_{\theta_0,t}(j|S)$ defined in Eq. (3).

Our objective is to design policy π such that the regret

$$\text{Regret}(\{S_t\}_{t=1}^T) = \mathbb{E}^\pi \sum_{t=1}^T R_t(S_t^*) - R_t(S_t) \quad \text{where } S_t^* = \arg\max_{S\subseteq[N],|S|\leq K} R_t(S) \tag{5}$$

is minimized. Here, S_t^* is an optimal assortment chosen when the full knowledge of choice probabilities is available (i.e., θ_0 is known).

3. An MLE-UCB policy and its regret

We propose an MLE-UCB policy, described in Algorithm 1.

The policy can be roughly divided into two phases. In the first *pure exploration* phase, the policy selects assortments of size K uniformly at random. The objective of the pure exploration is to establish a "pilot" estimator of the unknown coefficient θ_0, i.e., a good initial estimator for θ_0. In the second phase, we use a UCB-type approach that selects S_t as the assortment maximizing an *upper bound* $\overline{R}_t(S_t)$ of the expected revenue $R_t(S_t)$.

1 Initialize: $T_0 = \lceil \sqrt{T} \rceil$, $\tau = T^{-1/8}$;

2 Pure exploration: for $t = 1, \cdots, T_0$, pick S_t consisting of K products sampled uniformaly at random from $\{1, 2, \cdots, N\}$, and record purchasing actions $(i_1, \cdots, i_{T_0})$;

3 Compute a pilot estimator using global MLE:
$\theta^* \in \arg\max_{\|\theta\|_2 \le 1} \sum_{t'=1}^{T_0} \log p_{\theta,t}(i_{t'}|S_{t'})$;

4 **for** $t = T_0 + 1$ *to* T **do**

5 Observe revenue parameters $\{r_{tj}\}_{j=1}^N$ and preference features $\{v_{tj}\}_{j=1}^N$ at time t;

6 Compute local MLE $\widehat{\theta}_{t-1} \in \arg\max_{\|\theta - \theta^*\|_2 \le \tau} \sum_{t'=1}^{t-1} \log p_{\theta,t}(i_{t'}|S_{t'})$;

7 For every assortment $S \subseteq [N]$, $|S| \le K$, compute its upper confidence bound

$$\overline{R}_t(S) := \mathbb{E}_{\widehat{\theta}_{t-1},t}[r_{tj}|S] + \min\left\{1, \omega \sqrt{\|\widehat{I}_{t-1}^{-1/2}(\widehat{\theta}_{t-1}) \widehat{M}_t(\widehat{\theta}_{t-1}|S) \widehat{I}_{t-1}^{-1/2}(\widehat{\theta}_{t-1})\|_{\mathrm{op}}}\right\};$$

$$\widehat{I}_{t-1}(\theta) := \sum_{t'=1}^{t-1} \widehat{M}_{t'}(\theta|S_{t'}); \quad \widehat{M}_t(\theta|S) := \mathbb{E}_{\theta,t}[v_{tj} v_{tj}^\top|S] - \{\mathbb{E}_{\theta,t}[v_{tj}|S]\}\{\mathbb{E}_{\theta,t}[v_{tj}|S]\}^\top;$$

8 $\omega = \sqrt{d \log(TK)}$;

9 Pick $S_t \in \arg\max_{S \subseteq [N], |S| \le K} \overline{R}_t(S)$ and observe purchasing action $i_t \in S_t \cup \{0\}$;

10 **end**

11 Remark: the expectations admit the following closed-form expressions:

$$\mathbb{E}_{\theta,t}[r_{tj}|S] = \textstyle\sum_{j \in S} p_{\theta,t}(j|S) r_{tj} = \frac{\sum_{j \in S} r_{tj} \exp\{v_{tj}^\top \theta\}}{1 + \sum_{j \in S} \exp\{v_{tj}^\top \theta\}};$$

$$\mathbb{E}_{\theta,t}[v_{tj}|S] = \textstyle\sum_{j \in S} p_{\theta,t}(j|S) v_{tj} = \frac{\sum_{j \in S} v_{tj} \exp\{v_{tj}^\top \theta\}}{1 + \sum_{j \in S} \exp\{v_{tj}^\top \theta\}};$$

$$\mathbb{E}_{\theta,t}[v_{tj} v_{tj}^\top|S] = \textstyle\sum_{j \in S} p_{\theta,t}(j|S) v_{tj} v_{tj}^\top = \frac{\sum_{j \in S} v_{tj} v_{tj}^\top \exp\{v_{tj}^\top \theta\}}{1 + \sum_{j \in S} \exp\{v_{tj}^\top \theta\}}.$$

Algorithm 1: The MLE-UCB policy for dynamic assortment optimization with changing features

Such upper bounds are built using a *local Maximum Likelihood Estimation (MLE)* of θ_0. In particular, in Step 6, instead of computing an MLE, we compute a local MLE, where the estimator $\widehat{\theta}_{t-1}$ lies in a ball centered at the pilot estimator θ^* with a radius τ. This localization also simplifies the technical analysis based on Taylor expansion, which benefits from the constraint that $\widehat{\theta}_{t-1}$ is not too far away from θ^*.

To construct the confidence bound, we introduce the matrices $\widehat{M}_t(\widehat{\theta}_{t-1}|S)$ and $\widehat{I}_{t-1}(\widehat{\theta}_{t-1})$ in Step 7 of Algorithm 1, which are empirical estimates of the *Fisher's information* matrices $-\mathbb{E}[\nabla_\theta^2 \log p(\cdot|\theta)]$ corresponding to the MNL choice model $p(\cdot|S_t)$. The population version of the Fisher's information matrices are presented in Eq. (8) in Sec. 3.2.2. These quantities play an essential role in classical statistical analysis of maximum likelihood estimators (see, e.g., (Van der Vaart, 2000)).

In the rest of this section, we give a regret analysis that shows an $\widetilde{O}(d\sqrt{T})$ upper bound on the regret of the MLE-UCB policy. Additionally, we prove a lower bound of $\widetilde{O}(d\sqrt{T}/K)$ in Sec. 4 and show how the combinatorial optimization in Step 7 can be approximately computed efficiently in Sec. 5.

3.1 Regret analysis

To establish rigorous regret upper bounds on Algorithm 1, we impose the following assumptions:

(A1) $\{v_{tj}\}_{t,j=1}^{T,N}$ are i.i.d. generated from an unknown distribution supported on $\{v \in \mathbb{R}^d : \|v\|_2 \le 1\}$ with the density μ satisfying that $\lambda_{\min}(\mathbb{E}_\mu(v-a)(v-a)^\top) \ge \lambda_0$ for some constant $\lambda_0 > 0$ and all $a \in \mathbb{R}^d$, $\|a\|_2 \le 1$;

(A2) The underlying linear model $\theta_0 \in \mathbb{R}^d$ satisfies $\|\theta_0\|_2 \le 1$.

The item (A1) assumes that the contextual information vectors $\{v_{tj}\}$ are *randomly generated* from a compactly supported and non-degenerate density. (A2) further assumes that the regression model θ_0 is bounded in ℓ_2 norm, a standard assumption in the contextual bandit literature (Chu et al., 2011; Filippi et al., 2010). We further note that the upper bound 1 on $\|v\|_2$ and $\|\theta_0\|_2$ in (A1) and (A2) can be replaced by any constant. We choose 1 here only for the simplicity of presentation.

We also remark that it is possible to relax Assumption (A1) in the following two ways:

1. Assumption (A1) can be relaxed so that it only needs to hold during the pure-exploration phase of our proposed algorithm (more specifically, the first $T_0 = \lceil\sqrt{T}\rceil$ selling periods), instead of being imposed over the entire T periods. This is because, after the pilot estimator θ^* is obtained, the rest of the algorithm and the analysis no longer requires Assumption (A1).

2. Instead of assuming the context vectors $\{v_{t,j}\}$ are i.i.d. from an underlying distribution, it is sufficient to adopt the weaker deterministic condition that, for an "exploration" assortment $S \subseteq [N]$ it holds that $\mathbb{E}_{\theta_0}[(v_{tj} - \mathbb{E}_{\theta_0} v_{tj})(v_{tj} - \mathbb{E}_{\theta_0} v_{tj})^\top] \succeq \lambda_0 I$ for all t, where all expectations are conditioned on the assortment S. Such a "eigenvalue" deterministic condition is implied by Assumption (A1), and is sufficient to prove Lemma 3 on the statistical property of the pilot estimator.

In general, Assumption (A1) can be relaxed to be only imposed on the pure-exploration phase of our algorithm that implies the exploration algorithm can get information about θ_0 along *all* directions in $\mathbb{R}^d$. This is necessary for the pilot estimator after the exploration phase to converge to θ_0 in the ℓ_2 vector norm, because otherwise some directions in $\mathbb{R}^d$ will not receive sufficient information which prevents the pilot estimator to converge to θ_0 in all directions.

We are now ready to state our main result that upper bounds the worst-case accumulated regret of the proposed MLE-UCB policy in Algorithm 1.

Theorem 1 *Suppose that $T \gtrsim \lambda_0^{-8} d^8 \log^8 T + K^8$. Then the regret of the MLE-UCB policy in Algorithm 1 is upper bounded by*

$$\mathrm{Regret}(\{S_t\}_{t=1}^T) = O\left(d\sqrt{T}\log(\lambda_0^{-1}TK)\right) \tag{6}$$

Remark 2 *In the O-notation in Theorem 1 only universal constants are hidden.*

One remarkable aspect of Theorem 1 is the fact that the regret upper bound has *no* dependency on the total number of items N (even in a logarithmic term). This is an attractive property of the proposed policy, which allows N to be very large, even exponentially large in d and K.

Comparing with the regret obtained for plain MNL models (Agrawal et al., 2017, 2019), our regret upper bound in Theorem 1 achieves an $\widetilde{O}(\sqrt{N})$ improvement. To understand, intuitively, why an improvement of $\widetilde{O}(\sqrt{N})$ is possible with contextual information, one can think from an information-theoretical perspective on the difference between the two models: without contextual information each of the N product will have a completely independent utility/preference parameter needed to be estimated, and therefore the regret will unavoidably scale polynomially with N. On the other hand, with contextual information all the utility information can be summarized in an $\mathbb{R}^d$ context vector and the only thing an algorithm needs to learn is an $\mathbb{R}^d$ unknown regression coefficient vector θ_0. Hence, contextual dynamic assortment optimization has the potential of a vast $\widetilde{O}(\sqrt{N})$ improvement over plain MNL-Bandit models because of the additional contextual information available.

3.2 Proof sketch of Theorem 1

We provide a proof sketch of Theorem 1 in this section. The proofs of technical lemmas are relegated to the online supplement.

The proof is divided into four steps. In the first step, we analyze the pilot estimator θ^* obtained from the pure exploration phase of Algorithm 1, and show as a corollary that the true model θ_0 is feasible to all subsequent local MLE formulations with high probability (see Corollary 4). In the second step, we use an ε-net argument to analyze the estimation error of the local MLE. Afterwards, we show in the third step that an upper bound on the estimation error $\widehat{\theta}_{t-1} - \theta_0$ implies an upper bound on the estimation error of the expected revenue $R_t(S)$, hence showing that $\overline{R}_t(S)$ are valid upper confidence bounds. Finally, we apply the *elliptical potential lemma*, which also plays a key role in linear stochastic bandit and its variants, to complete our proof.

3.2.1 Analysis of pure exploration and the pilot estimator

Our first step is to establish an upper bound on the estimation error $\|\theta^* - \theta_0\|_2$ of the pilot estimator θ^*, built using pure exploration data. Because the assortments $\{S_t\}_{t=1}^{T_0}$ in the pure-exploration phase consist of products/items chosen uniformly at random, the analysis on the pilot estimator error $\|\theta^* - \theta_0\|_2$ becomes much easier than the analysis of local MLEs to be conducted in the next section. More specifically, we have the following lemma, which is proved in the appendix:

Lemma 3 *With probability* $1 - O(T^{-1})$ *it holds that*

$$\|\theta^* - \theta_0\|_2 = O((\lambda_0^{-1} d \log T / T_0)^{1/4}). \tag{7}$$

The following corollary immediately follows Lemma 3, by lower bounding $\lambda_{\min}(V)$ using standard matrix concentration inequalities. Its proof is again deferred to the appendix.

Corollary 4 *There exists a universal constant $C_0 > 0$ such that, if $T \geq C_0 \lambda_0^{-8} d^8 \log^8 T$ then with probability $1 - O(T^{-1})$, $\|\theta^* - \theta_0\|_2 \leq T^{-1/8} = \tau$.*

The purpose of Corollary 4 is to establish a connection between the number of pure exploration iterations T_0 and the critical radius τ used in the local MLE formulation. It shows a lower bound on T_0 in order for the estimation error $\|\theta^* - \theta_0\|_2$ to be upper bounded by τ with high probability, which certifies that the true model θ_0 is also a *feasible* local estimator in our MLE-UCB policy. This is an important property for later analysis of local MLE solutions $\widehat{\theta}_{t-1}$.

3.2.2 Analysis of the local MLE

The following lemma upper bounds a Mahalanobis distance between $\widehat{\theta}_t$ and θ_0. For convenience, we adopt the notation that $r_{t0} = 0$ and $v_{t0} = 0$ for all t throughout this section. We also define

$$\begin{aligned} I_t(\theta) &:= \sum_{t'=1}^{t} M_{t'}(\theta), \qquad (8) \\ M_{t'}(\theta) &:= -\mathbb{E}_{\theta_0,t'}[\nabla_\theta^2 \log p_{\theta,t'}(j|S_{t'})] \\ &= \mathbb{E}_{\theta_0,t'}[v_{t'j} v_{t'j}^\top] - \{\mathbb{E}_{\theta_0,t'} v_{t'j}\}\{\mathbb{E}_{\theta,t'} v_{t'j}\}^\top - \{\mathbb{E}_{\theta,t'} v_{t'j}\}\{\mathbb{E}_{\theta_0,t'} v_{t'j}\}^\top + \{\mathbb{E}_{\theta,t'} v_{t'j}\}\{\mathbb{E}_{\theta,t'} v_{t'j}\}^\top \end{aligned}$$

where $\mathbb{E}_{\theta,t'}$ denotes the expectation evaluated under the law $j \sim p_{\theta,t'}(\cdot|S_{t'})$; that is, $p_{\theta,t'}(j|S_{t'}) = \exp\{v_{t'j}^\top \theta\}/(1 + \sum_{k \in S_{t'}} \exp\{v_{t'j}^\top \theta\})$ for $j \in S_{t'}$ and $p_{\theta,t'}(j|S_{t'}) = 0$ for $j \notin S_{t'}$.

Lemma 5 *Suppose $\tau \leq 1/(15K)$. Then there exists a universal constant $C > 0$ such that with probability $1 - O(T^{-1})$ the following holds uniformly over all $t = T_0, \cdots, T-1$:*

$$(\widehat{\theta}_t - \theta_0)^\top I_t(\theta_0)(\widehat{\theta}_t - \theta_0) \leq C \cdot d \log(TK). \qquad (9)$$

Remark 6 *For $\theta = \theta_0$, the expression of $M_{t'}(\theta)$ can be simplified as $M_{t'}(\theta_0) = \mathbb{E}_{\theta_0,t'}[v_{t'j} v_{t'j}^\top] - \{\mathbb{E}_{\theta_0,t'} v_{t'j}\}\{\mathbb{E}_{\theta_0,t'} v_{t'j}\}^\top$.*

The complete proof of Lemma 5 is given in the appendix, and here we provide some high-level ideas behind our proof.

Our proof is inspired by the classical convergence rate analysis of M-estimators (Van der Vaart, 2000, Sec. 5.8). The main technical challenge is to provide *finite-sample* analysis of several components in the proof of (Van der Vaart, 2000, Sec. 5.8).

In particular, for any $\theta \in \mathbb{R}^d$, consider

$$F_t(\theta) := \sum_{t' \leq t} f_{t'}(\theta) \quad \text{where} \quad f_{t'}(\theta) := \mathbb{E}_{\theta_0,t'}\left[\log \frac{p_{\theta,t'}(j|S_{t'})}{p_{\theta_0,t'}(j|S_{t'})}\right] = \sum_{j \in S_{t'} \cup \{0\}} p_{\theta_0,t'}(j|S_{t'}) \log \frac{p_{\theta,t'}(j|S_{t'})}{p_{\theta_0,t'}(j|S_{t'})}$$

and its "sample" version

$$\widehat{F}_t(\theta) := \sum_{t' \le t} \widehat{f}_{t'}(\theta) \quad \text{where } \widehat{f}_{t'}(\theta) := \log \frac{p_{\theta,t'}(i_{t'}|S_{t'})}{p_{\theta_0,t'}(i_{t'}|S_{t'})}.$$

It is easy to verify by definition that $F_t(\widehat{\theta}_t) \ge F_t(\theta_0) = 0$ and $\widehat{F}_t(\widehat{\theta}_t) \le \widehat{F}_t(\theta_0) = 0$, because $F_t(\cdot)$ is a Kullback-Leibler divergence, θ_0 is feasible to the local MLE formulation and $\widehat{\theta}_{t-1}$ is the optimal solution. On the other hand, it can be proved that $|F_t(\theta) - \widehat{F}_t(\theta)|$ is small for all θ with high probability, by using concentration inequalities for self-normalized empirical process (note that $\mathbb{E}\widehat{f}_{t'}(\theta) = f_{t'}(\theta)$ for any θ). Moreover, by constructing a local quadratic approximation of $F_t(\cdot)$ around θ_0, we can show that $F_t(\theta) - F_t(\theta_0)$ is large when θ is far away from θ_0.

Following the above observations, we can use proof by contradiction to prove Lemma 5, which essentially claims that $\widehat{\theta}_t$ and θ_0 are close under the quadratic distance $\|\cdot\|_{I_t(\theta_0)}$. Suppose by contradiction that $\widehat{\theta}_t$ and θ_0 are far apart, which implies that $|F_t(\widehat{\theta}_t) - F_t(\theta_0)|$ is large. On the other hand, by the fact that $\widehat{F}_t(\widehat{\theta}_t) \le 0 = F_t(\theta_0) \le F_t(\widehat{\theta}_t)$, we have

$$|F_t(\widehat{\theta}_t) - F_t(\theta_0)| = |F_t(\widehat{\theta}_t)| \le |F_t(\widehat{\theta}_t) - \widehat{F}_t(\widehat{\theta}_t)|.$$

By the established concentration result, we have $|F_t(\theta) - \widehat{F}_t(\theta)|$ is small for all θ with high probability (including $\theta = \widehat{\theta}_t$). This leads to the desired contradiction.

3.2.3 Analysis of upper confidence bounds

The following technical lemma shows that the upper confidence bounds constructed in Algorithm 1 are valid with high probability. Additionally, we establish an upper bound on the discrepancy between $\overline{R}_t(S)$ and the true value $R_t(S)$ defined in Eq. (4).

