Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Rate-Optimal Online Learning for Dynamic Assortment Selection with Positioning

Yiyun Luo, Will Wei Sun, Yufeng Liu

To cite this article:
Yiyun Luo, Will Wei Sun, Yufeng Liu (2026) Rate-Optimal Online Learning for Dynamic Assortment Selection with Positioning. Operations Research 74(1):224-242. https://doi.org/10.1287/opre.2024.1556

https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 74, No. 1, January–February 2026, pp. 224–242
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Crosscutting Areas

Rate-Optimal Online Learning for Dynamic Assortment Selection with Positioning

Yiyun Luo,[a] Will Wei Sun,[b] Yufeng Liu[c,*]

[a] School of Statistics and Data Science, Shanghai University of Finance and Economics, Shanghai 200433, China; [b] Daniels School of Business, Purdue University, West Lafayette, Indiana 47907; [c] Department of Statistics and Operations Research, Department of Genetics, Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599

*Corresponding author

Contact: luoyiyun@sufe.edu.cn, https://orcid.org/0009-0004-3788-1902 (YiL); sun244@purdue.edu, https://orcid.org/0000-0002-8412-6430 (WWS); yfliu@email.unc.edu, https://orcid.org/0000-0002-1686-0545 (YuL)

Received: March 4, 2024
Revised: December 12, 2024; April 21, 2025; June 12, 2025
Accepted: June 23, 2025
Published Online in Articles in Advance: August 11, 2025

Area of Review: Machine Learning and Data Science

https://doi.org/10.1287/opre.2024.1556

Abstract. In online retailing, the seller aims to offer assortment of items with maximized revenue. We introduce a new online learning problem called dynamic assortment selection with positioning (DAP) that additionally learns the optimal positioning within the assortment. Specifically, the customers make purchases based on the item attractiveness as the product of the position effect and unknown preference parameter through a multinomial logit choice model. We first demonstrate that any assortment-only algorithm that neglects position effects results in linear regrets. To address this gap, we propose the truncated linear regression upper confidence bound (TLR-UCB) policy. TLR-UCB utilizes a novel geometric linear bandit–type feedback structure for UCB construction under random and adaptive position effects. In addition, TLR-UCB conducts well-designed truncations before applying linear regression to handle conditional geometric responses. In theory, we establish a regret upper bound of $\tilde{O}(T^{1/2})$ for TLR-UCB, matching our derived $\Omega(T^{1/2})$ lower bound. Moreover, we develop an explore-in-TLR-UCB (EI-TLR) policy to tackle unknown position effects. It first conducts a joint learning procedure to estimate unknown preferences and position effects, and then implements a generalized TLR-UCB procedure driven by estimated position effects. Extensive experiments demonstrate the superior performance of TLR-UCB and EI-TLR over other benchmark policies.

Funding: This research was partially supported by the National Science Foundation [Grant NSF-SES 2217440].

Supplemental Material: All supplemental materials, including the code, data, and files required to reproduce the results, are available at https://doi.org/10.1287/opre.2024.1556.

Keywords: dynamic assortment optimization • position bias • regret analysis • upper confidence bound

1. Introduction

In both brick-and-mortar and online retailing, sellers strive to provide customers with an optimal assortment of products. Various approaches exist for modeling customer preferences and understanding how these preferences, in conjunction with the offered assortment, impact customers' choices (Rusmevichientong et al. 2010, Aouad et al. 2021). One popular approach is to use probabilistic models, where customers' purchasing probabilities are expressed as functions of the assortment of products and their preference parameters. The multinomial logit (MNL) model, an example that relies on the preference parameters, is derived from the random utility theory and has been widely adopted and studied (Agrawal et al. 2019, Wang et al. 2022, Ke et al. 2023, Shen et al. 2023). However, in many scenarios, these customer preferences are not known in advance, leading to the dynamic assortment selection problem, which falls into the category of an online learning problem (Zhalechian et al. 2022). In this problem, the seller aims to maximize revenue within a finite selling horizon through sequential assortment decisions and learning customer preferences. To achieve this, an effective dynamic assortment selection policy must strike a balance between exploring customer preferences and exploiting acquired knowledge.

Many studies have focused on designing policies for the dynamic assortment optimization problem (Caro and Gallien 2007, Rusmevichientong et al. 2010, Sauré and Zeevi 2013, Agrawal et al. 2017, 2019, Aznag et al. 2021, Chen et al. 2021a, Foussoul et al. 2023, Li et al. 2025). However, little attention has been given to the product presentation within the assortment, which is an important aspect in practice. The arrangement of products within the assortment can significantly impact customers' purchasing behaviors, making it a valuable

second-level decision-making process. Considering these second-level effects can lead to better decisions and improved revenues.

One significant second-level effect is the position effect, commonly observed in real-world applications (Craswell et al. 2008, Chen et al. 2023). For example, users often pay more attention to the initial products displayed on the Amazon website, and online advertisements placed in different positions yield varying boosting effects. The position bias has recently received attention in the recommendation and ranking literature. Cascading bandits (Kveton et al. 2015, Li et al. 2016, Cheung et al. 2019), for instance, address the online decision-making problem of recommending a list of K items that maximize user clicks. They employ a position-aware cascade model, where users examine the list from the first item to the last and choose the first appealing item. In the static offline assortment selection problem, Abeliuk et al. (2016) determines the optimal assortment and positioning assuming known preference parameters and position effects. However, in real-world applications, customer preferences are typically unknown in advance, requiring the seller to learn these parameters through sequential customer interactions. To the best of our knowledge, none of the existing literature on dynamic assortment selection has explored the position effect.