Lemma 7 *Suppose τ satisfies the condition in Lemma 5. With probability $1 - O(T^{-1})$ the following holds uniformly for all $t > T_0$ and $S \subseteq [N]$, $|S| \le K$ such that*

1. $\overline{R}_t(S) \ge R_t(S)$;
2. $|\overline{R}_t(S) - R_t(S)| \lesssim \min\{1, \omega\sqrt{\|I_{t-1}^{-1/2}(\theta_0) M_t(\theta_0|S) I_{t-1}^{-1/2}(\theta_0)\|_{\mathrm{op}}}\}$.

At a higher level, the proof of Lemma 7 can be regarded as a "finite-sample" version of the classical *Delta's method*, which upper bounds estimation error of some functional φ of parameters, i.e., $|\varphi(\widehat{\theta}_{t-1}) - \varphi(\theta_0)|$ using the estimation error of the parameters themselves $\widehat{\theta}_{t-1} - \theta_0$. The complete proof is relegated to the appendix.

3.2.4 The elliptical potential lemma

Let S_t^* be the assortment that maximizes the expected revenue $R_t(\cdot)$ (defined in Eq. (4)) at time period t, and S_t be the assortment selected by Algorithm 1. Because $R_t(S) \le \overline{R}_t(S)$ for all S (see Lemma 7), we have the following upper bound for each term in the regret (see Eq. (5)):

$$R_t(S_t^*) - R_t(S_t) \le (\overline{R}_t(S_t^*) - \overline{R}_t(S_t)) + (\overline{R}_t(S_t) - R_t(S_t)) \le \overline{R}_t(S_t) - R_t(S_t), \tag{10}$$

where the last inequality holds because $\overline{R}_t(S_t^*) - \overline{R}_t(S_t) \le 0$ (note that S_t maximizes $\overline{R}_t(\cdot)$).

Subsequently, invoking Lemma 7 and the Cauchy-Schwarz inequality, we have

$$\sum_{t=T_0+1}^{T} R_t(S_t^*) - R_t(S_t) \lesssim \sqrt{d\log(TK)} \cdot \sum_{t=T_0+1}^{T} \sqrt{\min\{1, \|I_{t-1}^{-1/2}(\theta_0) M_t(\theta_0|S_t) I_{t-1}^{-1/2}(\theta_0)\|_{\mathrm{op}}\}}$$
$$\lesssim \sqrt{dT\log(TK) \cdot \sum_{t=T_0+1}^{T} \min\{1, \|I_{t-1}^{-1/2}(\theta_0) M_t(\theta_0|S_t) I_{t-1}^{-1/2}(\theta_0)\|_{\mathrm{op}}^2\}}. \tag{11}$$

The following lemma is a key result that upper bounds $\sum_{t=T_0+1}^{T} \min\{1, \|I_{t-1}^{-1/2}(\theta_0) M_t(\theta_0|S_t) I_{t-1}^{-1/2}(\theta_0)\|_{\mathrm{op}}^2\}$. It is usually referred to as the *elliptical potential lemma* and has found many applications in contextual bandit-type problems (see, e.g., Dani et al. (2008); Rusmevichientong et al. (2010); Filippi et al. (2010); Li et al. (2017)).

Lemma 8 *It holds that*

$$\sum_{t=T_0+1}^{T} \min\{1, \|I_{t-1}^{-1/2}(\theta_0) M_t(\theta_0|S_t) I_{t-1}^{-1/2}(\theta_0)\|_{\mathrm{op}}^2\} \le 4\log\frac{\det I_T(\theta_0)}{\det I_{T_0}(\theta_0)} \lesssim d\log(\lambda_0^{-1}).$$

The proof of Lemma 8 is placed in the appendix. It is a routine proof following existing proofs of elliptical potential lemmas using matrix-determinant rank-1 updates.

We are now ready to give the final upper bound on $\mathrm{Regret}(\{S_t\}_{t=1}^T)$ defined in Eq. (5). Note that the total regret incurred by the pure exploration phase is upper bounded by T_0, because the revenue parameters r_{tj} are normalized so that they are upper bounded by 1. In addition, as the failure event of $\overline{R}_t(S) \le R_t(S)$ for some S occurs with probability $1 - O(T^{-1})$, the total regret accumulated under the failure event is $O(T^{-1}) \cdot T = O(1)$. Further invoking Eq. (11) and Lemma 8, we have

$$\begin{aligned}\mathrm{Regret}(\{S_t\}_{t=1}^T) &\le T_0 + O(1) + \mathbb{E}\sum_{t=T_0+1}^{T} R_t(S_t^*) - R_t(S_t) \\ &\lesssim \sqrt{T} + d\sqrt{T} \cdot \log(\lambda_0^{-1} TK) \\ &\lesssim d\sqrt{T} \cdot \log(\lambda_0^{-1} TK). \end{aligned} \tag{12}$$

4. Lower bound

To complement our regret analysis in Sec. 3.1, in this section we prove a *lower* bound for worst-case regret. Our lower bound is information theoretical, and therefore applies to *any* policy for dynamic assortment optimization with changing contextual features.

Theorem 9 *Suppose d is divisible by 4. There exists a universal constant $C_0 > 0$ such that for any sufficiently large T and policy π, there is a worst-case problem instance with $N \asymp K \cdot 2^d$ items and uniformly bounded feature and coefficient vector (i.e., $\|v_{ti}\|_2 \le 1$ and $\|\theta_0\|_2 \le 1$ for all $i \in [N]$, $t \in [T]$) such that the regret of π is lower bounded by $C_2 \cdot d\sqrt{T}/K$.*

Theorem 9 essentially implies that the $\widetilde{O}(d\sqrt{T})$ regret upper bound established in Theorem 1 is tight (up to logarithmic factors) in T and d. Although there is an $O(K)$ gap between the upper and lower regret bounds, in practical applications K is usually small and can be generally regarded as a constant. It is an interesting technical open problem to close this gap of $O(K)$.

We also remark that an $\Omega(d\sqrt{T})$ lower bound was established in (Dani et al., 2008) for contextual linear bandit problems. However, in assortment selection, the reward function is *not* coordinate-wise decomposable, making techniques in Dani et al. (2008) not directly applicable. In the following subsection, we provide a high-level proof sketch of Theorem 9, with complete proofs of technical lemmas relegated to the appendix.

4.1 Proof sketch of Theorem 9

At a higher level, the proof of Theorem 9 can be divided into three steps (separated into three different sub-sections below). In the first step, we construct an *adversarial parameter* set and reduce the task of lower bounding the *worst-case* regret of any policy to lower bounding the *Bayes risk* of the constructed parameter set. In the second step, we use a "counting argument" similar to the one developed in Chen and Wang (2018) to provide an explicit lower bound on the Bayes risk of the constructed adversarial parameter set, and finally we apply *Pinsker's inequality* (see, e.g., Tsybakov (2009)) to derive a complete lower bound.

4.1.1 Adversarial construction and the Bayes risk

Let $\epsilon \in (0, 1/d\sqrt{d})$ be a small positive parameter to be specified later. For every subset $W \subseteq [d]$, define the corresponding parameter $\theta_W \in \mathbb{R}^d$ as $[\theta_W]_i = \epsilon$ for all $i \in W$, and $[\theta_W]_i = 0$ for all $i \notin W$. The parameter set we consider is

$$\theta \in \Theta := \{\theta_W : W \in \mathcal{W}_{d/4}\} := \{\theta_W : W \subseteq [d], |W| = d/4\}. \tag{13}$$

Note that $d/4$ is a positive integer because d is divisible by 4, as assumed in Theorem 9. Also, to simplify notation, we use $\mathcal{W}_k$ to denote the class of all subsets of $[d]$ whose size is k.

The feature vectors $\{v_{ti}\}$ are constructed to be invariant across time iterations t. For each t and $U \in \mathcal{W}_{d/4}$, K identical feature vectors v_U are constructed as (recall that K is the maximum allowed assortment capacity)

$$[v_U]_i = 1/\sqrt{d} \quad \text{for } \ i \in U; \qquad [v_U]_i = 0 \quad \text{for } \ i \notin U. \tag{14}$$

It is easy to check that with the condition $\epsilon \in (0, 1/\sqrt{d})$, $\|\theta_W\|_2 \leq 1$ and $\|v_U\|_2 \leq 1$ for all $W, U \in \mathcal{W}_{d/4}$. Hence the worst-case regret of any policy π can be lower bounded by the worst-case regret of parameters belonging to Θ, which can be further lower bounded by the

"average" regret over a uniform prior over Θ:

$$\sup_{v,\theta} \mathbb{E}^{\pi}_{v,\theta} \sum_{t=1}^{T} R(S^*_\theta) - R(S_t) \geq \max_{\theta_W \in \Theta} \mathbb{E}^{\pi}_{v,\theta_W} \sum_{t=1}^{T} R(S^*_{\theta_W}) - R(S_t)$$
$$\geq \frac{1}{|\mathcal{W}_{d/4}|} \sum_{W \in \mathcal{W}_{d/4}} \mathbb{E}^{\pi}_{v,\theta_W} \sum_{t=1}^{T} R(S^*_{\theta_W}) - R(S_t). \tag{15}$$

Here S^*_θ is the optimal assortment of size at most K that maximizes (expected) revenue under parameterization θ. By construction, it is easy to verify that $S^*_{\theta_W}$ consists of all K items corresponding to feature v_W. We also employ constant revenue parameters $r_{ti} \equiv 1$ for all $t \in [T]$, $i \in [N]$.

4.1.2 The counting argument

In this section we drive an explicit lower bound on the Bayes risk in Eq. (15). For any sequences $\{S_t\}_{t=1}^T$ produced by the policy π, we first describe an alternative sequence $\{\widetilde{S}_t\}_{t=1}^T$ that provably enjoys less regret under parameterization θ_W, while simplifying our analysis.

Let $v_{U_1}, \cdots, v_{U_L}$ be the distinct feature vectors contained in assortment S_t (if $S_t = \emptyset$ then one may choose an arbitrary feature v_U) with $U_1, \cdots, U_L \in \mathcal{W}_{d/4}$. Let U^* be the subset among $U_1, \cdots, U_L$ that maximizes $\langle v_{U^*}, \theta_W \rangle$, where θ_W is the underlying parameter. Let $\widetilde{S}_t$ be the assortment consisting of all K items corresponding to feature v^*_U. We then have the following observation:

Proposition 10 *$R(S_t) \leq R(\widetilde{S}_t)$ under θ_W.*

Proof. Because $r_{tj} \equiv 1$ in our construction, we have $R(S_t) = (\sum_{j \in S_t} u_j)/(1 + \sum_{j \in S_t} u_j)$ where $u_j = \exp\{v_j^\top \theta_W\}$ under θ_W. Clearly $R(S)$ is a monotonically non-decreasing function in u_j. By replacing all $v_j \in S_t$ with $v_{U^*} \in \widetilde{S}_t$, the u_j values do not decrease and therefore the Proposition holds true. □

To simplify notation we also use $\widetilde{U}_t$ to denote the unique $U^* \in \mathcal{W}_{d/4}$ in $\widetilde{S}_t$. We also use $\mathbb{E}_W$ and $\mathbb{P}_W$ to denote the law parameterized by θ_W and policy π. The following lemma gives a lower bound on $R(S^*_{\theta_W}) - R(\widetilde{S}_t)$ by comparing it with W, which is also proved in the appendix.

Lemma 11 *Suppose $\epsilon \in (0, 1/d\sqrt{d})$ and define $\delta := d/4 - |\widetilde{U}_t \cap W|$. Then*

$$R(S^*_{\theta_W}) - R(\widetilde{S}_t) \geq \frac{\delta\epsilon}{4K\sqrt{d}}.$$

Define random variables $\widetilde{N}_i := \sum_{t=1}^T \mathbf{1}\{i \in \widetilde{U}_t\}$. Lemma 11 immediately implies

$$\mathbb{E}_W \sum_{t=1}^{T} R(S^*_{\theta_W}) - R(\widetilde{S}_t) \geq \frac{\epsilon}{4K\sqrt{d}} \left(\frac{dT}{4} - \sum_{i \in W} \mathbb{E}_W[\widetilde{N}_i] \right), \quad \forall W \in \mathcal{W}_{d/4}. \tag{16}$$

Denote $\mathcal{W}_{d/4}^{(i)} := \{W \in \mathcal{W}_{d/4} : i \in W\}$ and $\mathcal{W}_{d/4-1} := \{W \subseteq [d] : |W| = d/4 - 1\}$. Averaging both sides of Eq. (16) with respect to all $W \in \mathcal{W}_{d/4}$ and swapping the summation order, we have

$$\begin{aligned}
\frac{1}{|\mathcal{W}_{d/4}|}\sum_{W\in\mathcal{W}_{d/4}} \mathbb{E}_W \sum_{t=1}^{T} R(S^*_{\theta_W}) - R(S_t) &\geq \frac{\epsilon}{4K\sqrt{d}} \frac{1}{|\mathcal{W}_{d/4}|}\sum_{W\in\mathcal{W}_{d/4}} \left(\frac{dT}{4} - \sum_{i\in W} \mathbb{E}_W[\widetilde{N}_i]\right) \\
&= \frac{\epsilon}{4K\sqrt{d}} \left(\frac{dT}{4} - \frac{1}{|\mathcal{W}_{d/4}|}\sum_{i=1}^{d}\sum_{W\in\mathcal{W}_{d/4}^{(i)}} \mathbb{E}_W[\widetilde{N}_i]\right) \\
&= \frac{\epsilon}{4K\sqrt{d}} \left(\frac{dT}{4} - \frac{1}{|\mathcal{W}_{d/4}|}\sum_{W\in\mathcal{W}_{d/4-1}}\sum_{i\notin W} \mathbb{E}_{W\cup\{i\}}[\widetilde{N}_i]\right) \\
&\geq \frac{\epsilon}{4K\sqrt{d}} \left(\frac{dT}{4} - \frac{|\mathcal{W}_{d/4-1}|}{|\mathcal{W}_{d/4}|}\max_{W\in\mathcal{W}_{d/4-1}}\sum_{i\notin W} \mathbb{E}_{W\cup\{i\}}[\widetilde{N}_i]\right) \\
&= \frac{\epsilon}{4K\sqrt{d}} \left(\frac{dT}{4} - \frac{|\mathcal{W}_{d/4-1}|}{|\mathcal{W}_{d/4}|}\max_{W\in\mathcal{W}_{d/4-1}}\sum_{i\notin W} \mathbb{E}_W[\widetilde{N}_i] + \mathbb{E}_{W\cup\{i\}}[\widetilde{N}_i] - \mathbb{E}_W[\widetilde{N}_i]\right).
\end{aligned}$$

Note that for any fixed W, $\sum_{i\notin W} \mathbb{E}_W[\widetilde{N}_i] \leq \sum_{i=1}^{d} \mathbb{E}_W[\widetilde{N}_i] \leq dT/4$. Also, $|\mathcal{W}_{d/4-1}|/|\mathcal{W}_{d/4}| = \binom{d}{d/4-1}/\binom{d}{d/4} = \frac{d/4}{3d/4+1} \leq 1/3$. Subsequently,

$$\frac{1}{|\mathcal{W}_{d/4}|}\sum_{W\in\mathcal{W}_{d/4}} \mathbb{E}_W \sum_{t=1}^{T} R(S^*_{\theta_W}) - R(S_t) \geq \frac{\epsilon}{4K\sqrt{d}} \left(\frac{dT}{6} - \max_{W\in\mathcal{W}_{d/4-1}}\sum_{i\notin W} |\mathbb{E}_{W\cup\{i\}}[\widetilde{N}_i] - \mathbb{E}_W[\widetilde{N}_i]|\right). \tag{17}$$

4.1.3 Pinsker's inequality

In this section we concentrate on upper bounding $|\mathbb{E}_{W\cup\{i\}}[\widetilde{N}_i] - \mathbb{E}_W[\widetilde{N}_i]|$ for any $W \in \mathcal{W}_{d/4-1}$. Let $P = \mathbb{P}_W$ and $Q = \mathbb{P}_{W\cup\{i\}}$ denote the laws under θ_W and $\theta_{W\cup\{i\}}$, respectively. Then

$$\begin{aligned}
\left|\mathbb{E}_P[\widetilde{N}_i] - \mathbb{E}_Q[\widetilde{N}_i]\right| &\leq \sum_{j=0}^{T} j \cdot \left|P[\widetilde{N}_i = j] - Q[\widetilde{N}_i = j]\right| \\
&\leq T \cdot \sum_{j=0}^{T} \left|P[\widetilde{N}_i = j] - Q[\widetilde{N}_i = j]\right| \\
&\leq T \cdot \|P - Q\|_{\mathrm{TV}} \leq T \cdot \sqrt{\frac{1}{2}\mathrm{KL}(P\|Q)},
\end{aligned}$$

where $\|P - Q\|_{\mathrm{TV}} = \sup_A |P(A) - Q(A)|$ is the total variation distance between P, Q, $\mathrm{KL}(P\|Q) = \int(\log \mathrm{d}P/\mathrm{d}Q)\mathrm{d}P$ is the Kullback-Leibler (KL) divergence between P, Q, and the inequality $\|P - Q\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2}\mathrm{KL}(P\|Q)}$ is the celebrated Pinsker's inequality.

For every $i \in [d]$ define random variables $N_i := \sum_{t=1}^{T} \frac{1}{K} \sum_{v_U \in S_t} \mathbf{1}\{i \in U\}$. The next lemma upper bound the KL divergence, which is proved in the appendix.

Lemma 12 *For any $W \in \mathcal{W}_{d/4-1}$ and $i \in [d]$, $\mathrm{KL}(P_W \| P_{W \cup \{i\}}) \le C_{\mathrm{KL}} \cdot \mathbb{E}_W[N_i] \cdot \epsilon^2/d$ for some universal constant $C_{\mathrm{KL}} > 0$.*

Combining Lemma 12 and Eq. (17), we have

$$\frac{1}{|\mathcal{W}_{d/4}|} \sum_{W \in \mathcal{W}_{d/4}} \mathbb{E}_W \sum_{t=1}^{T} R(S^*_{\theta_W}) - R(S_t) \ge \frac{\epsilon}{4K\sqrt{d}} \left(\frac{dT}{6} - T \sum_{i=1}^{d} \sqrt{C_{\mathrm{KL}} \mathbb{E}_W[N_i] \epsilon^2/d} \right).$$

Further using Cauchy-Schwartz inequality, we have

$$\sum_{i=1}^{d} \sqrt{C_{\mathrm{KL}} \mathbb{E}_W[N_i] \epsilon^2/d} \le \sqrt{d} \cdot \sqrt{\sum_{i=1}^{d} C_{\mathrm{KL}} \mathbb{E}_W[N_i] \epsilon^2/d},$$

which is further upper bounded by $\sqrt{d} \cdot \sqrt{C_{\mathrm{KL}} T \epsilon^2/4}$ because $\sum_{i=1}^{d} \mathbb{E}_W[N_i] \le dT/4$. Subsequently,

$$\frac{1}{|\mathcal{W}_{d/4}|} \sum_{W \in \mathcal{W}_{d/4}} \mathbb{E}_W \sum_{t=1}^{T} R(S^*_{\theta_W}) - R(S_t) \ge \frac{\epsilon}{4K\sqrt{d}} \left(\frac{dT}{6} - T \sqrt{C'_{\mathrm{KL}} dT \epsilon^2} \right), \tag{18}$$

where $C'_{\mathrm{KL}} = C_{\mathrm{KL}}/4$. Setting $\epsilon = \sqrt{d/144 C'_{\mathrm{KL}} T}$ we complete the proof of Theorem 9.

5. The combinatorial optimization subproblem

The major computational bottleneck of our algorithm is its Step 8, which involves solving a *combinatorial* optimization problem. For notational simplicity, we equivalently reformulate this problem as follows:

$$\max_{S \subseteq [N], |S| \le K} \mathrm{ESTR}(S) + \min\{1, \omega \cdot \mathrm{CI}(S)\} \quad \text{where } \mathrm{ESTR}(S) := \frac{\sum_{j \in S} r_{tj} \widehat{u}_{tj}}{1 + \sum_{j \in S} \widehat{u}_{tj}} \text{ and } \tag{19}$$

$$\mathrm{CI}(S) := \sqrt{\left\| \frac{\sum_{j \in S} \widehat{u}_{tj} x_{tj} x_{tj}^\top}{1 + \sum_{j \in S} \widehat{u}_{tj}} - \left(\frac{\sum_{j \in S} \widehat{u}_{tj} x_{tj}}{1 + \sum_{j \in S} \widehat{u}_{tj}} \right) \left(\frac{\sum_{j \in S} \widehat{u}_{tj} x_{tj}}{1 + \sum_{j \in S} \widehat{u}_{tj}} \right)^\top \right\|_{\mathrm{op}}}.$$

Here $\widehat{u}_{tj} := \exp\{v_{tj}^\top \widehat{\theta}_{t-1}\}$ and $x_{tj} := \widehat{I}_{t-1}^{-1/2}(\widehat{\theta}_{t-1}) v_{tj}$, both of which can be pre-computed before solving Eq. (19).

A brute-force way to compute Eq. (19) is to enumerate all subsets $S \subseteq [N]$, $|S| \le K$ and select the one with the largest objective value. Such an approach is *not* an efficient (polynomial-time) algorithm and is therefore not scalable.

In this section we provide two alternative methods for (approximately) solving the combinatorial optimization problem in Eq. (19). Our first algorithm is based on discretized dynamic programming and enjoys rigorous approximation guarantees. The second algorithm is a computationally efficient greedy heuristic. Although the greedy heuristic does not have rigorous guarantees, our numerical result suggests it works reasonably well (see Sec. 6).

5.1 Approximation algorithms for assortment optimization

In this section we introduce algorithms with polynomial running times and rigorous approximation guarantees for the optimization task described in Eq. (19). We first formally introduce the concept of $(\alpha, \varepsilon, \delta)$-*approximation* to characterize the approximation performance, and show that such approximation guarantees imply certain upper bounds on the final regret.

Definition 13 ($(\alpha, \varepsilon, \delta)$-approximation) *Fix $\alpha \geq 1$, $\varepsilon \geq 0$ and $\delta \in [0, 1)$. An algorithm is an* $(\alpha, \varepsilon, \delta)$-approximation algorithm *if it produces $\widehat{S} \subseteq [N]$, $|\widehat{S}| \leq K$ such that with probability at least $1 - \delta$,*

$$\mathrm{ESTR}(\widehat{S}) + \min\{1, \alpha\omega \cdot \mathrm{CI}(\widehat{S})\} + \varepsilon \geq \mathrm{ESTR}(S^*) + \min\{1, \omega \cdot \mathrm{CI}(S^*)\}, \tag{20}$$

where S^* is the assortment set maximizing the actual objective in Eq. (19) [2].

The following lemma shows how $(\alpha, \varepsilon, \delta)$-approximation algorithms imply an upper bound on the accumulated. It is proved using standard analysis of UCB type algorithms, with the complete proof given in Sec. C in the appendix.

Lemma 14 *Suppose an $(\alpha, \varepsilon, \delta)$-approximation algorithm is used instead of exact optimization in the MLE-UCB policy at each time period t. Then its regret can be upper bounded by*

$$\alpha \cdot \mathrm{Regret}^* + \varepsilon T + \delta T^2 + O(1),$$

where Regret^* *is the regret upper bound shown by Theorem 1 for Algorithm 1 with exact optimization in Step 8.*

In the rest of this section we introduce our proposed approximation algorithm and the approximation guarantee. To highlight the main idea of the approximation algorithm, we only describe how the algorithm operates in the *univariate* ($d = 1$) case, while leaving the general multivariate ($d > 1$) case to Sec. D in the appendix.

Our approximation algorithm can be roughly divided into three steps. In the first step, we use a "discretization" trick to approximate the objective function using "rounded" parameter values. Such rounding motivates the second step, in which we define "reachable states" and present a simple yet computationally expensive brute-force method to enumerate all reachable states, and establish approximation guarantees for such methods. This brute-force method is only presented for illustration purposes and will be replaced by a dynamic programing algorithm proposed in the third step. In particular, a *dynamic programming* algorithm is developed to compute which states are "reachable" in polynomial time.

5.1.1 The discretization trick

In the univariate case, $\{x_{tj}\}$ are scalars and therefore $x_{tj}x_{tj}^\top$ is simply x_{tj}^2. Let $\Delta > 0$ be a small positive discretization parameter to be specified later. For all $i \in [N]$, define

$$\mu_i := \left\lceil \frac{\widehat{u}_{ti}}{\Delta} \right\rceil \Delta, \;\; \alpha_i := \left\lceil \frac{\widehat{u}_{ti} x_{ti}}{\Delta} \right\rceil \Delta, \;\; \beta_i := \left\lceil \frac{\widehat{u}_{ti} x_{ti}^2}{\Delta} \right\rceil \Delta, \;\; \gamma_i := \left\lceil \frac{\widehat{u}_{ti} r_{ti}}{\Delta} \right\rceil \Delta, \tag{21}$$

2. We slightly abuse the notation S^* here following the optimization convention that S^* denotes the optimal solution. Note that S^* is different from S_t^* in (5), where the latter means the assortment that maximizes the expected revenue at time t.

where $[a]$ denotes the nearest integer a real number a is rounded into. Intuitively, μ_i is the real number closest to $\widehat{u}_{ti}$ that is an *integer* multiple of the discretization parameter Δ, and similarly for $\alpha_i, \beta_i, \gamma_i$.

The motivation for the definitions of $\{\mu_i, \alpha_i, \beta_i, \gamma_i\}$ is their *sufficiency* in computing the objective function $\mathrm{ESTR}(S) + \min\{1, \omega \cdot \mathrm{CI}(S)\}$. Indeed, for any $S \subseteq [n]$, $|S| \le K$, define $\mu = \sum_{j\in S} \mu_j$, $\alpha = \sum_{j\in S} \alpha_j$, $\beta = \sum_{j\in S} \beta_j$, $\gamma = \sum_{j\in S} \gamma_j$ and

$$\widehat{\mathrm{ESTR}}(S) := \frac{\gamma}{1+\mu}, \qquad \widehat{\mathrm{CI}}(S) := \max\left\{0, \sqrt{\frac{\beta}{1+\mu} - \left(\frac{\alpha}{1+\mu}\right)^2}\right\}.$$

Following the definition of $\mathrm{ESTR}(S)$ and $\mathrm{CI}(S)$, it is easy to see that $\widehat{\mathrm{ESTR}}(S) \to \mathrm{ESTR}(S)$ and $\widehat{\mathrm{CI}}(S) \to \mathrm{CI}(S)$ as $\Delta \to 0^+$. The following lemma gives a more precise control of the error between $\widehat{\mathrm{ESTR}}(S)$, $\widehat{\mathrm{CI}}(S)$ and $\mathrm{ESTR}(S)$, $\mathrm{CI}(S)$ using the values of Δ and the maximum utility parameter in S.

Lemma 15 *For any $S \subseteq [N]$, $|S| \le K$, suppose $U = \max_{j\in S}\{1, \widehat{u}_{tj}\}$ and $\Delta = \epsilon_0 U/K$ for some $\epsilon_0 > 0$. Suppose also $|x_{tj}| \le 1$ for all t, j. Then*

$$\left|\mathrm{ESTR}(S) - \widehat{\mathrm{ESTR}}(S)\right| \le 6\epsilon_0 \quad \text{and} \quad \left|\mathrm{CI}(S) - \widehat{\mathrm{CI}}(S)\right| \le \sqrt{96\epsilon_0}, \tag{22}$$

The complete proof of Lemma 15 is relegated to Sec. C in the appendix.

5.1.2 Reachable states and a brute-force algorithm

To apply the estimation error bounds in Lemma 15 one needs to first enumerate $q \in [N]$ giving rise to the item in S with the largest utility parameter $\widehat{u}_{tq}$. After such an element q is enumerated, the discretization parameter $\Delta = \epsilon_0 U/K = \epsilon_0 \max\{1, \widehat{u}_{tq}\}/K$ can be determined and discretized values $\mu_i, \alpha_i, \beta_i, \gamma_i$ can be computed for all $i \in [N]/\backslash\{q\}$. It is also easy to verify that there are at most $O(K/\epsilon)$ possible values of μ_i, γ_i, $O(K/\epsilon)$ possible values of α_i and $O(K/\epsilon)$ possible values of β_i.

For any $i \in [N] \cup \{0\}$, $k \in [K] \cup \{0\}$ and $\mu, \alpha, \beta, \gamma \ge 0$ being *integer* multiples of Δ, we use a tuple $\varsigma_i^k(\mu, \alpha, \beta, \gamma)$ to denote a *state*. Here the indices i and k mean that the assortment $S \subseteq \{1, 2, \cdots, i\}$ and $|S| = k$. Clearly there are at most $O(NK^5/\epsilon^4)$ different types of states. A state $\varsigma_i^k(\mu, \alpha, \beta, \gamma)$ can be either *reachable* or *non-reachable*, as defined below:

Definition 16 *Let $q \in [N]$ be the enumerated item with maximal utility parameter and $U = \max\{1, \widehat{u}_{tq}\}$, $\Delta = \epsilon_0 U/K$. A state $\varsigma_i^k(\mu, \alpha, \beta, \gamma)$ is* reachable *if there exists $S \subseteq [N]$ satisfying the following:*

1. *$S \subseteq \{1, 2, \cdots, i\}$ and $|S| = k$;*
2. *$\widehat{u}_{tj} \le \widehat{u}_{tq}$ for all $j \in S$;*
3. *if $i \ge q$ then $q \in S$;*

4. $\mu = \sum_{j\in S}\mu_j$, $\alpha = \sum_{j\in S}\alpha_j$, $\beta = \sum_{j\in S}\beta_j$ *and* $\gamma = \sum_{j\in S}\gamma_j$.

On the other hand, a state $\varsigma_i^k(\mu,\alpha,\beta,\gamma)$ *is* non-reachable *if at least one condition above is violated.*

A simple way to find all reachable states is to enumerate all $S \subseteq [N]$, $|S| \le K$ and verify the three conditions in Definition 16. While such a procedure is clearly computationally intractable, in the next section we will present a dynamic programming approach to compute all reachable states in polynomial time. After all reachable states are computed, enumerate over every $q \in [N]$ and reachable $\zeta_N^k(\cdot,\cdot,\cdot,\cdot)$ for $k \in [K]$ and find $\widehat{S}$ that maximizes $\widehat{\mathrm{ESTR}}(\widehat{S}) + \min\{1, \omega \cdot \widehat{\mathrm{CI}}(\widehat{S})\}$. The following corollary establishes the approximation guarantee for $\widehat{S}$, following Lemma 15.

Corollary 17 *Let* $\widehat{S} \subseteq [N]$, $|\widehat{S}| \le K$ *be a subset corresponding to a reachable state* $\varsigma_N^k(\cdot,\cdot,\cdot,\cdot)$ *for some* $k \in [K]$, $q \in [N]$, *that maximizes* $\widehat{\mathrm{ESTR}}(\widehat{S}) + \min\{1, \omega \cdot \widehat{\mathrm{CI}}(\widehat{S})\}$. *Then*

$$\mathrm{ESTR}(\widehat{S}) + \min\{1, \omega \cdot \mathrm{CI}(\widehat{S})\} \ge \max_{S\subseteq[N],|S|\le K} \mathrm{ESTR}(S) + \min\{1, \omega \cdot \mathrm{CI}(S)\} - (6\epsilon_0 + \omega\sqrt{96\epsilon_0}).$$

Corollary 17 follows easily by plugging in the upper bounds of estimation error in Lemma 15. By setting $\epsilon_0 = \min\{\varepsilon/12, \varepsilon^2/(384\omega^2)\}$, the algorithm that produces $\widehat{S}$ satisfies $(1,\varepsilon,0)$-approximation as defined in Definition 13.

5.1.3 A dynamic programming method for computation of reachable states

In this section we describe a *dynamic programming* algorithm to compute reachable states in polynomial time. The dynamic programming algorithm is exact and deterministic, therefore approximation guarantees in Corollary 17 remain valid.

The first step is again to enumerate $q \in [N]$ corresponding to the item in S with the largest utility parameter $\widehat{u}_{tq}$, and calculating the discretization parameter $\Delta = \epsilon \max\{1, \widehat{u}_{tq}\}/K$. Afterwards, reachable states are computed in an iterative manner, from $i = 0, 1, \cdots$ until $i = N$. The initialization is that $\varsigma_0^0(0,0,0,0)$ is reachable. Once a state $\varsigma_i^k(\mu,\alpha,\beta,\gamma)$ is determined to be reachable, the following two states are *potentially* reachable:

$$\varsigma_{i+1}^k(\mu,\alpha,\beta,\gamma) \quad \text{and} \quad \varsigma_{i+1}^{k+1}(\mu+\mu_{i+1}, \alpha+\alpha_{i+1}, \beta+\beta_{i+1}, \gamma+\gamma_{i+1}).$$

The first future state $\varsigma_{i+1}^k(\mu,\alpha,\beta,\gamma)$ corresponds to the case of $i+1 \notin S$. To determine when such a state is reachable, we review the conditions in Definition 16 and observe that whenever $i+1 \ne q$, the decision $i+1 \notin S$ is legal because q must belong to S whenever $i \ge q$ (note that q is the item in S with the largest estimated utility). The second future state $\varsigma_{i+1}^{k+1}(\mu+\mu_{i+1}, \alpha+\alpha_{i+1}, \beta+\beta_{i+1}, \gamma+\gamma_{i+1})$ corresponds to the case of $i+1 \in S$. Reviewing again the conditions listed in Definition 16, such a state is reachable if $k+1 \le K$ (meaning that there is still room to include a new item in S) and $\widehat{u}_{t,i+1} \le \widehat{u}_{tq}$ (meaning that the new item $(i+1)$ to be included has an estimated utility smaller than $\widehat{u}_{tq}$). Combining both cases, we arrive at the following updated rule of reachability:

1. If $i+1 \ne q$, then $\varsigma_{i+1}^k(\mu,\alpha,\beta,\gamma)$ is reachable;

Input: $\{\widehat{u}_{ti}, r_{ti}, x_{ti}\}_{i=1}^N$, the designated maximum utility item q, and approximation parameter ϵ.
Output: An approximate maximizer $\widehat{S}$ of $\mathrm{ESTR}(\widehat{S}) + \min\{1, \omega \cdot \mathrm{CI}(\widehat{S})\}$

Initialization: compute $\mu_i, \alpha_i, \beta_i, \gamma_i$ for all $i \in [N]$ as in Eq. (21);
Declare $\varsigma_1^0(0,0,0,0)$ as reachable;
for $i = 0, 1, \ldots, N-1$ **do**
  for *all reachable states* $\varsigma_{i-1}^k(\mu, \alpha, \beta, \gamma)$ **do**
    if $i + 1 \neq q$ **then**
      Declare $\varsigma_{i+1}^k(\mu, \alpha, \beta, \gamma)$ as reachable;
    end
    if $\widehat{u}_{t,i+1} \leq \widehat{u}_{tq}$ *and* $k + 1 \leq K$ **then**
      Decare $\varsigma_{i+1}^{k+1}(\mu + \mu_{i+1}, \alpha + \alpha_{i+1}, \beta + \beta_{i+1}, \gamma + \gamma_{i+1})$ as reachable;
    end
  end
end
For all reachable states $\varsigma_N^k(\mu, \alpha, \beta, \gamma)$, trace back the actual assortment $S \subseteq [N]$, $|S| \leq K$ and select the one with the largest $\widehat{\mathrm{ESTR}}(S) + \min\{1, \omega \cdot \widehat{\mathrm{CI}}(S)\}$ as the output $\widehat{S}$.

Algorithm 2: Approximate combinatorial optimization, the univariate ($d = 1$) case, and with the designated maximum utility item.

Input: $\{\widehat{u}_{ti}, r_{ti}, x_{ti}\}_{i=1}^N$ and additive approximation parameter ϵ.
Output: An approximate maximizer $\widehat{S}$ of $\mathrm{ESTR}(\widehat{S}) + \min\{1, \omega \cdot \mathrm{CI}(\widehat{S})\}$.

for $i = 1, 2, \ldots, N$ **do**
  Invoke Algorithm 2 with parameters $q = i$ and ϵ and denote the returned assortment by $\widehat{S}_i$.
end
Among $\widehat{S}_1, \ldots, \widehat{S}_N$, select the one with the largest $\widehat{\mathrm{ESTR}}(S) + \min\{1, \omega \cdot \widehat{\mathrm{CI}}(S)\}$ as the output $\widehat{S}$.

Algorithm 3: Approximate combinatorial optimization, the univariate ($d = 1$) case.

2. If $k < K$ and $\widehat{u}_{t,i+1} \leq \widehat{u}_{tq}$, then $\varsigma_{i+1}^{k+1}(\mu + \mu_{i+1}, \alpha + \alpha_{i+1}, \beta + \beta_{i+1}, \gamma + \gamma_{i+1})$ is reachable.

Algorithms 3 and 2 give pseudo-codes for the proposed dynamic programming approach of computing reachable states and an approximate optimizer of $\widehat{\mathrm{ESTR}}(S) + \min\{1, \omega \cdot \widehat{\mathrm{CI}}(S)\}$.

Finally, we remark on the time complexity of the proposed algorithm. Because the items j we consider in the assortment satisfy $|\widehat{u}_{ti}| \leq U$, $|r_{ti}| \leq 1$, and $|x_{ti}| \leq \nu$, and all $\mu_i, \alpha_i, \beta_i, \gamma_i$ are integral multiples of Δ, we have (1) μ_i and γ_i take at most $O(K\epsilon_0^{-1})$ possible values; (2) α_i takes at most $(K\epsilon_0^{-1})$ possible values; and (3) β_i takes at most $(K\epsilon_0^{-1})$ values. Therefore, the total number of states $\varsigma_i^k(\cdot, \cdot, \cdot, \cdot)$ for fixed $i \in [N] \cup \{0\}$, $k \in [K]$ can be upper bounded by $O(K^8 \epsilon_0^{-4})$. The time complexity of Algorithm 3 is thus upper bounded by $O(K^9 N \epsilon_0^{-4})$.