Motivated from this, we introduce a new online learning problem called dynamic assortment selection with positioning (DAP). In DAP, the seller must determine both the product assortment and their positioning at each time period. To model the customers' purchasing probabilities based on assortment and positioning decisions, we employ a debiased multinomial logit (debiased MNL) model to effectively incorporate position effects. This model assigns customers' purchasing probabilities of products proportional to their attraction, calculated as the product's preference parameter multiplied by the position's boosting effect. Consequently, different positionings of the same assortment may result in varying purchasing probabilities. Therefore, optimal decisions require careful consideration of both the assortment and the positioning to maximize revenues.

In our proposed DAP problem, the presence of position effects renders any assortment-only algorithm that neglects these effects to incur linear regret (Lemma 1). Hence, it is crucial to develop efficient policies that can effectively handle these additional position effects and yield improved regret bounds. Unlike the classic dynamic assortment selection problem (Agrawal et al. 2019), a significant challenge arises from the random and adaptive position effects induced by an adaptive policy. The upper confidence bound (UCB) construction approach employed in Agrawal et al. (2019) is not applicable when faced with such random and adaptive position effects. One potential solution to address this challenge while still utilizing the UCB algorithm proposed in Agrawal et al. (2019) is to treat the product and position pairs as new products. However, this approach necessitates the development of new static assortment algorithms that are challenging to solve exactly. Moreover, this approach fails to leverage the multiplication modeling of the overall attraction. See Remark 2 for details. Thus, there is a great need to explore novel techniques in constructing UCBs for unknown preference parameters in the presence of both random and adaptive position effects.

To address the DAP problem with known position effects, we propose a new truncated linear regression upper confidence bound (TLR-UCB) policy. TLR-UCB operates in epochs of random lengths, where an assortment and positioning is repeatedly offered until a no-purchase outcome occurs and ends the epoch. Within each epoch, we select the optimal assortment and positioning based on the upper confidence bounds constructed for the items' preference parameters, using historical purchase data. As the original techniques in Agrawal et al. (2019) no longer works, one key aspect and novelty of TLR-UCB is the construction of these upper confidence bounds. To handle random and adaptive position effects, TLR-UCB adopts an idea from linear contextual bandits (Abbasi-Yadkori et al. 2011). Specifically, for each item, the numbers of purchases and corresponding position effects across the epochs are treated as the response variables and predictor variables. Then a further challenge arises as the number of purchases in DAP follows a conditional geometric distribution rather than a sub-Gaussian one, commonly considered in bandit algorithms (Lattimore and Szepesvári 2020). Consequently, deriving concentration inequalities for the linear regression estimate and constructing the upper confidence bounds require additional techniques. To overcome this issue, we truncate the response variable before applying linear regression. This allows us to derive concentration inequalities for the truncated linear regression estimate around the preference parameter. The truncation parameter sequence is carefully designed to balance the variance and bias terms for geometric tails, enabling effective estimation and variance control.

In theory, we establish an $\tilde{O}(T^{1/2})$ regret upper bound for our proposed TLR-UCB policy (Theorem 1). We show that the position effects, numbers of purchases, and associated filtration form a novel geometric linear bandit–type feedback structure. By leveraging this structure, we derive a general concentration result for the truncated linear regression estimate (Proposition 1). This helps us establish the validity of our constructed UCBs for the preference parameters (Proposition 2). Notably, we extend the preference vector in traditional dynamic assortment selection to

a preference matrix that captures the attraction of each product placed at different positions. Our analysis reveals an interesting structural dominance property (Lemma 3), indicating that the optimal assortment and positioning with respect to an entrywise greater preference matrix yields a higher revenue. This property effectively controls revenue loss, leading to the remarkable $\tilde{O}(T^{1/2})$ regret achieved by our TLR-UCB policy, a significant improvement over the linear regrets of existing assortment-only algorithms (Lemma 1). To summarize, three major technical novelties jointly contribute to deriving the regret upper bound.

1. **UCB construction under geometric linear bandit–type feedback structure.** The adaptive positioning decisions and position effects lead to nonidentically distributed and nonindependent geometric random variables, and hence a significantly different probabilistic environment compared with Agrawal et al. (2019). Thus, we propose geometric linear bandit–type feedback structure under which concentrations are derived for truncated linear regression estimates for UCB construction.

2. **Logarithmic truncations for geometric tails.** Different from no truncation in Abbasi-Yadkori et al. (2011) for sub-Gaussian tails and polynomial truncations in Bubeck et al. (2013), Medina and Yang (2016) for heavy tails, we implement a carefully designed logarithmic truncation to optimally balance the bias and variance in our concentration derivation under geometric tails.

3. **Structural dominance for preference matrices.** With the introduction of position effects, we are faced with the preference matrix $\boldsymbol{v}^*(\boldsymbol{\theta}^*)^\top$ instead of the vector $\boldsymbol{v}^*$. Thus, we develop a new structural dominance property to bound the difference between the optimal revenues with respect to the true preference matrix $\boldsymbol{v}^*(\boldsymbol{\theta}^*)^\top$, and a valid UCB matrix that dominates $\boldsymbol{v}^*(\boldsymbol{\theta}^*)^\top$ entrywise.

Moreover, we establish an $\Omega(T^{1/2})$ regret lower bound for the DAP problem (Theorem 2), and hence the optimality of TLR-UCB in terms of regret rates up to logarithmic terms. Additionally, through extensive simulations and real data analysis, we validate the superior performance of TLR-UCB over two benchmark algorithms that also use known position effects.

Finally, we propose an explore-in-TLR-UCB (EI-TLR) policy to tackle the case of unknown position effects. The EI-TLR policy involves two stages. In the first stage, it conducts an exploration phase and uses a newly developed joint learning procedure to derive estimates for unknown preferences and position effects. In the second stage, it uses the estimated position effects to drive a generalized TLR-UCB procedure for further preference learning. The joint learning procedure relies on minimizing a newly developed loss function motivated from the distribution of the purchase data. Moreover, to compute the joint estimate for unknown preferences and position effects, we develop an alternate minimization (AM) algorithm with convergence guarantees. Importantly, we prove the existence and consistency of the joint estimate used by the EI-TLR policy. The proofs for the existence, consistency, and AM algorithm's convergence are challenging due to the nonconvexity of our adopted loss function and the noncompact minimization domain. On the other hand, the generalized TLR-UCB procedure implemented in the second stage is established as an extension of TLR-UCB. More specifically, the generalized TLR-UCB procedure takes the estimated position effects as an input to replace the role of true position effects in TLR-UCB. In addition, it incorporates extra purchase data from the exploration phase to accelerate its preference learning. Extensive numerical studies are conducted to evaluate the practical performance of EI-TLR. Specifically, we show the superiority of EI-TLR over two benchmark policies that ignore position effects, thus demonstrating the importance of taking into account position effects even if they are unknown. Furthermore, in Section 6, we illustrate through simulations the error rates of the joint estimate used in EI-TLR and a sublinear pattern of EI-TLR's cumulative regrets over horizon lengths.

1.1. Related Work

Our work is closely related to the two fields of dynamic assortment selection without positioning and static assortment selection with position bias. In the following, we provide a brief review of these fields and highlight their differences from our work.

1.1.1. Dynamic Assortment Selection Without Positioning. Online learning algorithms for dynamic assortment selection have gained significant interest in recent research. In this problem, the customers' preferences for products are initially unknown and must be learned through a balance of exploration and exploitation to maximize overall revenue. The pioneering work by Caro and Gallien (2007) addressed this problem for independent demands across products, leading to further investigations on the capacitated dynamic assortment selection problem under the MNL model. Various policies, including ETC (Rusmevichientong et al. 2010, Sauré and Zeevi 2013), UCB (Agrawal et al. 2019), and Thompson sampling (Agrawal et al. 2017), have been developed to balance exploration and exploitation in online decision making for the classic MNL model. Additionally, researchers have explored extensions to other models, such as the uncapacitated MNL model (Chen et al. 2021a), the nested logit model (Chen et al. 2021b), and the joint dynamic pricing and assortment (Miao and Chao 2021). However, none of these approaches consider position effects. In contrast, our adopted customer choice model is a debiased MNL model that incorporates position effects. It is worth noting that the exploration procedure in existing dynamic

assortment selection approaches is not applicable when facing the random and adaptive position effects present in our DAP problem, necessitating the development of new approaches.

1.1.2. Static Assortment Selection with Position Bias.

Several studies have examined the impact of position bias in static assortment selection problems (Abeliuk et al. 2016, Gallego et al. 2020, Aouad and Segev 2021, Feldman and Segev 2022). Among these works, Abeliuk et al. (2016) proposed a position-aware MNL model, whereas Gallego et al. (2020) and Aouad and Segev (2021) explored product framing and display with position bias across different pages or vertically differentiated locations. In Feldman and Segev (2022), the authors considered sequential offerings of assortments. In these static assortment selection problems, all parameters are known, and there is no random feedback from customers to learn from. Consequently, these problems primarily focus on optimization without the need for a learning procedure. In contrast, our dynamic assortment selection with positioning problem requires simultaneous learning of unknown parameters and optimization of overall revenue within a finite selling horizon.

2. Problem Formulation

In this section, we introduce the mathematical formulations of our proposed DAP problem. The problem involves selecting from a set of N products and placing them in K available positions. Each product $i \in [N] = \{1, \dots, N\}$ is associated with an unknown preference parameter v_i^*, representing its level of attraction to customers. We assume that all preference parameters fall within a known upper bound v_0, that is, $v_i^* \in [0, v_0]$, $\forall i \in [N]$. In addition, each position $j \in [K] = \{1, \dots, K\}$ is characterized by a position effect $\theta_j^* \in (0, 1]$ that quantifies how a product's attraction would scale if placed at that position. Without loss of generality, we assume $1 \geq \theta_1^* \geq \theta_2^* \geq \cdots \geq \theta_K^* > 0$ and define $\theta_{\max}^* = \max_{j\in[K]} \theta_j^*$ and $\theta_{\min}^* = \min_{j\in[K]} \theta_j^*$ as the maximum and minimum position effect.

The seller and customers interact through the following mechanism in a finite time horizon $[T] = \{1, 2, \dots, T\}$. At each time period $1 \leq t \leq T$, a new customer arrives and the seller makes a bilevel decision on both a product assortment $S_t \subseteq \{1, \dots, N\}$ with $|S_t| \leq K$, and an injective positioning function $\sigma_t : S_t \to \{1, \dots, K\}$ that assigns a position to each product in S_t. Upon viewing the displayed products and their respective positions $\{i, \sigma_t(i)\}_{i\in S_t}$, the customer is faced with the decision of choosing a product from the set S_t, or alternatively, choosing not to make a purchase. The seller, in turn, observes the customer's decision and utilizes this feedback to inform their future assortment and positioning strategies.