Input: problem parameters $\{\widehat{u}_{ti}, r_{ti}, x_{ti}\}_{i=1}^{N}$.
Output: approximate maximizer $\widehat{S}$ of $\mathrm{ESTR}(\widehat{S}) + \min\{1, \omega \cdot \mathrm{CI}(\widehat{S})\}$.
Initialization: select $S \subseteq [N]$, $|S| = K$ uniformly at random;
while $\mathrm{ESTR}(S) + \min\{1, \omega \cdot \mathrm{CI}(S)\}$ *can be improved* **do**
  For every $i \notin S$ and $j \in S$, consider new candidate assortments $S' = S \cup \{i\} \backslash \{j\}$ (swapping), $S' = S \cup \{i\}$ if $|S| < K$ (addition) and $S' = S \backslash \{j\}$ if $|S| > 1$ (deletion);
  let S'_* be the considered assortments with the largest $\mathrm{ESTR}(S'_*) + \min\{1, \omega \cdot \mathrm{CI}(S'_*)\}$;
  If S can be improved update $S \leftarrow S'_*$;
end

Algorithm 4: A greedy heuristic for combinatorial assortment optimization

Alternatively, to achieve $(1, \varepsilon, 0)$-approximation, one may set $\epsilon_0 = \min\{\varepsilon/12, \varepsilon^2/(384\omega^2)\}$ as suggested by Corollary 17, resulting in a time complexity of $O(K^9 N \max\{\varepsilon^{-4}, 256\omega^8\varepsilon^{-8}\})$.

This dynamic programming based approximation algorithm can be extended to multivariate feature vector with $d > 1$. The details are presented in Sec. D in the appendix.

5.2 Greedy swapping heuristics

While the proposed approximation has rigorous approximation guarantees and runs in polynomial time, the large time complexity still prohibits its application to moderately large scale problem instances. In this subsection, we consider a practically efficient greedy swapping heuristic to approximately solve the combinatorial optimization problem in Eq. (19).

At a higher level, the heuristic algorithm is a "local search" method similar to the Lloyd's algorithm for K-means clustering (Lloyd, 1982), which continuously tries to improve an assortment solution by considering local swapping/addition/deletions until no further improvements are possible. A pseudo-code description of our heuristic method is given in Algorithm 4.

While the greedy heuristic does not have rigorous guarantees in general, we would like to mention a special case of $\omega = 0$, in which Algorithm 4 does converge to the optimal assortment S maximizing $\mathrm{ESTR}(S) + \min\{1, \omega \cdot \mathrm{CI}(S)\}$ in polynomial time. More specifically, we have the following proposition which is proved in the appendix.

Proposition 18 *If $\omega = 0$, then Algorithm 4 terminates in $O(N^4)$ iterations and produces an output S that maximizes* $\mathrm{ESTR}(S)$.

6. Numerical studies

In this section, we present numerical results of our proposed MLE-UCB algorithm. We use the greedy swapping heuristics (Algorithm 4) as the subroutine to solve the combinatorial optimization problem in Eq. (19). We will also study the quality of the solution of the greedy swapping heuristics.

T	percentile rank					mean relative difference in objective value
	94th	96th	98th	99th	99.5th	
50	0	0.0159	0.0293	0.0393	0.0687	0.00207
200	0	0.0001	0.0040	0.0080	0.0123	0.00024
800	0	0	0	0.0014	0.0037	0.00004

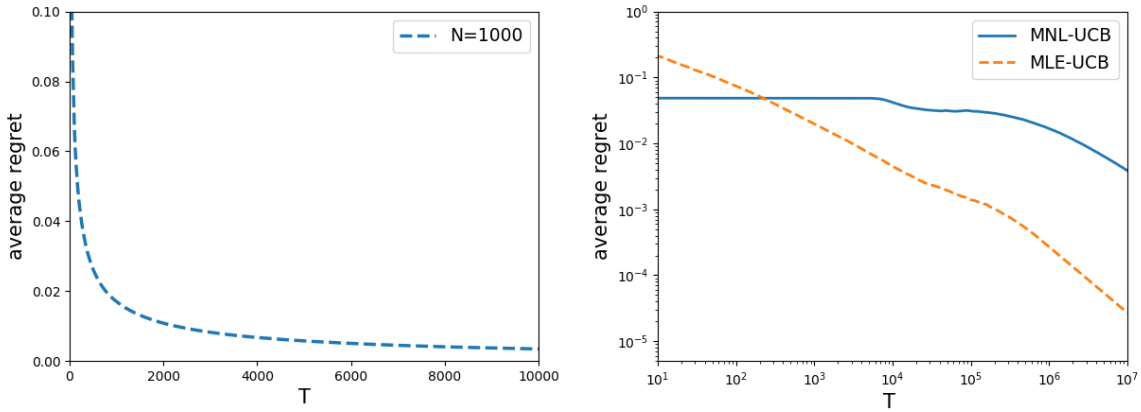
Table 1: relative differences in terms of objective value in Eq. (19) between the greedy swapping algorithm and the optimal solution.

Experiment setup. The unknown model parameter $\theta_0 \in \mathbb{R}^d$ is generated as a uniformly random unit d-dimensional vector. The revenue parameters $\{r_{tj}\}$ for $j \in [N]$ are independently and identically generated from the uniform distribution $[0.5, 0.8]$. For the feature vectors $\{v_{tj}\}$, each of them is independently generated as a uniform random vector v such that $\|v\| = 2$ and $v^\top \theta_0 < -0.6$. Here we set an upper bound of -0.6 for the inner product so that the utility parameters $u_{tj} = \exp\{v_{tj}^\top \theta_0\}$ are upper bounded by $\exp(-0.6) \approx 0.55$. We set such an upper bound because if the utility parameters are uniformly large, the optimal assortment is likely to pick very few items, leading to degenerated problem instances. In the implementation of our MLE-UCB algorithm, we simply set $\omega = \sqrt{d \ln(TK)}$.

The greedy swapping heuristics. We first numerically evaluate the solution quality of the greedy swapping heuristic algorithm by focusing on the optimization problem in Eq. (19). We compare the obtained objective values in Eq. (19) to the proposed greedy heuristic and the optimal solution (obtained by brute-force search). Instead of generating purely random instances, we consider more realistic instances generated from a dynamic assortment planning process. In particular, for a given T, we generate a dynamic assortment optimization problem with parameters $N = 10$, $K = 4$ and $d = 5$, and run the MLE-UCB algorithm till the T-th time period. Now the combinatorial optimization problem in Eq. (19) to be solved at the T-th time is kept as one testing instance for the greedy swapping algorithm.

For each $T \in \{50, 200, 800\}$, we generate 1000 such test instances, and compare the solution of the greedy swapping heuristics with the optimal solution obtained by brute-force search in terms of the objective value in Eq. (19). Table 1 shows the relative differences between the two solutions at several percentiles, and the mean relative differences. We can see that the approximation quality of the greedy swapping algorithm has already been desirable when $T = 50$, and becomes even better as T grows.

Performance of the MLE-UCB algorithm. In Figure 1a we plot the average regret (i.e. regret/T) of MLE-UCB algorithm with $N = 1000, K = 10, d = 5$ for the first $T = 10000$ time periods. For each experiment (in both Figure 1a and other figures), we repeat the experiment for 100 times and report the average value. In Figure 1b we compare our algorithm with the UCB algorithm for multinomial logit bandit (MNL-UCB) from Agrawal et al. (2019) without utilizing the feature information. Since the MNL-UCB algorithm assumes fixed item utilities that do not change over time, in this experiment we randomly generate one feature vector for each of the $N = 1000$ items and this feature vector will be fixed for the entire time span. We can observe that our MLE-UCB algorithm performs



(a) Average regret of MLE-UCB (b) Comparison of MLE-UCB and MNL-UCB

Figure 1: Illustration of the performance of MLE-UCB.

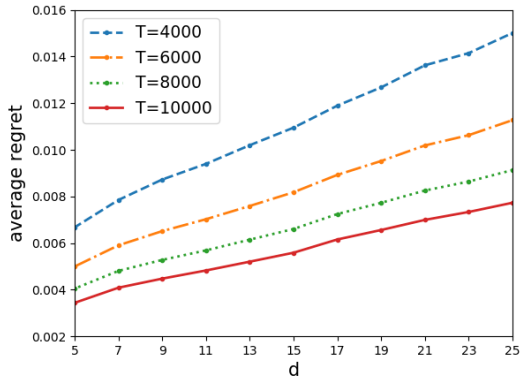


Figure 2: Average regret of MLE-UCB for various d's.

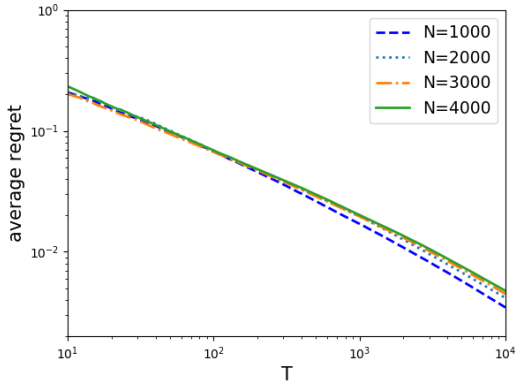


Figure 3: Average regret of MLE-UCB for various N's.

much better than MNL-UCB, which suggests the importance of taking the advantage of the contextual information.

Impact of the dimension size d. We study how the dimension of the feature vector impacts the performance of our MLE-UCB algorithm. We fix $N = 1000$ and $K = 10$ and test our algorithm for dimension sizes in $5, 7, 9, 11, \ldots, 25$. In Figure 2, we report the average regret at times $T \in \{4000, 6000, 8000, 10000\}$. We can see that the average regret increases approximately linearly with d. This phenomenon matches the linear dependency on d of the main term of the regret Eq. (6) of the MLE-UCB.

Impact of the number of items N. We compare the performance of our MLE-UCB algorithm for the varying number of items N. We fix $K = 10$ and $d = 5$, and test MLE-UCB for $N \in \{1000, 2000, 3000, 4000\}$. In Figure 3, we report the average regret for the first $T = 10000$ time periods. We observe that the regret of the algorithm is almost not

affected by a bigger N. This confirms the fact that the regret Eq. (6) of MLE-UCB is totally independent of N.

7. Conclusion and future directions

In this paper, we study the dynamic assortment planning problem under a contextual MNL model, which incorporates rich feature information into choice modeling. We propose an upper confidence bound (UCB) algorithm based on the local MLE that simultaneously learns the underlying coefficient and makes the decision on the assortment selection. We establish both the upper and lower bounds of the regret. Moreover, we develop an approximation algorithm and a greedy heuristic for solving the key optimization problem in our UCB algorithm.

There are a few possibilities for future work. Technically, there is still a gap of $1/K$ between our upper and lower bounds on regret. Although the cardinality constraint of an assortment K is usually small in practice, it is still a technically interesting question to close this gap. Second, introducing contextual information into choice model is a natural idea for many online applications. This paper explores the standard MNL model, and it would be interesting to extend this work to contextual nested logit and other popular choice models. Finally, it is interesting to incorporate other operational considerations into the model, such as prices or inventory constraints.

Acknowledgments

The authors would like to thank Vineet Goyal for helpful discussions, and Zikai Xiong for helping with the numerical studies. Xi Chen would like to thank the support of NSF (via Grant IIS-1845444). Yuan Zhou is supported by NSF via Grant CCF-2006526.

Appendix A. Proofs of technical lemmas for Theorem 1 (upper bound)

A.1 Proof of Lemma 3

Lemma 19 (restated) *With probability* $1 - O(T^{-1})$ *it holds that*

$$\|\theta^* - \theta_0\|_2 = O((\lambda_0^{-1} d \log T/T_0)^{1/4}). \tag{23}$$

Proof. The proof is built on the same framework as in the proof of Lemma 5, to be presented in later sections of this file. It is easy to verify that all steps up to Eq. (29) remain valid, and we have

$$F_t(\widehat{\theta}_t) = -\frac{1}{2}(\widehat{\theta}_t - \theta_0)^\top I_t(\overline{\theta}_t)(\widehat{\theta}_t - \theta_0),$$

for some $\overline{\theta}_t = \alpha\theta_0 + (1-\alpha)\widehat{\theta}_t$, $\alpha \in (0,1)$. Recall also the definition that $I_t(\overline{\theta}_t) = \sum_{t'=1}^{t} -\nabla_\theta^2 f_{t'}(\overline{\theta}_t)$ where

$$\begin{aligned} -\nabla_\theta^2 f_{t'}(\overline{\theta}_t) = &-\mathbb{E}_{\theta_0,t'}[v_{t'j} v_{t'j}^\top] + \{\mathbb{E}_{\theta_0,t'} v_{t'j}\}\{\mathbb{E}_{\overline{\theta}_t,t'} v_{t'j}\}^\top \\ &+ \{\mathbb{E}_{\overline{\theta}_t,t'} v_{t'j}\}\{\mathbb{E}_{\theta_0,t'} v_{t'j}\}^\top - \{\mathbb{E}_{\overline{\theta}_t,t'} v_{t'j}\}\{\mathbb{E}_{\overline{\theta}_t,t'} v_{t'j}\}^\top. \end{aligned}$$

Using elementary algebra, we can further show that

$$\begin{aligned}
&\mathbb{E}_{\theta_0,t'}[v_{t'j}v_{t'j}^\top] - \mathbb{E}_{\theta_0,t'}[v_{t'j}]\mathbb{E}_{\overline{\theta}_t,t'}[v_{t'j}]^\top - \mathbb{E}_{\overline{\theta}_t,t'}[v_{t'j}]\mathbb{E}_{\theta_0,t'}[v_{t'j}]^\top + \mathbb{E}_{\overline{\theta}_t,t'}[v_{t'j}]\mathbb{E}_{\overline{\theta}_t,t'}[v_{t'j}]^\top \\
&= \mathbb{E}_{\theta_0,t'}[v_{t'j}v_{t'j}^\top] - \mathbb{E}_{\overline{\theta}_t,t'}[v_{t'j}]\mathbb{E}_{\theta_0,t'}[v_{t'j}]^\top + (\mathbb{E}_{\theta_0,t'}[v_{t'j}] - \mathbb{E}_{\overline{\theta}_t,t'}[v_{t'j}])(\mathbb{E}_{\theta_0,t'}[v_{t'j}] - \mathbb{E}_{\overline{\theta}_t,t'}[v_{t'j}])^\top \\
&\succeq \mathbb{E}_{\theta_0,t'}[v_{t'j}v_{t'j}^\top] - \mathbb{E}_{\theta_0,t'}[v_{t'j}]\mathbb{E}_{\theta_0,t'}[v_{t'j}]^\top = \mathbb{E}_{\theta_0,t'}\{(v_{t'j} - \mathbb{E}_{\theta_0,t'}v_{t'j})(v_{t'j} - \mathbb{E}_{\theta_0,t'}v_{t'j})^\top\} \\
&\succeq \frac{\lambda_0 K}{K+e} I_{d\times d},
\end{aligned}$$

where the last inequality holds thanks to Assumption (A1) and the fact that S_t consists of K items sampled uniformly at random from $[N]$. Subsequently, we have that

$$\frac{1}{2}(\widehat{\theta}_t - \theta_0)^\top I_t(\overline{\theta}_t)(\widehat{\theta}_t - \theta_0) \succeq \frac{\lambda_0 K T_0}{2(K+e)} I_{d\times d} \succeq \frac{\lambda_0 T_0}{8} I_{d\times d}. \tag{24}$$

Additionally, because $p_{\theta,t'}(j|S_{t'}) \geq 1/[e(1+eK)]$ for all $\|\theta\|_2 \leq 1$, t' and $|S_{t'}| \leq K$, we have that $|\widehat{f}_{t'}(\theta)| \leq O(\ln K)$ almost surely. Using standard Hoeffding's inequality together with covering numbers of $\{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$, we have with probability $1 - O(T^{-1})$ that

$$\sup_{\|\theta\|_2\leq 1} \left|\widehat{F}_t(\theta) - F_t(\theta)\right| \leq O(\sqrt{dT_0 \log T}). \tag{25}$$

Combining Eqs. (24,25) and noting that $\widehat{F}_t(\widehat{\theta}_t) \geq 0$, we have with probability $1 - O(T^{-1})$ that

$$\|\widehat{\theta}_t - \theta_0\|_2 \leq O\left(\left[\frac{d\log T}{\lambda_0 T_0}\right]^{1/4}\right),$$

which is to be demonstrated. □

A.2 Proof of Lemma 5

Lemma 20 (restated) *Suppose $\tau \leq 1/(15K)$. Then there exists a universal constant $C > 0$ such that with probability $1 - O(T^{-1})$ the following holds uniformly over all $t = T_0, \cdots, T-1$:*

$$(\widehat{\theta}_t - \theta_0)^\top I_t(\theta_0)(\widehat{\theta}_t - \theta_0) \leq C \cdot d\log(TK). \tag{26}$$

Proof. For any $\theta \in \mathbb{R}^d$ define

$$f_{t'}(\theta) := \mathbb{E}_{\theta_0,t'}\left[\log \frac{p_{\theta,t'}(j|S_{t'})}{p_{\theta_0,t'}(j|S_{t'})}\right] = \sum_{j\in S_{t'}\cup\{0\}} p_{\theta_0,t'}(j|S_{t'}) \log \frac{p_{\theta,t'}(j|S_{t'})}{p_{\theta_0,t'}(j|S_{t'})}.$$

By simple algebra calculations, the first and second order derivatives of $f_{t'}$ with respect to θ can be computed as

$$\nabla_\theta f_{t'}(\theta) = \mathbb{E}_{\theta_0,t'}[v_{t'j}] - \mathbb{E}_{\theta,t'}[v_{t'j}]; \tag{27}$$

$$\begin{aligned}
\nabla_\theta^2 f_{t'}(\theta) = &-\mathbb{E}_{\theta_0,t'}[v_{t'j}v_{t'j}^\top] + \{\mathbb{E}_{\theta_0,t'}v_{t'j}\}\{\mathbb{E}_{\theta,t'}v_{t'j}\}^\top \\
&+ \{\mathbb{E}_{\theta,t'}v_{t'j}\}\{\mathbb{E}_{\theta_0,t'}v_{t'j}\}^\top - \{\mathbb{E}_{\theta,t'}v_{t'j}\}\{\mathbb{E}_{\theta,t'}v_{t'j}\}^\top.
\end{aligned} \tag{28}$$

In the rest of the section we drop the subscript in ∇_θ, ∇^2_θ, and the ∇, ∇^2 notations should always be understood as with respect to θ.

Define $F_t(\theta) := \sum_{t'=1}^{t} f_{t'}(\theta)$. It is easy to verify that $-F_t(\theta)$ is the Kullback-Leibler divergence between the conditional distribution of $(i_1, \cdots, i_t)$ parameterized by θ and θ_0, respectively. Therefore, $F_t(\theta)$ is always non-positive. Note also that $F_t(\theta_0) = 0$, $\nabla F_t(\theta_0) = 0$, $\nabla^2 f_{t'}(\theta) = -M_{t'}(\theta)$ and $\nabla^2 F_t(\theta) \equiv -I_t(\theta)$. By Taylor expansion with Lagrangian remainder, there exists $\overline{\theta}_t = \alpha\theta_0 + (1-\alpha)\widehat{\theta}_t$ for some $\alpha \in (0,1)$ such that

$$F_t(\widehat{\theta}_t) = -\frac{1}{2}(\widehat{\theta}_t - \theta_0)^\top I_t(\overline{\theta}_t)(\widehat{\theta}_t - \theta_0). \tag{29}$$

Our next lemma shows that, if $\overline{\theta}_t$ is close to θ_0 (guaranteed by the constraint that $\|\widehat{\theta}_t - \theta^*\|_2 \le \tau$), then $I_t(\overline{\theta}_t)$ can be *spectrally* lower bounded by $I_t(\theta_0)$. It is proved in the appendix.