Now we introduce the probabilistic model of the customer purchasing decisions. Given an assortment S and a positioning function σ, the overall attraction of a product $i \in S$ is $\theta_{\sigma(i)}^* v_i^*$, that is, the multiplication of its original preference parameter v_i^* and the position effect $\theta_{\sigma(i)}^*$ of its allocated position $\sigma(i)$. Denote the purchased product as c, where the decision not to make a purchase is encoded as $c = 0$. Based on the overall attractions $\{\theta_{\sigma(i)}^* v_i^*\}_{i\in S}$, we adopt a position-aware debiased MNL model to determine the purchasing probabilities. Specifically, the probability of purchasing product $i \in \{0\} \cup [N]$ is

$$\mathbb{P}(c = i \mid S, \sigma) = \begin{cases} \dfrac{\theta_{\sigma(i)}^* v_i^*}{1 + \sum_{j\in S} \theta_{\sigma(j)}^* v_j^*} & \text{if } i \in S, \\ \dfrac{1}{1 + \sum_{j\in S} \theta_{\sigma(j)}^* v_j^*} & \text{if } i = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Under the MNL model of the classical dynamic assortment selection problem (Rusmevichientong et al. 2010, Agrawal et al. 2019), there are no position effects and the purchasing probability of product $i \in S$ given an assortment decision S is modeled as $v_i^*/(1 + \sum_{j\in S} v_j^*)$. As a result, our adopted position-aware debiased MNL model reduces to the classical one when all position effects are equal to one. The distinguishing factor in our position-aware approach is that the position effect directly influences the overall attraction, which in turn determines the customer's purchasing probabilities.

Denote r_i as the collected revenue when product $i \in [N]$ is sold. Then by the purchasing probabilities under our adopted position-aware MNL model, the expected revenue for any assortment S and positioning σ is given by $R(S, \sigma, \boldsymbol{v}^*, \boldsymbol{\theta}^*) = \sum_{i\in S} r_i \mathbb{P}(c = i \mid S, \sigma) = \sum_{i\in S} r_i \theta_{\sigma(i)}^* v_i^* / (1 + \sum_{j\in S} \theta_{\sigma(j)}^* v_j^*)$, where $\boldsymbol{v}^* = (v_1^*, \dots, v_N^*)^\top$ denotes the true preference vector and $\boldsymbol{\theta}^* = (\theta_1^*, \dots, \theta_K^*)^\top$ denotes the true position effect vector. The optimal assortment and positioning decision with respect to the true preference vector $\boldsymbol{v}^*$ and position effect vector $\boldsymbol{\theta}^*$ is denoted as $(S^*, \sigma^*) = \arg\max_{|S|\leq K, \sigma} R(S, \sigma, \boldsymbol{v}^*, \boldsymbol{\theta}^*)$. The seller's decision (S_t, σ_t) at any time period t may incur a revenue loss with respect to the optimal revenue $R(S^*, \sigma^*, \boldsymbol{v}^*, \boldsymbol{\theta}^*)$. To evaluate the performance of any DAP policy π, we define its expected cumulative regret $\mathrm{Reg}_\pi(T)$ as

$$\mathbb{E}_\pi \left(\sum_{t=1}^{T} \left(R(S^*, \sigma^*, \boldsymbol{v}^*, \boldsymbol{\theta}^*) - R(S_t, \sigma_t, \boldsymbol{v}^*, \boldsymbol{\theta}^*) \right) \right).$$

This regret criterion $\mathrm{Reg}_\pi(T)$ captures the total revenue loss in the horizon $[T]$ with respect to an optimal

policy that always offers the best decision (S^*, σ^*). Here $\mathbb{E}_\pi(\cdot)$ denotes the expectation with respect to the probability measure induced by the policy π. Our aim is to design a policy π to minimize the regret and equivalently maximize the expected revenue.

2.1. Linear Regret for Assortment-Only Algorithms

In this section, we emphasize the significance of incorporating position effects into dynamic assortment selection. Although various algorithms have been proposed for the classical dynamic assortment selection problem (Rusmevichientong et al. 2010; Sauré and Zeevi 2013; Agrawal et al. 2017, 2019), none of these algorithms considers the impact of position effects or provides positioning decisions. We introduce the following definition to encompass algorithms that account for such considerations.

Definition 1. An algorithm is assortment-only if it only offers assortment decisions.

In the DAP problem, each time period requires a bilevel decision that involves both assortment and positioning. Consequently, an algorithm focused solely on assortment decisions can only be utilized if accompanied by a positioning function. In such cases, a fair and natural approach is to employ uniformly random positioning across all possibilities. This choice is reasonable since assortment-only algorithms lack awareness of position effects, and from their perspective, all positions are considered equal. However, it has been discovered that any assortment-only algorithm, when applied to the nondegenerate DAP problem, can lead to a linear regret, as demonstrated by Lemma 1. Before proceeding, we introduce the definition of the nondegenerate DAP setting as follows.

Definition 2. A DAP problem instance $\{\boldsymbol{v}^*, \boldsymbol{\theta}^*, r\}$ is nondegenerate if the position effects $\{\theta_i^*\}_{i\in[K]}$ are not all identical.

Lemma 1. *Consider a nondegenerate DAP problem instance $\{\boldsymbol{v}^*, \boldsymbol{\theta}^*, r\}$. Assume $K \geq 3$ positions and the K lines $\{f_i(\lambda) = v_i^*(r_i - \lambda)\}_{i\in S}$ do not intersect at a single point for any assortment S with $|S| = K$. Then any assortment-only algorithm with random positioning incurs a linear regret under this problem instance.*

Remark 1. The conditions in Lemma 1 are mild. The first condition of $K \geq 3$ positions only eliminates the cases with two positions. The second assumption is a concise and reasonable condition to ensure that there is no assortment that generates identical revenue for any positioning. This condition is easily met because it is rare for three or more lines to intersect at a single point. Consequently, Lemma 1 establishes a linear regret of assortment-only algorithms for a wide range of settings.

Lemma 1 implies that the assortment-only algorithms would perform poorly when position effects exist. Therefore, it is important to develop policies that well adapt to position effects and achieve sublinear regrets.

3. Proposed Policy

In this section, we present our TLR-UCB policy designed to address the DAP problem with known position effects $\boldsymbol{\theta}^*$. To provide a clearer understanding, we begin by outlining the algorithm for TLR-UCB, which is depicted in Algorithm 1. Subsequently, we introduce the essential UCB construction procedure in Algorithm 2 as a separate component.