Lemma 21 *Suppose $\tau \le 1/(8K)$. Then $I_t(\overline{\theta}_t) \succeq \frac{1}{2} I_t(\theta_0)$ for all t.*

As a corollary of Lemma 21, we have

$$F_t(\widehat{\theta}_t) \le -\frac{1}{4}(\widehat{\theta}_t - \theta_0)^\top I_t(\theta_0)(\widehat{\theta}_t - \theta_0). \tag{30}$$

On the other hand, consider the "empirical" version $\widehat{F}_t(\theta) := \sum_{t'=1}^{t} \widehat{f}_{t'}(\theta)$, where

$$\widehat{f}_{t'}(\theta) := \log \frac{p_{\theta,t'}(i_{t'}|S_{t'})}{p_{\theta_0,t'}(i_{t'}|S_{t'})}. \tag{31}$$

It is easy to verify that $\widehat{F}_t(\theta_0) = 0$ remains true; in addition, for any *fixed* $\theta \in \mathbb{R}^d$, $\{\widehat{F}_t(\theta)\}_t$ forms a *martingale* [3] and satisfies $\mathbb{E}\widehat{F}_t(\theta) = F_t(\theta)$ for all t. This leads to our following lemma, which upper bounds the *uniform* convergence of $\widehat{F}_t(\theta)$ towards $F_t(\theta)$ for all $\|\theta - \theta_0\| \le 2\tau$.

Lemma 22 *Suppose $\tau \le 1/(15K)$. Then there exists a universal constant $C > 0$ such that with probability $1 - O(T^{-1})$ the following holds uniformly for all $t \in \{T_0 + 1, \cdots, T\}$ and $\|\theta - \theta_0\|_2 \le 2\tau$:*

$$\left|\widehat{F}_t(\theta) - F_t(\theta)\right| \le C\left[d\log(TK) + \sqrt{|F_t(\theta)| d\log(TK)}\right]. \tag{32}$$

Lemma 22 can be proved by using a standard ε-net argument. Since the complete proof is quite involved, we defer it to the appendix.

We are now ready to prove Lemma 5. By Eq. (32) and the fact that $\widehat{F}_t(\widehat{\theta}_t) \le 0 \le F_t(\widehat{\theta}_t)$, we have

$$|F_t(\widehat{\theta}_t)| \le |\widehat{F}_t(\widehat{\theta}_t) - F_t(\widehat{\theta}_t)| \lesssim d\log(TK) + \sqrt{|F_t(\widehat{\theta}_t)| d\log(TK)}. \tag{33}$$

Subsequently,

$$|F_t(\widehat{\theta}_t)| \lesssim d\log(NT). \tag{34}$$

3. $\{X_k\}_k$ forms a martingale if $\mathbb{E}[X_{k+1}|X_1, \cdots, X_k] = X_k$ for all k.

In addition, because $F_t(\widehat{\theta}_t) \leq 0$, by Eq. (30) we have

$$-\frac{1}{2}(\widehat{\theta}_t - \theta_0)^\top I_t(\theta_0)(\widehat{\theta}_t - \theta_0) \geq F_t(\widehat{\theta}_t) \geq d\log(TK). \tag{35}$$

Lemma 5 is thus proved. □

Proof of Lemma 21

Lemma 23 (restated) *Suppose $\tau \leq 1/(8K)$. Then $I_t(\overline{\theta}_t) \succeq \frac{1}{2}I_t(\theta_0)$ for all t.*

Proof. Because $\widehat{\theta}_t$ is a feasible solution of the local MLE, we know $\|\widehat{\theta}_t - \theta^*\|_2 \leq \tau$. Also by Corollary 4 we know that $\|\theta^* - \theta_0\|_2 \leq \tau$ with high probability. By triangle inequality and the definition of $\overline{\theta}_t$ we have that $\|\overline{\theta}_t - \theta_0\|_2 \leq 2\tau$.

To prove $I_t(\overline{\theta}_t) \succeq \frac{1}{2}I_t(\theta_0)$ we only need to show that $M_{t'}(\overline{\theta}_t) - M_{t'}(\theta_0) \preceq \frac{1}{2}M_{t'}(\theta_0)$ for all $1 \leq t' \leq t$. This reduces to proving

$$\{\mathbb{E}_{\overline{\theta}_t,t'}v_{t'j} - \mathbb{E}_{\theta_0,t'}v_{t'j}\}\{\mathbb{E}_{\overline{\theta}_t,t'}v_{t'j} - \mathbb{E}_{\theta_0,t'}v_{t'j}\}^\top \preceq \frac{1}{2}\mathbb{E}_{\theta_0,t'}\left[(v_{t'j} - \mathbb{E}_{\theta_0,t'}v_{t'j})(v_{t'j} - \mathbb{E}_{\theta_0,t'}v_{t'j})^\top\right]. \tag{36}$$

Fix arbitrary $S_{t'} \subseteq [N]$, $|S_{t'}| = J \leq K$ and for convenience denote $x_1, \cdots, x_J \in \mathbb{R}^d$ as the feature vectors of items in $S_{t'}$ (i.e., $\{v_{t'j}\}_{j\in S_{t'}}$). Let also $p_{\theta_0}(j)$ and $p_{\overline{\theta}_t}(j)$ be the probability of choosing action $j \in [J]$ corresponding to x_j parameterized by θ_0 or $\overline{\theta}_t$. Define $\overline{x} := \sum_{j=1}^J p_{\theta_0}(j)x_j$, $w_j := x_j - \overline{x}$ and $\delta_j := p_{\overline{\theta}_t}(j) - p_{\theta_0}(j)$. Recall also that $x_0 = 0$ and $w_0 = -\overline{x}$. Eq. (36) is then equivalent to

$$\left\{\sum_{j=0}^J \delta_j w_j\right\}\left\{\sum_{j=0}^J \delta_j w_j\right\}^\top \preceq \frac{1}{2}\sum_{j=0}^J p_{\theta_0}(j)w_j w_j^\top. \tag{37}$$

Let $L = \mathrm{span}\{w_j\}_{j=0}^J$ and $H \in \mathbb{R}^{L\times d}$ be a whitening matrix such that $H(\sum_j p_{\theta_0}(j)w_j w_j^\top)H^\top = I_{L\times L}$, where $I_{L\times L}$ is the identity matrix of size L. Denote $\widetilde{w}_j := Hw_j$. We then have $\sum_{j=0}^J p_{\theta_0}(j)\widetilde{w}_j\widetilde{w}_j^\top = I_{L\times L}$. Eq. (37) is then equivalent to

$$\left\|\sum_{j=0}^J \delta_j \widetilde{w}_j\right\|_2^2 \leq \frac{1}{2}. \tag{38}$$

On the other hand, by (A2) we know that $p_{\theta_0}(j) \geq 1/(e^2K)$ for all j and therefore $\|\widetilde{w}_j\|_2 \leq \sqrt{e^2K}$ for all j. Subsequently, we have

$$\left\|\sum_{j=0}^J \delta_j \widetilde{w}_j\right\|_2^2 \leq \left(\max_j |\delta_j| \cdot \sum_{j=0}^J \|\widetilde{w}_j\|_2\right)^2 \leq \max_j |\delta_j|^2 \cdot e^2K^2. \tag{39}$$

Recall that $\delta_i = p_{\overline{\theta}_t}(i) - p_{\theta_0}(i)$ where $p_\theta(i) = \exp\{x_i^\top\theta\}/(1 + \sum_{j\in S_{t'}}\exp\{x_i^\top\theta\})$. Simple algebra yields that $\nabla_\theta p_\theta(i) = p_\theta(i)[x_i - \mathbb{E}_\theta x_j]$, where $\mathbb{E}_\theta x_j = \sum_{j\in S_{t'}} p_\theta(j)x_j$. Using the mean-value theorem, there exists $\widetilde{\theta}_t = \widetilde{\alpha}\overline{\theta}_t + (1-\widetilde{\alpha})\theta_0$ for some $\widetilde{\alpha} \in (0,1)$ such that

$$\delta_i = \langle \nabla_\theta p_{\widetilde{\theta}_t}(i), \widehat{\theta}_t - \theta_0\rangle = p_{\widetilde{\theta}_t}(i)\langle x_i - \mathbb{E}_{\widetilde{\theta}_t}x_j, \overline{\theta}_t - \theta_0\rangle. \tag{40}$$

Because $\|x_{ti}\|_2 \le 1$ almost surely for all $t \in [T]$ and $i \in [N]$, we have

$$\max_j |\delta_j|^2 \cdot eK^2 \le 4 \cdot \max_i \|x_i\|_2^2 \cdot \|\overline{\theta}_t - \theta_0\|_2^2 \cdot e^2K^2 \le 4e^2K^2 \cdot \tau^2. \tag{41}$$

The lemma is then proved by plugging in the condition on τ. □

Proof of Lemma 22

Lemma 24 (restated) *Suppose $\tau \le 1/(15K)$. Then there exists a universal constant $C > 0$ such that with probability $1 - O(T^{-1})$ the following holds uniformly for all $t \in \{T_0 + 1, \cdots, T\}$ and $\|\theta - \theta_0\|_2 \le 2\tau$:*

$$\left|\widehat{F}_t(\theta) - F_t(\theta)\right| \le C\left[d\log(TK) + \sqrt{|F_t(\theta)| d\log(TK)}\right]. \tag{42}$$

Proof. We first consider a *fixed* $\theta \in \mathbb{R}^d$, $\|\theta - \theta_0\|_2 \le 2\tau$. Define

$$\mathcal{M} := \max_{t' \le t} |\widehat{f}_{t'}(\theta)| \quad \text{and} \quad \mathcal{V}^2 := \sum_{t'=1}^{t} \mathbb{E}_{j \sim \theta_0, t'} \left|\log \frac{p_{\theta,t'}(j|S_{t'})}{p_{\theta_0,t'}(j|S_{t'})}\right|^2. \tag{43}$$

Using an Azuma-Bernstein type inequality (see, for example, (Fan et al., 2015, Theorem A), (Freedman, 1975, Theorem (1.6))), we have

$$\left|\widehat{F}_t(\theta) - F_t(\theta)\right| \lesssim \mathcal{M}\log(1/\delta) + \sqrt{\mathcal{V}^2 \log(1/\delta)} \quad \text{with probability } 1 - \delta. \tag{44}$$

The following lemma upper bounds $\mathcal{M}$ and $\mathcal{V}^2$ using $F_t(\theta)$ and the fact that θ is close to θ_0. It will be proved right after this proof.

Lemma 25 *If $\tau \le 1/(15K)$ then $\mathcal{M} \le 1$ and $\mathcal{V}^2 \le 8|F_t(\theta)|$.*

Corollary 26 *Suppose τ satisfies the condition in Lemma 25. Then for any $\|\theta - \theta_0\|_2 \le 2\tau$,*

$$\left|\widehat{F}_t(\theta) - F_t(\theta)\right| \lesssim \log(1/\delta) + \sqrt{|F_t(\theta)| \log(1/\delta)} \quad \textit{with probability } 1 - \delta. \tag{45}$$

Our next step is to construct an ϵ-net over $\{\theta \in \mathbb{R}^d : \|\theta - \theta_0\|_2 \le 2\tau\}$ and apply union bound on the constructed ϵ-net. This together with a deterministic perturbation argument delivers *uniform* concentration of $\widehat{F}_t(\theta)$ towards $F_t(\theta)$.

For any $\epsilon > 0$, let $\mathcal{H}(\epsilon)$ be a finite covering of $\{\theta \in \mathbb{R}^d : \|\theta - \theta_0\|_2 \le 2\tau\}$ in $\|\cdot\|_2$ up to precision ϵ. That is, $\sup_{\|\theta - \theta_0\|_2 \le 2\tau} \min_{\theta' \in \mathcal{H}(\epsilon)} \|\theta - \theta'\|_2 \le \epsilon$. By standard covering number arguments (e.g., (van de Geer, 2000)), such a finite covering set $\mathcal{H}(\epsilon)$ exists whose size can be upper bounded by $\log|\mathcal{H}(\epsilon)| \lesssim d\log(\tau/\epsilon)$. Subsequently, by Corollary 26 and the union bound, we have with probability $1 - O(T^{-1})$ that

$$\left|\widehat{F}_t(\theta) - F_t(\theta)\right| \lesssim d\log(T/\epsilon) + \sqrt{|F_t(\theta)| d\log(T/\epsilon)} \quad \forall T_0 < t \le T, \theta \in \mathcal{H}(\epsilon). \tag{46}$$

On the other hand, with probability $1 - O(T^{-1})$ such that Eq. (40) holds, we have for arbitrary $\|\theta - \theta'\|_2 \le \epsilon$ that

$$\begin{aligned}
\left|\widehat{F}_t(\theta) - \widehat{F}_t(\theta')\right| &\le t \cdot \sup_{t' \le t, j \in S_{t'} \cup \{0\}} \left| \log \frac{p_{\theta,t'}(j|S_{t'})}{p_{\theta',t'}(j|S_{t'})} \right| \\
&\le t \cdot \sup_{t' \le t, j \in S_{t'} \cup \{0\}} \frac{|p_{\theta,t'}(j|S_{t'}) - p_{\theta',t'}(j|S_{t'})|}{p_{\theta',t'}(j|S_{t'})} \qquad (47) \\
&\le 2e^2 TK \cdot \sup_{t' \le t, j \in S_{t'} \cup \{0\}} \left|p_{\theta,t'}(j|S_{t'}) - p_{\theta',t'}(j|S_{t'})\right| \qquad (48) \\
&\le 2e^2 TK \cdot \sup_{t' \le t, j \in [N]} 4\|v_{t'j}\|_2^2 \cdot \|\theta - \theta'\|_2 \\
&\lesssim TK\epsilon. \qquad (49)
\end{aligned}$$

Here Eq. (47) holds because $\log(1+x) \le x$; Eq. (48) holds because $p_{\theta',t'}(j|S_{t'}) \ge p_{\theta_0,t'}(j|s_{t'}) - |p_{\theta',t'}(j|S_{t'}) - p_{\theta_0,t'}(j|S_{t'})| \ge 1/(2e^2K)$ thanks to (A2) and Eq. (41).

Combining Eqs. (46,49) and setting $\epsilon \asymp 1/(TK)$ we have with probability $1 - O(T^{-1})$ that

$$\left|\widehat{F}_t(\theta) - F_t(\theta)\right| \lesssim d\log(TK) + \sqrt{|F_t(\theta)| d\log(TK)} \quad \forall T_0 < t \le T, \|\theta - \theta_0\|_2 \le 2\tau, \qquad (50)$$

which is to be demonstrated in Lemma 22. □

Proof of Lemma 25

Lemma 27 (restated) *If $\tau \le 1/(15K)$ then $\mathcal{M} \le 1$ and $\mathcal{V}^2 \le 8|F_t(\theta)|$.*

Proof. We first derive an upper bound for M. By (A2), we know that $p_{\theta_0,t'}(j|S_{t'}) \ge 1/(e^2K)$ for all j. Also, Eqs. (40,41) shows that $|p_{\theta,t'}(j|S_{t'}) - p_{\theta_0,t'}(j|S_{t'})| \le 4\tau^2$. If $\tau^2 \le 1/(8e^2K)$ we have $|p_{\theta,t'}(j|S_{t'}) - p_{\theta_0,t'}(j|S_{t'})| \le 0.5 p_{\theta_0,t'}(j|S_{t'})$ and therefore $|\widehat{f}_{t'}(\theta)| \le \log^2 2 \le 1$.

We next give upper bounds on $\mathcal{V}^2$. Fix arbitrary t', and for notational simplicity let $p_j = p_{\theta_0,t'}(j|S_{t'})$ and $q_j = p_{\theta,t'}(j|S_{t'})$. Because $\log(1+x) \le x$ for all $x \in (-1, \infty)$, we have

$$\mathbb{E}_{j \sim \theta_0,t'} \left| \log \frac{p_{\theta,t'}(j|S_{t'})}{p_{\theta_0,t'}(j|S_{t'})} \right|^2 = \sum_{j \in S_{t'} \cup \{0\}} p_j \log^2 \left(1 + \frac{q_j - p_j}{p_j}\right) \le \sum_{j \in S_{t'} \cup \{0\}} \frac{(q_j - p_j)^2}{p_j}. \qquad (51)$$

On the other hand, by Taylor expansion we know that for any $x \in (-1, \infty)$, there exists $\overline{x} \in (0, x)$ such that $\log(1+x) = x - x^2/2(1+\overline{x})^2$. Subsequently,

$$\begin{aligned}
-f_{t'}(\theta) = -\mathbb{E}_{j \sim \theta_0,t'} \left[\log \frac{p_{\theta,t'}(j|S_{t'})}{p_{\theta_0,t'}(j|S_{t'})}\right] &= -\sum_{j \in S_{t'} \cup \{0\}} p_j \log\left(1 + \frac{q_j - p_j}{p_j}\right) \qquad (52) \\
&= -\sum_{j \in S_{t'} \cup \{0\}} p_j \left(\frac{q_j - p_j}{p_j} - \frac{1}{2(1+\overline{\delta}_j)^2} \frac{|q_j - p_j|^2}{p_j^2}\right) \qquad (53) \\
&\ge \frac{1}{2(1 + \max_j |p_j - q_j|/p_j)^2} \cdot \sum_{j \in S_{t'} \cup \{0\}} \frac{(q_j - p_j)^2}{p_j}. \qquad (54)
\end{aligned}$$

Here $\overline{\delta}_j \in (0, (q_j - p_j)/p_j)$ and the last inequality holds because $\sum_j p_j = \sum_j q_j = 1$.

By Eqs. (40) and (41), we have that $|q_j - p_j|^2 \le 4\tau^2$. In addition, (A2) implies that $p_j \ge 1/(e^2 K)$ for all j. Therefore, if $\tau \le 1/(15K)$ we have $|p_j - q_j|/p_j \le 1$ for all j and hence

$$\mathbb{E}_{j\sim\theta_0,t'}\left|\log\frac{p_{\theta,t'}(j|S_{t'})}{p_{\theta_0,t'}(j|S_{t'})}\right|^2 \le \sum_{j\in S_{t'}\cup\{0\}}\frac{(q_j-p_j)^2}{p_j} \le 8|f_{t'}(\theta)|. \tag{55}$$

Summing over all $t' = 1, \cdots, t$ and noting that $f_{t'}(\theta)$ is always non-positive, we complete the proof of Lemma 25. □

A.3 Proof of Lemma 7

Lemma 28 (restated) *Suppose τ satisfies the condition in Lemma 5. With probability $1 - O(T^{-1})$ the following holds uniformly for all $t > T_0$ and $S \subseteq [N]$, $|S| \le K$ such that*

1. $\overline{R}_t(S) \ge R_t(S)$;
2. $|\overline{R}_t(S) - R_t(S)| \lesssim \min\{1, \omega\sqrt{\|I_{t-1}^{-1/2}(\theta_0) M_t(\theta_0|S) I_{t-1}^{-1/2}(\theta_0)\|_{\rm op}}\}$.