3.1. TLR-UCB: Algorithm Outline

Algorithm 1 illustrates the primary structure of the TLR-UCB policy. In the following paragraphs, we will provide a comprehensive explanation of each component.

Algorithm 1 (TLR-UCB)

Input: Truncation parameter sequence $\{\alpha_k = \log_{(1+1/v_0)}(k)\}_{k\in\mathbb{N}^+}$; regularization parameter $\lambda \in \mathbb{R}^+$; horizon length T

Initialization: Time period $t = 1$; Epoch $\ell = 1$; Offered times $T_i(0) = 0,\ \forall i \in [N]$

For $\ell = 1, 2, \ldots,$ **do**

  If $\ell \leq \tilde{L} \triangleq \lceil N/K \rceil$ **do** (Decision of $(\dot{S}_\ell, \dot{\sigma}_\ell)$: initialization epochs)

    Set the assortment $\dot{S}_\ell = \{([(\ell-1)K + i - 1] \bmod N) + 1 | i \in [K]\}$, and the positioning

    $\dot{\sigma}_\ell : \dot{S}_\ell \to [K]$ such that $\dot{\sigma}_\ell[([(\ell-1)K + i - 1] \bmod N) + 1] = i,\ \forall i \in [K]$.

  Else (Decision of $(\dot{S}_\ell, \dot{\sigma}_\ell)$: postinitialization epochs)

    For $i = 1, 2, \ldots, N$ **do** (UCB construction)

      Apply the UCB Construction Algorithm 2 with inputs $\{\{\tilde{v}_{i,k}\}_{k\in[T_i(\ell-1)]},$

      $\{\tilde{\theta}_{i,k}\}_{k\in[T_i(\ell-1)]}, \{\alpha_k\}_{k\in\mathbb{N}^+}, \lambda, T\}$; obtain its output $v_{i,\ell-1}^{\mathrm{UCB}}$.

    End for

    Set the assortment and positioning $(\dot{S}_\ell, \dot{\sigma}_\ell) = \arg\max_{|S|\leq K, \sigma} R(S, \sigma, \boldsymbol{v}_{\ell-1}^{\mathrm{UCB}}, \boldsymbol{\theta}^*)$,

    where $\boldsymbol{v}_{\ell-1}^{\mathrm{UCB}} = (v_{1,\ell-1}^{\mathrm{UCB}}, v_{2,\ell-1}^{\mathrm{UCB}}, \ldots, v_{N,\ell-1}^{\mathrm{UCB}})^\top$.

  End if

  Repeat (Epoch-based offering)

    Offer assortment and positioning $(\dot{S}_\ell, \dot{\sigma}_\ell)$.

17: Observe the customer purchasing decision $c_t \in \{0\} \cup \dot{S}_\ell$.
18: Update $\mathcal{E}_\ell = \mathcal{E}_\ell \cup \{t\}$, $t = t+1$.
19: **Until** $c_t = 0$
20: Compute the number of purchases of product i as $\hat{v}_{i,\ell} = \sum_{t\in\mathcal{E}_\ell} \mathbb{1}_{\{c_t=i\}}$, $\forall i \in \dot{S}_\ell$.
21: Update $T_i(\ell) = T_i(\ell-1)+1$, $\forall i \in \dot{S}_\ell$, and $T_i(\ell) = T_i(\ell-1)$, $\forall i \notin \dot{S}_\ell$.
22: Denote $\tilde{v}_{i,T_i(\ell)} = \hat{v}_{i,\ell}, \tilde{\theta}_{i,T_i(\ell)} = \theta^*_{\dot{\sigma}_\ell(i)}$, $\forall i \in \dot{S}_\ell$. (Reindexing)
23: **End for**

3.1.1. Input (Line 1). The TLR-UCB policy takes two inputs. The first input is a sequence of truncation parameters, denoted as $\{\alpha_k\}_{k\in\mathbb{N}^+}$. These truncation parameters are utilized to construct an estimate of v_i^* that acts as a baseline term in the UCB construction in Algorithm 2. Throughout the paper, we use the choice $\alpha_k = \log_{(1+1/v_0)}(k)$, which is carefully designed to ensure a valid and tight UCB, leading to an optimal $\tilde{O}(T^{1/2})$ regret for TLR-UCB. Further details regarding the truncation design can be found in Section 3.2 and after the key concentration Proposition 1. The second input is a regularization parameter $\lambda \in \mathbb{R}^+$, which is utilized in the ridge regression to estimate v_i^*.

3.1.2. Initialization (Line 2). The TLR-UCB algorithm operates in epochs, with the current time period and epoch denoted as t and ℓ respectively. Initially, we set $t = 1$ and $\ell = 1$. Furthermore, we denote $T_i(\ell)$ as the number of epochs, up to epoch ℓ, in which product i is offered. Namely, $T_i(\ell) = |\mathcal{T}_i(\ell)|$, where $\mathcal{T}_i(\ell) = \{\tau \le \ell : i \in \dot{S}_\tau\}$, and $T_i(0)$ is initialized as zero. At the end of each epoch ℓ, TLR-UCB updates the value of $T_i(\ell)$ from $T_i(\ell-1)$ based on whether product i is offered in epoch ℓ, as indicated in line 18 of the algorithm.

3.1.3. Decision of $(\dot{S}_\ell, \dot{\sigma}_\ell)$ (Lines 3–11). At the beginning of epoch ℓ, a combination of assortment and positioning $(\dot{S}_\ell, \dot{\sigma}_\ell)$ are determined. The initialization and postinitialization epochs entail two distinct decision-making approaches for assortment and positioning decisions.

- **Decision of $(\dot{S}_\ell, \dot{\sigma}_\ell)$: initialization epochs (Lines 4 and 5).** TLR-UCB includes $\tilde{L} = \lceil N/K \rceil$ initialization epochs. During these initial $\tilde{L}$ epochs, the assortment and positioning $(\dot{S}_\ell, \dot{\sigma}_\ell)$ are determined using a rolling approach, as depicted in line 5 of the algorithm. This rolling approach ensures that each product obtains purchase data during the initialization epochs, which prevents any singularities from occurring during the subsequent estimation stage.
- **Decision of $(\dot{S}_\ell, \dot{\sigma}_\ell)$: postinitialization epochs (Lines 6–10).** For postinitialization epochs $\ell \ge \tilde{L}+1$, we employ a novel UCB construction to determine $(\dot{S}_\ell, \dot{\sigma}_\ell)$. In lines 7 and 8 of the algorithm, we utilize the UCB construction from Algorithm 2 for each product i individually, resulting in upper confidence bounds $\{v^{\mathrm{UCB}}_{i,\ell-1}\}_{i\in[N]}$ for the preference parameters $\{v_i^*\}_{i\in[N]}$. Subsequently, in line 10, $(\dot{S}_\ell, \dot{\sigma}_\ell)$ is determined as $\arg\max_{|S|\le K,\sigma} R(S,\sigma,\boldsymbol{v}^{\mathrm{UCB}}_{\ell-1}, \boldsymbol{\theta}^*)$, representing the optimal assortment and positioning based on the constructed UCBs for the preference parameters. Although Algorithm 1 does not delve into the details of constructing upper confidence bounds for the preference parameters, the key UCB construction procedure is outlined in Algorithm 2, where novel concepts are introduced. It is important to note that the UCB construction method described in Agrawal et al. (2019) is no longer applicable due to the presence of position effects and adaptive positioning decisions in previous epochs. Additional details on this important step are provided in Section 3.2.

3.1.4. Epoch-Based Offering and Reindexing (Lines 12–19). After obtaining the assortment and positioning $(\dot{S}_\ell, \dot{\sigma}_\ell)$, in lines 12–16 of the algorithm, we repeatedly offer this assortment and positioning until a no-purchase outcome occurs. Subsequently, we process the observed purchasing decisions made by customers throughout the epoch. Specifically, we can obtain the number of purchases $\hat{v}_{i,\ell}$ for each product $i \in \dot{S}_\ell$, as indicated in line 17. To facilitate UCB constructions and theoretical derivations, we reindex $\hat{v}_{i,\ell}$ based on the offering times $T_i(\ell)$ of product i up to epoch ℓ. This is done in line 19, where we denote $\tilde{v}_{i,T_i(\ell)} = \hat{v}_{i,\ell}$ and the corresponding position effect $\tilde{\theta}_{i,T_i(\ell)} = \theta^*_{\dot{\sigma}_\ell(i)}$. Thus, $\{\tilde{v}_{i,k}\}_{k\in[T_i(\ell)]}$ and $\{\theta_{i,k}\}_{k\in[T_i(\ell)]}$ represent the $T_i(\ell)$ numbers of purchases and corresponding position effects of product i in the first ℓ epochs. They serve as the relevant data for product i that can be utilized in the UCB construction of its preference parameter v_i in the subsequent epoch, as illustrated in line 8.

3.2. TLR-UCB: Key Step of UCB Construction

In this section, we present the crucial step of a novel UCB construction method outlined in Algorithm 2. It comprises two major steps: a truncation step and a UCB construction step.

Firstly, we establish that the number of purchases $\hat{v}_{i,\ell}$ in line 17 of Algorithm 1 follows a conditional geometric distribution. This observation serves as a motivation for the truncation step, which is conducted to facilitate the subsequent UCB construction.

Lemma 2. *For $i \in \dot{S}_\ell$, $\hat{v}_{i,\ell}$ follows a geometric distribution of* $\mathrm{Geo}(1/(1+\theta^*_{\dot{\sigma}_\ell(i)}v_i^*))$ *conditional on $(\dot{S}_\ell, \dot{\sigma}_\ell)$. Namely,* $\mathbb{P}(\hat{v}_{i,\ell} = k | \dot{S}_\ell, \dot{\sigma}_\ell) = (1 - 1/(1+\theta^*_{\dot{\sigma}_\ell(i)}v_i^*))^k \cdot (1/(1+\theta^*_{\dot{\sigma}_\ell(i)}v_i^*))$ *for any* $k \in \mathbb{N}^+ \cup \{0\}$.