Proof. Without explicit clarification, all statements are conditioned on the success event in Lemma 5, which occurs with probability $1 - O(T^{-1})$ if τ is sufficiently large and satisfies the condition in Lemma 5.

We present below a key technical lemma in the proof of Lemma 7, which is an upper bound on the absolute value difference between $R_t(S) := \mathbb{E}_{\theta_0,t}[r_{tj}|S]$ and $\widehat{R}_t(S) := \mathbb{E}_{\widehat{\theta}_{t-1},t}[r_{tj}|S]$ using $I_{t-1}(\theta_0)$ and $M_t(\theta_0|S)$, where $I_{t-1}(\theta) = \sum_{t'=1}^{t-1} M_{t'}(\theta)$ and $M_{t'}(\theta) = \mathbb{E}_{\theta_0,t'}[v_{t'j}v_{t'j}^\top] - \{\mathbb{E}_{\theta_0,t'}v_{t'j}\}\{\mathbb{E}_{\theta,t'}v_{t'j}\}^\top - \{\mathbb{E}_{\theta,t'}v_{t'j}\}\{\mathbb{E}_{\theta_0,t'}v_{t'j}\}^\top + \{\mathbb{E}_{\theta,t'}v_{t'j}\}\{\mathbb{E}_{\theta,t'}v_{t'j}\}^\top$. This key lemma can be regarded as a finite sample version of the celebrated *Delta's method* (e.g., (Van der Vaart, 2000)) used widely in classical statistics to estimate and/or infer a functional of unknown quantities.

Lemma 29 *For all $t > T_0$ and $S \subseteq [N]$, $|S| \le K$, it holds that $|\widehat{R}_t(S) - R_t(S)| \lesssim \sqrt{d\log(TK)} \cdot \sqrt{\|I_{t-1}^{-1/2}(\theta_0) M_t(\theta_0|S) I_{t-1}^{-1/2}(\theta_0)\|_{\rm op}}$, where in $\lesssim$ notation we only hide numerical constants.*

Below we state our proof of Lemma 29, while deferring the proof of some detailed technical lemmas to the appendix. Fix $S \subseteq [N]$. We use $\mathfrak{R}_t(\theta) = \mathbb{E}_{\theta,t}[r_{tj}] = [\sum_{j\in S} r_{tj}\exp\{v_{tj}^\top\theta\}]/[1 + \sum_{j\in S}\exp\{v_{tj}^\top\theta\}]$ to denote the expected revenue of assortment S at time t, evaluated using a specific model $\theta \in \mathbb{R}$. Then

$$\begin{aligned}\nabla_\theta\mathfrak{R}_t(\theta) &= \frac{\sum_{j\in S} r_{tj}\exp\{v_{tj}^\top\theta\}(1+\sum_{j\in S}\exp\{v_{tj}^\top\theta\})^2 - (\sum_{j\in S} r_{tj}\exp\{v_{tj}^\top\theta\})(\sum_{j\in S}\exp\{v_{tj}^\top\theta\})}{(1+\sum_{j\in S}\exp\{v_{tj}^\top\theta\})^2} \\ &= \mathbb{E}_{\theta,t}[r_{tj}v_{tj}] - \{\mathbb{E}_{\theta,t}r_{tj}\}\{\mathbb{E}_{\theta,t}v_{tj}\}. \end{aligned} \tag{56}$$

By the mean value theorem, there exists $\widetilde{\theta}_{t-1} = \theta_0 + \xi(\widehat{\theta}_{t-1} - \theta_0)$ for some $\xi \in (0,1)$ such that

$$\begin{aligned}\left|\widehat{R}_t(S) - R_t(S)\right| &= \left|\mathfrak{R}_t(\widehat{\theta}_{t-1}) - \mathfrak{R}_t(\theta_0)\right| = \left|\langle \nabla\mathfrak{R}_t(\widetilde{\theta}_{t-1}), \widehat{\theta}_{t-1} - \theta_0\rangle\right| \\ &= \sqrt{(\widehat{\theta}_{t-1} - \theta_0)^\top [\nabla\mathfrak{R}_t(\widetilde{\theta}_{t-1})\nabla\mathfrak{R}_t(\widetilde{\theta}_{t-1})^\top)](\widehat{\theta}_{t-1} - \theta_0)}. \end{aligned} \tag{57}$$

Recall that $\nabla\mathfrak{R}_t(\widetilde{\theta}_{t-1}) = \mathbb{E}_{\widetilde{\theta}_{t-1},t}[r_{tj}v_{tj}] - \{\mathbb{E}_{\widetilde{\theta}_{t-1},t} r_{tj}\}\{\mathbb{E}_{\widetilde{\theta}_{t-1},t} v_{tj}\} = \mathbb{E}_{\widetilde{\theta}_{t-1},t}[(r_{tj} - \mathbb{E}_{\widetilde{\theta}_{t-1},t} r_{tj})(v_{tj} - \mathbb{E}_{\widetilde{\theta}_{t-1},t} v_{tj})]$. Subsequently, by Jenson's inequality and the fact that $r_{tj} \in [0,1]$ almost surely,

$$\begin{aligned}\nabla\mathfrak{R}_t(\widetilde{\theta}_{t-1})\nabla\mathfrak{R}_t(\widetilde{\theta}_{t-1})^\top &\preceq \mathbb{E}_{\widetilde{\theta}_{t-1,t}}\left[(r_{tj} - \mathbb{E}_{\widetilde{\theta}_{t-1},t} r_{tj})^2 (v_{tj} - \mathbb{E}_{\widetilde{\theta}_{t-1},t} v_{tj})(v_{tj} - \mathbb{E}_{\widetilde{\theta}_{t-1},t} v_{tj})^\top\right] \\ &\preceq \mathbb{E}_{\widetilde{\theta}_{t-1,t}}\left[(v_{tj} - \mathbb{E}_{\widetilde{\theta}_{t-1},t} v_{tj})(v_{tj} - \mathbb{E}_{\widetilde{\theta}_{t-1},t} v_{tj})^\top\right] = \widehat{M}_t(\widetilde{\theta}_{t-1}|S). \end{aligned} \tag{58}$$

Define $\widehat{M}_t(\theta|S) := \mathbb{E}_{\theta,t}[(v_{tj} - \mathbb{E}_{\theta,t} v_{tj})(v_{tj} - \mathbb{E}_{\theta,t} v_{tj})^\top]$, where $S \subseteq [N]$ is the assortment supplied at iteration t. Combining Eqs. (57,58) with Lemma 5, we have

$$\left|\widehat{R}_t(S) - R_t(S)\right| \lesssim \sqrt{d\log(NT)} \cdot \sqrt{\|I_{t-1}(\theta_0)^{-1/2} \widehat{M}_t(\widetilde{\theta}_{t-1}|S) I_{t-1}(\theta_0)^{-1/2}\|_{\mathrm{op}}}. \tag{59}$$

It remains to show that $\widehat{M}_t(\widetilde{\theta}_{t-1}|S)$ and $M_t(\theta_0|S)$ are close, for which we first recall the definitions of both quantities:

$$\begin{aligned}\widehat{M}_t(\widetilde{\theta}_{t-1}|S) &= \mathbb{E}_{\widetilde{\theta}_{t-1,t}}\left[(v_{tj} - \mathbb{E}_{\widetilde{\theta}_{t-1},t} v_{tj})(v_{tj} - \mathbb{E}_{\widetilde{\theta}_{t-1},t} v_{tj})^\top\right]; \\ M_t(\theta_0|S) &= \mathbb{E}_{\theta_0,t}[v_{tj} v_{tj}^\top] - \{\mathbb{E}_{\theta_0,t} v_{tj}\}\{\mathbb{E}_{\theta_0,t} v_{tj}\}^\top = \widehat{M}_t(\theta_0|S).\end{aligned}$$

The next lemma shows that under suitable conditions $\widehat{M}_t(\widetilde{\theta}_{t-1}|S)$ is close to $\widehat{M}_t(\theta_0|S) = M_t(\theta_0|S)$, implying that $\frac{1}{4}M_t(\theta_0|S) \preceq \widehat{M}_t(\widetilde{\theta}_{t-1}|S) \preceq 4M_t(\theta_0|S)$. It is proved in the appendix.

Lemma 30 *Suppose $\tau \le 1/(30K)$. Then $\frac{1}{4}M_t(\theta_0|S) \preceq \widehat{M}_t(\widetilde{\theta}_{t-1}|S) \preceq 4M_t(\theta_0|S)$ for all t, S and θ.*

As a consequence of Lemma 30, the right-hand side of Eq. (59) can be upper bounded by

$$\sqrt{d\log(TK)} \cdot \sqrt{4\|I_{t-1}(\theta_0)^{-1/2} M_t(\theta_0|S) I_{t-1}(\theta_0)^{-1/2}\|_{\mathrm{op}}}.$$

Lemma 29 is thus proved. We are now ready to prove Lemma 7. By Lemma 29, we know that with high probability

$$\left|\widehat{R}_t(S) - R_t(S)\right| \lesssim \sqrt{d\log(TK)} \cdot \sqrt{\|I_{t-1}(\theta_0)^{-1/2} M_t(\theta_0|S) I_{t-1}(\theta_0)^{-1/2}\|_{\mathrm{op}}} \tag{60}$$

In addition, by Lemma 30 and the fact that $\|\widehat{\theta}_{t-1} - \theta_0\|_2 \le \tau$ thanks to the local MLE formulation, we have $\frac{1}{4}M_t(\theta_0|S) \preceq \widehat{M}_t(\widehat{\theta}_{t-1}|S) \preceq 4M_t(\theta_0|S)$ and subsequently $\frac{1}{4}I_{t-1}(\theta_0) \preceq \widehat{I}_{t-1}(\widehat{\theta}_{t-1}) \preceq 4I_{t-1}(\theta_0)$ because $I_{t-1}(\cdot)$ and $\widehat{I}_{t-1}(\cdot)$ are summations of $M_{t'}(\cdot)$ and $\widehat{M}_{t'}(\cdot)$ terms. Setting $\omega \gtrsim \sqrt{d\log(TK)}$ we proved that $\overline{R}_t(S) \ge R_t(S)$. The second property of Lemma 7 can be proved similarly, by invoking the spectral similarities between $I_{t-1}(\cdot)$, $M_{t'}(\cdot)$ and $\widehat{I}_{t-1}(\cdot)$, $\widehat{M}_{t'}(\cdot)$. □

Proof of Lemma 30

Lemma 31 (restated) *Suppose $\tau \le 1/(30K)$. Then $\frac{1}{4}M_t(\theta_0|S) \preceq \widehat{M}_t(\widetilde{\theta}_{t-1}|S) \preceq 4M_t(\theta_0|S)$ for all t, S and θ.*

Proof. Define $\overline{M}_t(\theta|S) := \mathbb{E}_{\theta_0,t}[(v_{tj} - \mathbb{E}_{\theta,t}v_{tj})(v_{tj} - \mathbb{E}_{\theta,t}v_{tj})^\top]$, where only the outermost expectation is replaced by taking with respect to the probability law under θ_0. Denote also $\widetilde{w}_j := v_{tj} - \mathbb{E}_{\theta,t}v_{tj}$. Then $\overline{M}_t(\theta|S) = \sum_j p_{\theta_0,t}(j)\widetilde{w}_j\widetilde{w}_j^\top$ and $\overline{M}_t(\theta|S) - \widehat{M}_t(\theta|S) = \sum_j \delta_j \widetilde{w}_j\widetilde{w}_j^\top$, where $\delta_j = p_{\theta_0,t}(j) - p_{\theta,t}(j)$. By Eq. (40) and the fact that $\|v_{ti}\|_2 \le 1$, $\|\theta - \theta_0\|_2 \le \tau$, we have

$$\max_j |\delta_j| \le 2\tau. \tag{61}$$

On the other hand, by (A2) we know that $\min_j p_{\theta_0,t}(j) \ge 1/(e^2K)$ and therefore

$$\overline{M}_t(\theta|S) = \sum_j p_{\theta_0,t}\widetilde{w}_j\widetilde{w}_j^\top \succeq \frac{1}{e^2K}\sum_j \widetilde{w}_j\widetilde{w}_j^\top. \tag{62}$$

Combining Eqs. (61,62) and the fact that $\overline{M}_t(\theta|S) - \widehat{M}_t(\theta|S) = \sum_j \delta_j\widetilde{w}_j\widetilde{w}_j^\top$, we have $\overline{M}_t(\theta|S) - \widehat{M}_t(\theta|S) \preceq \overline{M}_t(\theta|S)/2$ and $\widehat{M}_t(\theta|S) - \overline{M}_t(\theta|S) \preceq \overline{M}_t(\theta|S)/2$, provided that $\tau \le 1/(30K)$. This also implies $\frac{1}{2}\overline{M}_t(\theta|S) \preceq \widehat{M}_t(\theta|S) \preceq 2\overline{M}_t(\theta|S)$.

We next prove that $\frac{1}{2}M_t(\theta_0|S) \preceq \overline{M}_t(\theta|S) \preceq 2M_t(\theta_0|S)$ which, together with $\frac{1}{2}\overline{M}_t(\theta|S) \preceq \widehat{M}_t(\theta|S) \preceq 2\overline{M}_t(\theta|S)$ established in the previous section, implies Lemma 30. Recall the definitions that

$$M_t(\theta_0|S) = \mathbb{E}_{\theta_0,t}\left[(v_{tj} - \mathbb{E}_{\theta_0,t}v_{tj})(v_{tj} - \mathbb{E}_{\theta_0,t}v_{tj})^\top\right];$$
$$\overline{M}_t(\theta|S) = \mathbb{E}_{\theta_0,t}\left[(v_{tj} - \mathbb{E}_{\theta,t}v_{tj})(v_{tj} - \mathbb{E}_{\theta,t}v_{tj})^\top\right].$$

Adding and subtracting $\mathbb{E}_{\theta,t}v_{tj}, \mathbb{E}_{\theta_0,t}v_{tj}$ terms, we have

$$\begin{aligned}
&\overline{M}_t(\theta|S) - M_t(\theta_0|S)\\
&= \mathbb{E}_{\theta_0,t}\left[(v_{tj} - \mathbb{E}_{\theta_0,t}v_{tj} + \mathbb{E}_{\theta_0,t}v_{tj} - \mathbb{E}_{\theta,t}v_{tj})(v_{tj} - \mathbb{E}_{\theta_0,t}v_{tj} + \mathbb{E}_{\theta_0,t}v_{tj} - \mathbb{E}_{\theta,t}v_{tj})^\top\right]\\
&\quad - \mathbb{E}_{\theta_0,t}\left[(v_{tj} - \mathbb{E}_{\theta_0,t}v_{tj})(v_{tj} - \mathbb{E}_{\theta_0,t}v_{tj})^\top\right]\\
&= \mathbb{E}_{\theta_0,t}\left[(\mathbb{E}_{\theta_0,t}v_{tj} - \mathbb{E}_{\theta,t}v_{tj})(v_{tj} - \mathbb{E}_{\theta_0,t}v_{tj})^\top\right] + \mathbb{E}_{\theta_0,t}\left[(v_{tj} - \mathbb{E}_{\theta_0,t}v_{tj})(\mathbb{E}_{\theta_0,t}v_{tj} - \mathbb{E}_{\theta,t}v_{tj})^\top\right]\\
&\quad + (\mathbb{E}_{\theta_0,t}v_{tj} - \mathbb{E}_{\theta,t}v_{tj})(\mathbb{E}_{\theta_0,t}v_{tj} - \mathbb{E}_{\theta,t}v_{tj})^\top\\
&= (\mathbb{E}_{\theta_0,t}v_{tj} - \mathbb{E}_{\theta,t}v_{tj})(\mathbb{E}_{\theta_0,t}v_{tj} - \mathbb{E}_{\theta,t}v_{tj})^\top.
\end{aligned}$$

By Eq. (36) in the proof of Lemma 21, we have that

$$(\mathbb{E}_{\theta_0,t}v_{tj} - \mathbb{E}_{\theta,t}v_{tj})(\mathbb{E}_{\theta_0,t}v_{tj} - \mathbb{E}_{\theta,t}v_{tj})^\top \lesssim \frac{1}{2}\mathbb{E}_{\theta_0,t}[(v_{tj} - \mathbb{E}_{\theta_0,t}v_{tj})(v_{tj} - \mathbb{E}_{\theta_0,t}v_{tj})^\top] = \frac{1}{2}M_t(\theta_0|S)$$

provided that $\tau \le 1/(30K)$, thus implying $\frac{1}{2}M_t(\theta_0|S) \preceq \overline{M}_t(\theta|S) \preceq 2M_t(\theta_0|S)$. □

A.4 Proof of Lemma 8

Lemma 32 (restated) *It holds that*

$$\sum_{t=T_0+1}^{T} \min\{1, \|I_{t-1}^{-1/2}(\theta_0) M_t(\theta_0|S_t) I_{t-1}^{-1/2}(\theta_0)\|_{\mathrm{op}}^2\} \le 4 \log \frac{\det I_T(\theta_0)}{\det I_{T_0}(\theta_0)} \lesssim d \log(\lambda_0^{-1}).$$

Proof. Denote $A_t := I_{t-1}^{-1/2}(\theta_0) M_t(\theta_0|S_t) I_{t-1}^{-1/2}(\theta_0)$ as d-dimensional positive semi-definite matrices with eigenvalues sorted as $\sigma_1(A_t) \ge \cdots \ge \sigma_d(A_t) \ge 0$. By simple algebra,

$$\begin{aligned}\sum_{t=T_0+1}^{T} \min\{1, \|I_{t-1}^{-1/2}(\theta_0) M_t(\theta_0|S_t) I_{t-1}^{-1/2}(\theta_0)\|_{\mathrm{op}}^2\} &= \sum_{t=T_0+1}^{T} \min\{1, \sigma_1(A_t)^2\} \\ &\le \sum_{t=T_0+1}^{T} 2\log(1+\sigma_1(A_t)^2) \le \sum_{t=T_0+1}^{T} 4\log(1+\sigma_1(A_t)). \end{aligned} \tag{63}$$

On the other hand, note that $I_t(\theta_0) = I_{t-1}(\theta_0) + M_t(\theta_0|S_t) = I_{t-1}(\theta_0)^{1/2}[I_{d\times d} + A_t] I_{t-1}(\theta_0)^{1/2}$. Hence,

$$\log\det I_t(\theta_0) = \log\det I_{t-1}(\theta_0) + \sum_{j=1}^{d} \log(1+\sigma_j(A_t)). \tag{64}$$

Comparing Eqs. (63) and (64), we have

$$\sum_{t=T_0+1}^{T} \min\{1, \|I_{t-1}^{-1/2}(\theta_0) M_t(\theta_0|S_t) I_{t-1}^{-1/2}(\theta_0)\|_{\mathrm{op}}^2\} \le 4 \log \frac{\det I_T(\theta_0)}{\det I_{T_0}(\theta_0)}, \tag{65}$$

which proves the first inequality in Lemma 8.