The observation in Lemma 2 aligns with our intuition, as a higher overall attraction $\theta^*_{\sigma_\ell(i)}v^*_i$ corresponds to a larger expected number of purchases $\hat{v}_{i,\ell}$. Next we formalize the probabilistic environment. Let $\mathcal{F}_\ell = \sigma(\dot{S}_1, \dot{\sigma}_1, \{\hat{v}_{i,1}\}_{i\in\dot{S}_1}, \ldots, \dot{S}_\ell, \dot{\sigma}_\ell, \{\hat{v}_{i,\ell}\}_{i\in\dot{S}_\ell}, \dot{S}_{\ell+1}, \dot{\sigma}_{\ell+1})$ be the sigma-field that encompasses the assortment and positioning decisions up to the $(\ell+1)$th epoch, as well as the summarized purchase data $\{\hat{v}_{i,\ell}\}_{i\in\dot{S}_\ell}$ up to the ℓth epoch. Based on Lemma 2, we find that $\hat{v}_{i,\ell}$ follows a conditional geometric distribution of $\mathrm{Geo}(1/(1+\theta^*_{\sigma_\ell(i)}v^*_i))$ given $\mathcal{F}_{\ell-1}$. Let epoch $\tau_{i,k}$ denote the kth epoch in which product i is offered, that is, $\tau_{i,k} = \min\{\ell : T_i(\ell) = k\}$. We define the product-specific filtration $\{\mathcal{G}_{i,k-1}\}_{k\geq 1}$ by $\mathcal{G}_{i,k-1} = \mathcal{F}_{\tau_{i,k}-1}$. It is important to note that the event $\{\tau_{i,k} - 1 = \ell\}$, indicating that epoch $\ell+1$ is the kth epoch in which product i is offered, is fully determined by $\mathcal{F}_\ell$. Hence, $\tau_{i,k} - 1$ is a stopping time and we validate that $\{\mathcal{G}_{i,k-1}\}_{k\geq 1}$ form a filtration. According to the definition of the reindexed numbers of purchases $\{\tilde{v}_{i,k}\}_{k\in[T_i(\ell)]}$ and corresponding position effects $\{\tilde{\theta}_{i,k}\}_{k\in[T_i(\ell)]}$ for the product i, as indicated in line 19 of Algorithm 1, we obtain that $\tilde{v}_{i,k} = \hat{v}_{i,\tau_{i,k}} \sim \mathrm{Geo}(1/(1+\theta^*_{\dot{\sigma}_{\tau_{i,k}}(i)}v^*_i)) = \mathrm{Geo}(1/(1+\tilde{\theta}_{i,k}v^*_i))$ conditional on $\mathcal{G}_{i,k-1} = \mathcal{F}_{\tau_{i,k}-1}$, and additionally $\tilde{\theta}_{i,k} = \theta^*_{\dot{\sigma}_{\tau_{i,k}}(i)} \in \mathcal{F}_{\tau_{i,k}-1} = \mathcal{G}_{i,k-1}$ and $\tilde{v}_{i,k} = \hat{v}_{i,\tau_{i,k}} \in \mathcal{F}_{\tau_{i,k}} \subseteq \mathcal{F}_{\tau_{i,k+1}-1} = \mathcal{G}_{i,k}$.

Algorithm 2 (UCB Construction for TLR-UCB)

1: **Input:** $\{\tilde{v}_{i,k}\}_{k\in[T_i(\ell-1)]}, \{\tilde{\theta}_{i,k}\}_{k\in[T_i(\ell-1)]}$, that is, numbers of purchases and corresponding position effects of product i in the first $\ell-1$ epochs; truncation parameter sequence $\{\alpha_k = \log_{(1+1/v_0)}(k)\}_{k\in\mathbb{N}^+}$; regularization parameter $\lambda \in \mathbb{R}^+$; horizon length T

2: Truncate the number of purchases $\{\tilde{v}_{i,k}\}_{k\in[T_i(\ell-1)]}$ toward $\{\alpha_k\}_{k\in[T_i(\ell-1)]}$, that is,

$$\dot{v}_{i,k} = \tilde{v}_{i,k}\mathbb{1}_{\{|\tilde{v}_{i,k}|\leq\alpha_k\}} + \alpha_k\mathbb{1}_{\{|\tilde{v}_{i,k}|>\alpha_k\}}.$$

3: Compute an estimate of v^*_i as $\overline{v}_{i,T_i(\ell-1)} = (\sum_{k=1}^{T_i(\ell-1)} \tilde{\theta}_{i,k}\dot{v}_{i,k})/(\lambda + \sum_{j=1}^{T_i(\ell-1)} \tilde{\theta}^2_{i,k})$.

4: Construct the upper confidence bound of v^*_i as

$$v^{\mathrm{UCB}}_{i,\ell-1} = \bar{v}_{i,T_i(\ell-1)} + \frac{\frac{2\log T}{\log\left(1+\frac{1}{v_0}\right)}\sqrt{2\log(NT)+\log\left(1+\frac{T_i(\ell-1)(\theta^*_{\max})^2}{\lambda}\right)} + \frac{(v_0+1)\pi}{\sqrt{6}} + \sqrt{\lambda}v_0}{\theta^*_{\min}\sqrt{T_i(\ell-1)}}.$$

5: **Output:** $v^{\mathrm{UCB}}_{i,\ell-1}$, that is, constructed upper confidence bound of v^*_i

Now we are ready to introduce Algorithm 2 with the inputs inherited from line 8 of Algorithm 1. The goal of Algorithm 2 is to construct the UCB for v^*_i. The algorithm proceeds by first building an estimate for v^*_i using the provided inputs, and then establishing concentration bounds for this estimate around v^*_i. A key observation is that $\tilde{v}_{i,k}$ follows a conditional geometric distribution of $\mathrm{Geo}(1/(1+\tilde{\theta}_{i,k}v^*_i))$ conditional on $\mathcal{G}_{i,k-1}$, and $\mathbb{E}(\tilde{v}_{i,k}|\mathcal{G}_{i,k-1}) = \tilde{\theta}_{i,k}v^*_i$. Therefore, a standard estimate for v^*_i can be obtained by performing linear regression with the responses $\{\tilde{v}_{i,k}\}_{k\leq T_i(\ell)}$ and predictors $\{\tilde{\theta}_{i,k}\}_{k\leq T_i(\ell)}$. However, because $\{\tilde{v}_{i,k}\}_{k\leq T_i(\ell)}$ follow a conditional geometric distribution rather than a conditional sub-Gaussian distribution, deriving concentration bounds for this standard estimate becomes challenging, which in turn limits a provable UCB construction for v^*_i. To overcome the issue of the geometric tail, we employ the concept of truncation before applying linear regression. Specifically, we define the truncation parameter sequence as $\{\alpha_k = \log_{(1+1/v_0)}(k)\}_{k\in\mathbb{N}^+}$ and truncate the nonnegative number of purchases $\hat{v}_{i,k}$ toward α_k, as shown in line 2 of Algorithm 2.