We next prove the second inequality in Lemma 8. Because assortments have size 1 throughout the pure exploration phase ($t \le T_0$), we have

$$I_{T_0}(\theta_0) = \sum_{t=1}^{T_0} p_{\theta_0,t}(j_t)(1-p_{\theta_0,t}(j_t))^2 v_{t,j_t} v_{t,j_t}^\top \ge \frac{1}{(1+e^2)^3} \cdot \sum_{t=1}^{T_0} v_{t,j_t} v_{t,j_t}^\top, \tag{66}$$

where the last inequality holds thanks to assumption (A2), which implies $p_{\theta_0,t}(j_t) \in [1/(1+e), e/(1+e^{-1})]$. In addition, by the proof of Corollary 4, with high probability $\lambda_{\min}(\sum_{t=1}^{T_0} v_{t,j_t} v_{t,j_t}^\top) \ge 0.5 T_0 \lambda_0$, where $\lambda_0 > 0$ is a parameter specified in assumption (A1). Therefore,

$$\det I_{T_0}(\theta_0) \gtrsim [T_0\lambda_0]^d. \tag{67}$$

On the other hand, because $\max_{t,j} \|v_{tj}\|_2 \le 1$ we have $I_T(\theta_0) \lesssim T$ and subsequently

$$\det I_T(\theta_0) \lesssim T^d. \tag{68}$$

Combining Eqs. (67) and (68) we proved the second inequality in Lemma 8. □

Appendix B. Proofs of technical lemmas for Theorem 9 (lower bound)

B.1 Proof of Lemma 11

Lemma 33 (restated) *Suppose $\epsilon \in (0, 1/d\sqrt{d})$ and define $\delta := d/4 - |\widetilde{U}_t \cap W|$. Then*

$$R(S^*_{\theta_W}) - R(\widetilde{S}_t) \geq \frac{\delta\epsilon}{4K\sqrt{d}}.$$

Proof. Let $v = v_W$ and $\widehat{v} = v_{\widetilde{U}_t}$ be the corresponding feature vectors. Then

$$\begin{aligned} R(S^*_{\theta_W}) - R(\widetilde{S}_t) &= \frac{K\exp\{v^\top\theta_W\}}{1+K\exp\{v^\top\theta_W\}} - \frac{K\exp\{\widehat{v}^\top\theta_W\}}{1+K\exp\{\widehat{v}^\top\theta_W\}} \\ &= \frac{K[\exp\{v^\top\theta_W\} - \exp\{\widehat{v}^\top\theta_W\}]}{(1+K\exp\{v^\top\theta_W\})(1+K\exp\{\widehat{v}^\top\theta_W\})} \\ &\geq \frac{\exp\{v^\top\theta_W\} - \exp\{\widehat{v}^\top\theta_W\}}{2Ke}. \end{aligned}$$

Here the last inequality holds because $\max(\exp\{v^\top\theta_W\}, \exp\{\widehat{v}^\top\theta_W\}) \leq e$. In addition, by Taylor expansion we know that $1+x \leq e^x \leq 1+x+x^2/2$ for all $x \in [0,1]$. Subsequently,

$$R(S^*_{\theta_W}) - R(\widetilde{S}_t) \geq \frac{(v-\widehat{v})^\top\theta_W - (\widehat{v}^\top\theta_W)^2/2}{2Ke} \geq \frac{\delta\epsilon/\sqrt{d} - (\sqrt{d}\epsilon)^2/2}{2Ke}.$$

Finally, noting that $d\epsilon^2/2 \leq \delta\epsilon/2\sqrt{d}$ provided that $\epsilon \in (0, 1/d\sqrt{d})$, we finish the proof of Lemma 11. □

B.2 Proof of Lemma 12

Lemma 34 (restated) *For any $W \in \mathcal{W}_{d/4-1}$ and $i \in [d]$, $\mathrm{KL}(P_W \| P_{W\cup\{i\}}) \leq C_{\mathrm{KL}} \cdot \mathbb{E}_W[N_i] \cdot \epsilon^2/d$ for some universal constant $C_{\mathrm{KL}} > 0$.*

Proof. Fix a time t with policy's assortment choice S_t, and define $n_i(S_t) := \sum_{v_U \in S_t} \mathbf{1}\{i \in U\}/K$. Let $\{p_j\}_{j\in S_t\cup\{0\}}$ and $\{q_j\}_{j\in S_t\cup\{0\}}$ be the probabilities of purchasing item j under parameterization θ_W and $\theta_{W\cup\{i\}}$, respectively. Then

$$\mathrm{KL}(P_W(\cdot|S_t)\|P_{W\cup\{i\}}(\cdot|S_t)) = \sum_{j\in S_t\cup\{0\}} p_j \log\frac{q_j}{p_j} \leq \sum_j p_j \frac{p_j - q_j}{q_j} \leq \sum_j \frac{|p_j - q_j|^2}{q_j}, \quad (69)$$

where the only inequality holds because $\log(1+x) \leq x$ for all $x > -1$. Because $q_j \geq e^{-1}/(1+Ke) \geq 1/(2Ke^2)$ for all $j \in S_t \cup \{0\}$, Eq. (69) is reduced to

$$\mathrm{KL}(P_W(\cdot|S_t)\|P_{W\cup\{i\}}(\cdot|S_t)) \leq 2e^2K \cdot \sum_{j\in S_t\cup\{0\}} |p_j - q_j|^2. \quad (70)$$

We next upper bound $|p_j - q_j|$ separately. First consider $j = 0$. We have

$$\begin{aligned}
|p_j - q_j| &= \left| \frac{1}{1 + \sum_{j \in S_t} \exp\{v_j^\top \theta_W\}} - \frac{1}{1 + \sum_{j \in S_t} \exp\{v_j^\top \theta_{W \cup \{i\}}\}} \right| \\
&\leq \frac{1}{(1 + K/e)^2} \cdot 2 \sum_{j \in S_t} \left| v_j^\top (\theta_W - \theta_{W \cup \{i\}}) \right| \\
&\leq \frac{2K n_i(S_t)\epsilon/\sqrt{d}}{(1 + K/e)^2} \leq \frac{8e^2 n_i(S_t)\epsilon}{K\sqrt{d}}.
\end{aligned}$$

Here the first inequality holds because $e^x \leq 1 + 2x$ for all $x \in [0, 1]$.
For $j > 0$ corresponding to $v_j = v_U$ where $i \notin U$, we have

$$\begin{aligned}
|p_j - q_j| &= \left| \frac{\exp\{v_U^\top \theta_W\}}{1 + \sum_{j \in S_t} \exp\{v_j^\top \theta_W\}} - \frac{\exp\{v_U^\top \theta_{W \cup \{i\}}\}}{1 + \sum_{j \in S_t} \exp\{v_j^\top \theta_{W \cup \{i\}}\}} \right| \\
&\leq \left| \frac{1}{1 + \sum_{j \in S_t} \exp\{v_j^\top \theta_W\}} - \frac{1}{1 + \sum_{j \in S_t} \exp\{v_j^\top \theta_{W \cup \{i\}}\}} \right| \\
&\leq \frac{8e^2 n_i(S_t)\epsilon}{K\sqrt{d}}.
\end{aligned}$$

Here the first inequality holds because $\exp\{v_U^\top \theta_W\} = \exp\{v_U^\top \theta_{W \cup \{i\}}\} \leq 1$, since $i \notin U$.
For $j > 0$ corresponding to $v_j = v_U$ and $i \in U$, we have

$$\begin{aligned}
|p_j - q_j| &= \left| \frac{\exp\{v_U^\top \theta_W\}}{1 + \sum_{j \in S_t} \exp\{v_j^\top \theta_W\}} - \frac{\exp\{v_U^\top \theta_{W \cup \{i\}}\}}{1 + \sum_{j \in S_t} \exp\{v_j^\top \theta_{W \cup \{i\}}\}} \right| \\
&\leq \exp\{v_u^\top \theta_{W \cup \{i\}}\} \cdot \left| \frac{1}{1 + \sum_{j \in S_t} \exp\{v_j^\top \theta_W\}} - \frac{1}{1 + \sum_{j \in S_t} \exp\{v_j^\top \theta_{W \cup \{i\}}\}} \right| \\
&\quad + \left| \exp\{v_u^\top \theta_W\} - \exp\{v_u^\top \theta_{W \cup \{i\}}\} \right| \cdot \left| \frac{1}{1 + \sum_{j \in S_t} \exp\{v_j^\top \theta_W\}} \right| \\
&\leq \frac{8e^2 n_i(S_t)\epsilon}{K\sqrt{d}} + \frac{\epsilon}{\sqrt{d}} \cdot \frac{1}{1 + K/e}. \leq \frac{8e^2 n_i(S_t)\epsilon}{K\sqrt{d}} + \frac{2e\epsilon}{K\sqrt{d}}.
\end{aligned}$$

Combining all upper bounds on $|p_j - q_j|$ and Eq. (70), we have

$$\begin{aligned}
\mathrm{KL}(P_W(\cdot|S_t) \| P_{W \cup \{i\}}(\cdot|S_t)) &\leq 2e^2 K \cdot \left[\frac{128 e^4 n_i(S_t)^2 \epsilon^2}{K^2 d}(1 + K) + K n_i(S_t) \cdot \frac{8e^4 \epsilon^2}{K^2 d} \right] \\
&\lesssim n_i(S_t)\epsilon^2/d.
\end{aligned}$$

Here the last inequality holds because $n_i(S_t) \leq 1$. Note also that $N_i = \sum_{t=1}^T n_i(S_t)$ by definition, and subsequently summing over all $t = 1$ to T we have

$$\mathrm{KL}(P_W \| P_{W \cup \{i\}}) \lesssim \mathbb{E}_W[N_i] \cdot \epsilon^2/d,$$

which is to be demonstrated. □

Appendix C. Proofs of approximation algorithms

C.1 Proof of Lemma 14

Lemma 35 (restated) *Suppose an $(\alpha, \varepsilon, \delta)$-approximation algorithm is used instead of exact optimization in the MLE-UCB policy at each time period t. Then its regret can be upper bounded by*

$$\alpha \cdot \mathrm{Regret}^* + \varepsilon T + \delta T^2 + O(1),$$

where Regret^* *is the regret upper bound shown by Theorem 1 for Algorithm 1 with exact optimization in Step 8.*

Proof. By union bound, we know the approximation guarantee in Eq. (20) *for all* t with probability at least $1-\delta T$. In the event of failure, the accumulated regret is upper bounded by T almost surely, because the regret incurred by each time period t is at most 1. This gives rises to the δT^2 term in Lemma 14, and in the rest of the proof we shall assume Eq. (20) holds for all t.

Let S_t^* be the solution to the exact optimization problem in Step 8 of Algorithm 1, $S_t^\#$ be the assortment with the optimal revenue the same step, and $\widehat{S}_t$ be the solution by an $(\alpha, \varepsilon, \delta)$-approximation algorithm.

For each $t > T_0$, we bound the expected regret incurred at time t by

$$\begin{aligned}
&\mathrm{ESTR}(S_t^\#) - \mathrm{ESTR}(S_t) \\
=&\Big(\mathrm{ESTR}(S_t^\#) + \min\{1 + \omega \cdot \mathrm{CI}(S_t^\#)\}\Big) - (\mathrm{ESTR}(S_t^*) + \min\{1 + \omega \cdot \mathrm{CI}(S_t^*)\}) \\
&\quad + (\mathrm{ESTR}(S_t^*) + \min\{1 + \omega \cdot \mathrm{CI}(S_t^*)\}) - (\mathrm{ESTR}(S_t) + \min\{1 + \alpha\omega \cdot \mathrm{CI}(S_t)\} - \varepsilon) \\
&\quad + \Big(\varepsilon + \min\{1 + \alpha\omega \cdot \mathrm{CI}(S_t)\} - \min\{1 + \omega \cdot \mathrm{CI}(S_t^\#)\}\Big) \\
\le&\varepsilon + \min\{1 + \alpha\omega \cdot \mathrm{CI}(S_t)\} - \min\{1 + \omega \cdot \mathrm{CI}(S_t^\#)\} \le \varepsilon + \min\{1 + \alpha\omega \cdot \mathrm{CI}(S_t)\}.
\end{aligned}$$

Therefore, the total expected regret is bounded by

$$T_0 + \sum_{t=T_0+1}^{T} (\varepsilon + \min\{1 + \alpha\omega \cdot \mathrm{CI}(S_t)\}) \le \varepsilon T + T_0 + \alpha \sum_{t=T_0+1}^{T} \min\{1 + \omega \cdot \mathrm{CI}(S_t)\},$$

which, by the same analysis in Section 3.2.4, can be bounded by $\alpha \cdot \mathrm{Regret}^* + \varepsilon T$. □

C.2 Proof of Lemma 15

Lemma 36 (restated) *For any $S \subseteq [N]$, $|S| \le K$, suppose $U = \max_{j\in S}\{1, \widehat{u}_{tj}\}$ and $\Delta = \epsilon_0 U/K$ for some $\epsilon_0 > 0$. Suppose also $|x_{tj}| \le \nu$ for all t, j. Then*

$$\big|\mathrm{ESTR}(S) - \widehat{\mathrm{ESTR}}(S)\big| \le 6\epsilon_0 \quad \text{and} \quad \big|\mathrm{CI}(S) - \widehat{\mathrm{CI}}(S)\big| \le \sqrt{24\epsilon_0}(1+\nu), \tag{71}$$

Proof. We first prove the upper bound on $|\mathrm{ESTR}(S) - \widehat{\mathrm{ESTR}(S)}|$, which is

$$\left|\frac{\sum_{i\in S} \widehat{u}_{ti} r_{ti}}{1 + \sum_{i\in S} \widehat{u}_{ti}} - \frac{\sum_{i\in S} \gamma_i}{1 + \sum_{i\in S} \mu_i}\right| \le 6\epsilon_0, \tag{72}$$

where $\mu_i = [\widehat{u}_{ti}/\Delta]\cdot\Delta$, $\gamma_i = [\widehat{u}_{ti}r_{ti}/\Delta]\cdot\Delta$.

Denote $A := \sum_{i\in S}\widehat{u}_{ti}r_{ti}$ and $B := 1+\sum_{i\in S}\widehat{u}_{ti}$. Because $r_{ti}\le 1$, we have $A\le B$. Let also $\tau_1 := \sum_{i\in S}\gamma_i - A$ and $\tau_2 := 1+\sum_{i\in S}\mu_i - B$. Because $\max\{|\gamma_i-\widehat{u}_{ti}r_{ti}|, |\mu_i-\widehat{u}_{ti}|\}\le\Delta$, we have $\max\{|\tau_1|,|\tau_2|\}\le\Delta\cdot K$. Subsequently,

$$\begin{aligned}\left|\frac{\sum_{i\in S}\widehat{u}_{ti}r_{ti}}{1+\sum_{i\in S}\widehat{u}_{ti}} - \frac{\sum_{i\in S}\gamma_i}{1+\sum_{i\in S}\mu_i}\right| &= \left|\frac{A}{B}-\frac{A+\tau_1}{B+\tau_2}\right| = \left|\frac{A\tau_2-B\tau_1}{B(B+\tau_2)}\right| = \left|\frac{A\tau_2-B\tau_2+B\tau_2-B\tau_1}{B(B+\tau_2)}\right| \\ &\le \left|\frac{(A-B)\tau_2}{B(B+\tau_2)}\right| + \frac{|\tau_1|+|\tau_2|}{B-|\tau_2|} \le \frac{|\tau_1|+2|\tau_2|}{B-|\tau_2|},\end{aligned}$$

where the last inequality holds because $A\le B$. Using $B = 1+\sum_{i\in S}\widehat{u}_{ti}\ge 1+\widehat{u}_{tq}\ge U$ (since $q\in S$ and $U=\max\{1,u_{tq}\}$, and $\max\{|\tau_1|,|\tau_2|\}\le\Delta\cdot K=\epsilon_0 U$, we have

$$\frac{|\tau_1|+2|\tau_2|}{B-|\tau_2|}\le\frac{3\epsilon_0 U}{U-\epsilon_0 U}\le 6\epsilon_0,$$

provided that $\epsilon_0\in(0,1/2]$. Eq. (72) is thus proved.

We next prove the upper bound on $|\mathrm{CI}(S)-\widehat{\mathrm{CI}}(S)|$, which is

$$\left|\sqrt{\frac{\sum_{i\in S}\widehat{u}_{ti}x_{it}^2}{1+\sum_{i\in S}\widehat{u}_{ti}} - \left(\frac{\sum_{i\in S}\widehat{u}_{ti}x_{ti}}{1+\sum_{i\in S}\widehat{u}_{ti}}\right)^2} - \sqrt[*]{\frac{\sum_{i\in S}\beta_i}{1+\sum_{i\in S}\mu_i} - \left(\frac{\sum_{i\in S}\alpha_i}{1+\sum_{i\in S}u_i}\right)^2}\right| \le \sqrt{24\epsilon_0}(1+\nu), \tag{73}$$

where $\sqrt[*]{\cdot} = \sqrt{\max\{0,\cdot\}}$, $\mu_i = [\widehat{u}_{ti}/\Delta]\cdot\Delta$, $\alpha_i = [\widehat{u}_{ti}x_{ti}/\Delta]\cdot\Delta$, $\beta_i = [\widehat{u}_{ti}x_{ti}^2/\Delta]\cdot\Delta$.

Denote $C := \frac{\sum_{i\in S}\widehat{u}_{ti}x_{ti}}{1+\sum_{i\in S}\widehat{u}_{ti}}$ and $D := \frac{\sum_{i\in S}\widehat{u}_{ti}x_{ti}^2}{1+\sum_{i\in S}\widehat{u}_{ti}}$. Because $|x_{ti}|\le\nu$ for all t and i, we have $C\in[-\nu,\nu]$ and $D\in[0,\nu^2]$. Denote also $\tau_3 := \frac{\sum_{i\in S}\alpha_i}{1+\sum_{i\in S}\mu_i} - C$ and $\tau_4 := \frac{\sum_{i\in S}\beta_i}{1+\sum_{i\in S}\mu_i} - D$. Using the same analysis as in the proof of Eq. (72), we have $|\tau_3|\le 6\epsilon_0(1+\nu)$ and $|\tau_4|\le 6\epsilon_0(1+\nu^2)$.

With the definitions of C, D, τ_3 and τ_4, the left-hand side of Eq. (73) can be re-written as

$$\left|\sqrt{D-C^2} - \sqrt[*]{(D+\tau_4)-(C+\tau_3)^2}\right|. \tag{74}$$

Case 1: $D-C^2 > -(\tau_4-2\tau_3 C-\tau_3^2)$**.** In this case, we have

$$\begin{aligned}\text{Eq. (74)} &= \frac{|\tau_4-2\tau_3C-\tau_3^2|}{\sqrt{D-C^2}+\sqrt{D-C^2+(\tau_4-2\tau_3C-\tau_3^2)}} \le \sqrt{|\tau_4-2\tau_3C-\tau_3^2|} \\ &\le \sqrt{6\epsilon_0(1+\nu^2)+2\cdot 6\epsilon_0(1+\nu)^2+6\epsilon_0(1+\nu)} \le \sqrt{24\epsilon_0}(1+\nu).\end{aligned}$$

Case 2: $D-C^2\le -(\tau_4-2\tau_3C-\tau_3^2)$**.** In this case, we have $(D+\tau_4)-(C+\tau_3)^2\le 0$ and subsequently

$$\text{Eq. (74)} = \sqrt{D-C^2} \le \sqrt{|\tau_4-2\tau_3C-\tau_3^2|} \le \sqrt{24\epsilon_0}(1+\nu).$$

Combining both cases we prove Eq. (73). □

C.3 Proof of Proposition 18

Proposition 37 (restated) *If $\omega = 0$, then Algorithm 4 terminates in $O(N^4)$ iterations and produces an output S that maximizes* $\mathrm{ESTR}(S)$.

Proof. We first show that when the algorithm terminates with $\mathrm{ESTR}(S) = r$, S is one of the optimal assortments. Suppose S is not an optimal assortment, i.e. there exists $S^\#$ such that $\mathrm{ESTR}(S^\#) > r$, we show that the algorithm will not terminate. By the definition of $\mathrm{ESTR}(\cdot)$ we have $\sum_{i\in S^\#} \widehat{u}_{ti}(r_{ti} - r) > r$ and $\sum_{i\in S} \widehat{u}_{ti}(r_{ti} - r) = r$. By comparing $S^\#$ and S, one can find a new candidate assortment S' via swapping, adding, or deleting an item from/to S such that $\sum_{i\in S'} \widehat{u}_{ti}(r_{ti} - r) > r$. Therefore, $\mathrm{ESTR}(S') > r$ and the algorithm will not terminate.

It remains to show that the algorithm terminates in $O(N^4)$ iterations.

For each $r \in [0, 1]$, we define a total order $\geq_r$ on $[N] \cup \{\perp\}$, where $[N]$ corresponds to the N items and $\perp$ is a special element with the definition $\widehat{u}_{t\perp} = r_{t\perp} = 0$ for convenience, as follows: $i \geq_r j$ if and only if $\widehat{u}_{ti}(r_{ti} - r) \geq \widehat{u}_{tj}(r_{tj} - r)$ (and consequently $i >_r j$ if and only if $\widehat{u}_{ti}(r_{ti} - r) > \widehat{u}_{tj}(r_{tj} - r)$). It is straightforward to verify that there exists $O(N^2)$ section points $\theta_0 < \theta_1 < \theta_2 < \cdots < \theta_{L-1} < \theta_L = 1$ so that for any two r_1, r_2 that sandwiched by the same pair of neighboring section points (i.e. $\exists \ell \in [L] : r_1, r_2 \in (\theta_{\ell-1}, \theta_\ell)$), we have $\geq_{r_1} \equiv \geq_{r_2}$. Indeed, one can set the section points to be the solutions to the equalities $\widehat{u}_{ti}(r_{ti} - r) = \widehat{u}_{tj}(r_{tj} - r)$ for every pair of $i, j \in [N] \cup \{\perp\}$.

We will show that if $\mathrm{ESTR}(S) \in (\theta_{\ell-1}, \theta_\ell)$ for some $\ell \in [L]$, after at most $O(N^2)$ iterations, either the algorithm terminates or $\mathrm{ESTR}(S) \geq \theta_\ell$. This directly leads to an $O(N^4)$ upper bound on the total number of iterations that the algorithm performs. We pick an arbitrary $r \in (\theta_{\ell-1}, \theta_\ell)$ and define the following two potential functions: $I(S) = |\{(i, j) : i \in S, j \in [N]\backslash S, i <_r j\}|$, and $J(S) = |\{i \in S : i <_r \perp\}|$. We have the following observations:

- When a swapping operation is performed on S, $I(S)$ strictly decreases and $J(S)$ does not increase.
- When a deletion operation is performed on S, $I(S)$ increases by at most N and $J(S)$ strictly decreases.
- When an addition operation is performed on S, $I(S)$ increases by at most N and $J(S)$ does not increase.

We let $F(S) = I(S) + (2N+1) \cdot J(S) \in [0, O(N^2)]$. Suppose there are a swapping operations, b deletion operations, and c addition operations done in total, S is the assortment that the algorithm begins with and T is the last assortment satisfying $\mathrm{ESTR}(T) < \theta_\ell$. Observe that there are at most $c \leq b + N$ addition operations. Together with the three observations above, we have

$$\begin{aligned} 0 \leq F(T) \leq F(S) - a + Nb - (2N+1)b + Nc &\leq F(S) - a + Nb - (2N+1)b + N(b+N) \\ &\leq F(S) + N^2 - a - b \leq O(N^2) - a - b. \end{aligned}$$

In total, we have $a + b \leq O(N^2)$. Therefore, the total number of iterations where $\mathrm{ESTR}(S) \in (\theta_{\ell-1}, \theta_\ell)$ is $a + b + c \leq a + 2b + N \leq O(N^2)$. □

Input: $\{\widehat{u}_{ti}, r_{ti}, x_{ti}\}_{i=1}^{N}$, multiplicative approximation parameter α, additive approximation parameter ϵ, repetition $L \in \mathbb{N}$.
Output: An approximate maximizer $\widehat{S}$ of $\mathrm{ESTR}(\widehat{S}) + \min\{1, \omega \cdot \mathrm{CI}(\widehat{S})\}$.
Generalize L vectors $y^{(1)}, \cdots, y^{(L)} \in \mathbb{R}^d$ independently and uniformly from the unit sphere;
for $\ell = 1, 2, \cdots, L$ **do**
  Replace each x_{ti} with $\langle x_{ti}, y^{(\ell)} \rangle$;
  Invoke Algorithm 3 on the reduced univariate problem instance, and let $\widehat{S}^{(\ell)}$ be the output;
end
Output $\widehat{S}^{(\ell)}$ that maximizes $\mathrm{ESTR}(\widehat{S}^{(\ell)}) + \min\{1, \alpha\omega \cdot \mathrm{CI}(\widehat{S}^{(\ell)})\}$.

Algorithm 5: Approximate combinatorial optimization, the multivariate ($d > 1$) case

Appendix D. Multivariate approximation algorithm

In this section we describe an approximation algorithm for the combinatorial optimization problem studied in Sec. 5.1 for the general multivariate ($d > 1$) case. The multivariate case is dealt with by *randomized* reductions to several univariate problem instances.

More specifically, for any $y \in \mathbb{R}^d$, $\|y\|_2 = 1$, a univariate problem instance can be constructed by replacing every occurrences of x_{ti} with $x_{ti}^\top y$. The univariate approximation Algorithm 3 is then invoked on L independent univariate problem instances, each corresponding to a y vector sampled uniformly at random from the d-dimensional unit sphere. The L output maximizers $\widehat{S}$ of Algorithm 3 are then compared against each other and the one leading to the largest value of $\mathrm{ESTR}(\widehat{S}) + \min\{1, \alpha\omega \cdot \mathrm{CI}(R)\}$ is selected, where α is the preset multiplicative approximation parameter. A pseudo-code description is given in Algorithm 5.

D.1 Approximation guarantees

The approximation performance of Algorithm 5 can be analyzed based on the following observation: if y is close to y^*, the leading eigenvector of

$$\frac{\sum_{j \in S^*} \widehat{u}_{tj} x_{tj} x_{tj}^\top}{1 + \sum_{j \in S^*} \widehat{u}_{tj}} - \left(\frac{\sum_{j \in S^*} \widehat{u}_{tj} x_{tj}}{1 + \sum_{j \in S^*} \widehat{u}_{tj}} \right) \left(\frac{\sum_{j \in S^*} \widehat{u}_{tj} x_{tj}}{1 + \sum_{j \in S^*} \widehat{u}_{tj}} \right)^\top,$$

where S^* is the exact maximizer of Eq. (19), then the reduction to a univariate problem instance $x_{tj} \mapsto x_{tj}^\top y$ does not lose much accuracy. More specifically, we have the following lemma:

Lemma 38 *Suppose there exists $\ell \in [L]$ such that $\langle y^{(\ell)}, y^* \rangle \geq 1/\alpha$ for some $\alpha \geq 1$ in Algorithm 5, then* $\mathrm{ESTR}(\widehat{S}^{(\ell)}) + \min\{1, \alpha\omega \cdot \mathrm{CI}(\widehat{S}^{(\ell)})\} + \varepsilon \geq \mathrm{ESTR}(S^*) + \min\{1, \omega \cdot \mathrm{CI}(S^*)\}$, *where $\varepsilon > 0$ is the approximation parameter of the univariate problem instances.*

Proof. For each assortment S, define $\mathrm{CI}^{(\ell)}(S)$ by

$$\mathrm{CI}^{(\ell)}(S) := {y^{(\ell)}}^\top \left(\frac{\sum_{j\in S} \widehat{u}_{tj} x_{tj} x_{tj}^\top}{1+\sum_{j\in S}\widehat{u}_{tj}} - \left(\frac{\sum_{j\in S}\widehat{u}_{tj}x_{tj}}{1+\sum_{j\in S}\widehat{u}_{tj}}\right)\left(\frac{\sum_{j\in S}\widehat{u}_{tj}x_{tj}}{1+\sum_{j\in S}\widehat{u}_{tj}}\right)^\top \right) y^{(\ell)},$$

Since $\mathrm{CI}^{(\ell)}(S) \le \mathrm{CI}(S)$, we have

$$\mathrm{ESTR}(\widehat{S}^{(\ell)}) + \min\{1, \alpha\omega\cdot \mathrm{CI}(\widehat{S}^{(\ell)})\} + \varepsilon \ge \mathrm{ESTR}(\widehat{S}^{(\ell)}) + \min\{1, \alpha\omega\cdot \mathrm{CI}^{(\ell)}(\widehat{S}^{(\ell)})\} + \varepsilon. \quad (75)$$

By the approximation guarantee of Algorithm 3, we have

$$\mathrm{ESTR}(\widehat{S}^{(\ell)}) + \min\{1, \alpha\omega\cdot \mathrm{CI}^{(\ell)}(\widehat{S}^{(\ell)})\} + \varepsilon \ge \mathrm{ESTR}(S^*) + \min\{1, \alpha\omega\cdot \mathrm{CI}^{(\ell)}(S^*)\}. \quad (76)$$

Since $\langle y^{(\ell)}, y^*\rangle \ge 1/\alpha$, we have $\mathrm{CI}^{(\ell)}(S^*) \ge (1/\alpha)\cdot \mathrm{CI}(S^*)$. Therefore,

$$\mathrm{ESTR}(S^*) + \min\{1, \alpha\omega\cdot \mathrm{CI}^{(\ell)}(S^*)\} \ge \mathrm{ESTR}(S^*) + \min\{1, \omega\cdot \mathrm{CI}(S^*)\}. \quad (77)$$

The lemma is proved by combining Eq. (75), Eq. (76), and Eq. (77). □

At a higher level, Lemma 38 shows that when the sampled vector $y^{(\ell)}$ is close to the underlying leading eigenvector y^* (in the sense that the inner product between $y^{(\ell)}$ and y^* is large), the produced subset $\widehat{S}^{(\ell)}$ will have good performance in maximizing the objective function $\mathrm{ESTR}(S) + \min\{1, \omega\cdot \mathrm{CI}(S)\}$.

The following proposition additionally gives the proximity between a random y and y^*.

Proposition 39 *Assume that $d \ge 2$. Let $y^* \in \mathbb{R}^d$, $\|y^*\|_2 = 1$ be fixed and y be sampled uniformly at random from the unit d-dimensional sphere. Then*

$$\Pr[\langle y, y^*\rangle \ge 1/\sqrt{d}] = \Omega(1) \quad \text{and} \quad \Pr[\langle y, y^*\rangle \ge 1/2] = \exp\{-O(d)\}.$$

Proof. Assume without loss of generality that $y^* = (1, 0, 0, \ldots, 0)$, and let y be sampled as follows. Sample $z_i \sim N(0,1)$ independently for each $i \in [d]$, and let $y = z/\|z\|_2$. Now, $\langle y, y^*\rangle = z_1/\|z\|_2$.

We first prove $\Pr[\langle y, y^*\rangle \ge 1/\sqrt{d}] = \Pr[z_1/\|z\|_2 \ge 1/\sqrt{d}] = \Omega(1)$. Note that when $z_1 \ge 10$ and $\sqrt{z_2^2+\cdots+z_d^2} \le 5\sqrt{d}$, we have $z_1/\|z\|_2 = 1/\sqrt{1+(z_2^2+\cdots+z_d^2)/z_1^2} \ge 1/\sqrt{1+(5\sqrt{d})^2/10^2} \ge 1/\sqrt{d}$, where the last inequality holds for $d \ge 2$. Therefore,

$$\begin{aligned}\Pr[z_1/\|z\|_2 \ge 1/\sqrt{d}] &\ge \Pr\left[z_1 \ge 10 \wedge \sqrt{z_2^2+\cdots+z_d^2} \le 5\sqrt{d}\right] \\ &= \Pr[z_1 \ge 10]\cdot \Pr\left[\sqrt{z_2^2+\cdots+z_d^2} \le 5\sqrt{d}\right] = \Omega(1).\end{aligned}$$

Now we prove $\Pr[\langle y, y^* \rangle \geq 1/2] = \Pr[z_1/\|z\|_2 \geq 1/2] = \exp\{-O(d)\}$. Similarly, when $z_1 \geq 5\sqrt{d}$ and $\sqrt{z_2^2 + \cdots + z_d^2} \leq 5\sqrt{d}$, we have $z_1/\|z\|_2 = 1/\sqrt{1 + (z_2^2 + \cdots + z_d^2)/z_1^2} \geq 1/\sqrt{1+1} > 1/2$. Therefore,

$$\begin{aligned}
\Pr[z_1/\|z\|_2 \geq 1/2] &\geq \Pr\left[z_1 \geq 5\sqrt{d} \wedge \sqrt{z_2^2 + \cdots + z_d^2} \leq 5\sqrt{d}\right] \\
&= \Pr[z_1 \geq 5\sqrt{d}] \cdot \Pr\left[\sqrt{z_2^2 + \cdots + z_d^2} \leq 5\sqrt{d}\right] \\
&= \exp\{-O(d)\} \cdot \Omega(1) = \exp\{-O(d)\}.
\end{aligned}$$

□

Combining Lemma 38 and Proposition 39 we can give some recommendations on the choice of L in Algorithm 5, which is the number of random $y^{(\ell)}$ vectors sampled. First, if $L \asymp \log(1/\delta)$ initializations are taken, then with probability $1-\delta$ Lemma 38 is satisfied with $\alpha = \sqrt{d}$, yielding a $(\sqrt{d}, \epsilon, \delta)$-approximation. Additionally, if $L \asymp e^{O(d)} \log(1/\delta)$ initializations are taken, then with probability $1-\delta$ Lemma 38 is satisfied with $\alpha = 2$, yielding a $(2, \epsilon, \delta)$-approximation.

D.2 Time complexity analysis

To achieve a $(\sqrt{d}, \varepsilon, \delta)$-approximation L is set to $L \asymp \log(1/\delta)$ and the overall running time of Algorithm 5 $O(K^9 N \max\{\varepsilon^{-4}, \omega^8 \varepsilon^{-8}\} \log \delta^{-1})$. To achieve a $(2, \varepsilon, \delta)$-approximation L is set to $L \asymp e^{O(d)} \log(1/\delta)$ and the overall running time of Algorithm 5 is $e^{O(d)} K^9 N \max\{\varepsilon^{-4}, \omega^8 \varepsilon^{-8}\}$.

Now we use Algorithm 5 to solve the combinatorial optimization problem in Step 8 of Algorithm 1 and examine the cumulative regret. If we let Algorithm 5 achieve to $(\sqrt{d}, \varepsilon, \delta)$-approximation guarantee with $\varepsilon = T^{-1/2}$ and $\delta = T^{-2}$, the computational time complexity at each time slot will be $\widetilde{O}(K^9 N d^4 T^4)$,[4] and the cumulative regret will be upper bounded by $O(\sqrt{d}) \cdot \text{Regret}^*$. If we let Algorithm 5 to achieve $(1/2, \varepsilon, \delta)$-approximation guarantee with $\varepsilon = T^{-1/2}$ and $\delta = T^{-2}$, the computational time complexity at each time slot will be $e^{O(d)} \cdot \widetilde{O}(K^9 N d^4 T^4)$, and the cumulative regret will be upper bounded by $O(1) \cdot \text{Regret}^*$.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2011.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2013.

Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. Thompson sampling for MNL-bandit. In *Proccedings of the Conference on Learning Theory (COLT)*, 2017.

4. A poly-logarithmic factor dependent on T, K, δ^{-1} is hidden in the $\widetilde{O}(\cdot)$ notation.

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

Felipe Caro and Jérémie Gallien. Dynamic Assortment with Demand Learning for Seasonal Consumer Goods. *Management Science*, 53(2):276–292, 2007.

Xi Chen and Yining Wang. A note on tight lower bound for mnl-bandit assortment selection models. *Operations Research Letters*, 46(5):534–537, 2018.

Xi Chen, Will Ma, David Simchi-Levi, and Linwei Xin. Dynamic recommendation at checkout under inventory constraint. *Available at SSRN: https://www.ssrn.com/abstract=2853093*, 2016.

Xi Chen, Akshay Krishnamurthy, and Yining Wang. Robust dynamic assortment optimization in the presence of outlier customers. *arXiv preprint arXiv:1910.04183*, 2019.

Xi Chen, Zachary Owen, Clark Pixton, and David Simchi-Levi. A statistical learning approach to personalization in revenue management. *Management Science*, 2020a.

Xi Chen, Yining Wang, and Yuan Zhou. Dynamic assortment selection under nested logit models. *Production and Operations Management*, 2020b.

Wang Chi Cheung and David Simchi-Levi. Thompson sampling for online personalized assortment optimization problems with multinomial logit choice models. *Available at SSRN: https://papers.ssrn.com/?abstract_id=3075658*, 2017.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic Linear Optimization under Bandit Feedback. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2008.

Xiequan Fan, Ion Grama, and Quansheng Liu. Exponential inequalities for martingales with applications. *Electronic Journal of Probability*, 20(1):1–22, 2015.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2010.

David A Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1): 100–118, 1975.

Negin Golrezaei, Hamid Nazerzadeh, and Paat Rusmevichientong. Real-time optimization of personalized assortments. *Management Science*, 60(6):1532–1551, 2014.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of International Conference on Machine Learning (ICML)*, 2017.

Stuart Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

David McFadden. Conditional logit analysis of qualitative choice behaviour. In P. Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press New York, New York, NY, USA, 1973.

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

Paat Rusmevichientong, Zhuojun Max Shen, and David Shmoys. Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations Research*, 58(6):1666–1680, 2010.

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

Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.

Sara A. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

Aad W Van der Vaart. *Asymptotic statistics*. Cambridge university press, 2000.

Yining Wang, Xi Chen, and Yuan Zhou. Near-optimal policies for dynamic multinomial logit assortment selection models. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2018.