Using these truncated responses, we can formulate a truncated linear regression problem to obtain the estimate $\bar{v}_{i,T_i(\ell-1)}$ in line 3. Unlike the standard estimate, we can derive concentration bounds for this truncated linear regression estimate around v^*_i. Our approach involves first establishing concentration bounds for this estimate around its expectation (variance term) and then controlling the difference between its expectation and v^*_i (bias term). It is worth noting that the choice of the truncation parameter sequence $\{\alpha_k = \log_{(1+1/v_0)}(k)\}_{k\in\mathbb{N}^+}$ is carefully designed to balance the variance and bias terms for geometric tails. More detailed explanations regarding the truncation parameters and the rigorous implementation of this approach are provided in the discussion of Proposition 1 in Section 4.1.1. Finally, based on the derived concentration bounds, we can add a bonus term to the truncated linear regression estimate, resulting in the construction of the UCB for v^*_i as stated in line 4 of Algorithm 2. In Section 4.1.1, we will establish the validity of this constructed UCB and demonstrate that it leads to an optimal regret rate for TLR-UCB.

Remark 2 (Difference from the UCB Construction in Agrawal et al. (2019)). The TLR-UCB policy we propose takes a fundamentally different approach in constructing the UCBs for preference parameters compared with the UCB policy introduced in Agrawal et al. (2019), which we refer to as A-UCB hereafter. The distinctions arise due to the positioning decision and position effects introduced in the DAP problem. Specifically, in DAP, the conditional distributions of numbers of purchases for each product varies with its positions. Hence, the UCB construction in A-UCB, which uses the arithmetic average of numbers of purchases and leverages concentrations for independent and identically distributed geometric random variables, no longer works. Moreover, for an adaptive UCB-based algorithm, the positioning decision σ_ℓ at epoch ℓ depends on the random historical purchase

data in the past epochs. Thus σ_ℓ is indeed random and adaptive, resulting in random and adaptive position effects $\theta^*_{\sigma_\ell(i)}$ for $i \in S_\ell$. Therefore, a direct extension of A-UCB that uses the average of numbers of purchases scaled by the position effects, that is, $1/T_i(\ell) \times \sum_{\tau\in\mathcal{T}_i(\ell)} \hat{v}_{i,\tau}/\theta^*_{\sigma_\tau(i)}$, would introduce random and dependent denominators. In contrast, in the original dynamic assortment problem (Agrawal et al. 2019), all position effects degenerate to one and hence the estimate degenerates to $1/T_i(\ell) \times \sum_{\tau\in\mathcal{T}_i(\ell)} \hat{v}_{i,\tau}/1$. Such an intrinsic difference greatly increases the difficulty of concentration derivation for the extension of A-UCB. Therefore, to tackle random and adaptive position effects, we propose the geometric linear bandit–type feedback structure. Then we derive concentrations for truncated linear regression estimates under this structure in Proposition 1, which further enables the UCB construction in Proposition 2. In addition, we apply the truncating techniques to handle non–sub-Gaussian geometric tails. This also differs from the approach in A-UCB, which uses moment generating functions and conduct Chernoff bounding.

4. Theory

In this section, we present a comprehensive regret analysis for our proposed TLR-UCB policy. Firstly, we establish an upper bound on the regret of TLR-UCB, showing that it achieves an $\tilde{O}(T^{1/2})$ regret. This result demonstrates the effectiveness of TLR-UCB in policy learning for the DAP problem. Additionally, we derive a lower bound of $\Omega(T^{1/2})$ on the regret for the DAP problem. By combining these results, we establish the optimality of TLR-UCB in term of the regret bound as its regret matches the lower bound up to a logarithm term.

4.1. Regret Upper Bound

Before presenting the regret upper bound, we first state the required assumptions.

Assumption 1. *The revenue $r_i \in [0,1]$ for any product $i \in [N]$.*

Assumption 2. *There exists a known $v_0 > 0$ such that the preference parameter is $v_i^* \in (0, v_0]$, $\forall i \in [N]$.*

Assumption 3. *The position effect $\theta_j^* \in (0,1]$ for any position $j \in [K]$.*

Assumption 1 assumes a finite bound 1 on the revenues, which is a common assumption in the dynamic assortment literature (Agrawal et al. 2017, 2019; Chen et al. 2021b). Assumption 2 relaxes the common requirement, as seen in Agrawal et al. (2017, 2019), Aznag et al. (2021), and Foussoul et al. (2023), that the preference parameters are less than or equal to one. Instead, we assume a known upper bound for all preference parameters and use it as an input for our UCB construction subroutine (Algorithm 2) and hence TLR-UCB policy (Algorithm 1). Lastly, Assumption 3 assumes that position effects are bounded by one. Note that for both Assumptions 1 and 3, the upper bounds do not need to be exactly one, and boundedness would be enough. More detailed discussions on effects of v_0 in Assumption 2, and the boundedness of $\{v_i\}_{i\in[N]}$ and $\{\theta_k^*\}_{k\in[K]}$ in Assumptions 1 and 3 are presented in Section D of the Online Appendix. With these assumptions in place, we can now state our main result, Theorem 1, on regret upper bound for TLR-UCB.

Theorem 1. *Suppose Assumptions 1–3 hold and $NT \geq 2$, the regret of TLR-UCB satisfies* $\mathrm{Reg}_\pi(T) \leq \frac{4\theta^*_{\max}}{\theta^*_{\min}}\left(\frac{6\log^2((N+\frac{1}{\lambda})T)}{\log(1+\frac{1}{v_0})} + \frac{(v_0+1)\pi}{\sqrt{6}} + \sqrt{\lambda}v_0\right)\sqrt{KNT} + (v_0+1)(N+2K)$.

Theorem 1 proves that our proposed TLR-UCB policy achieves an $\tilde{O}(T^{1/2})$ regret on the DAP problem. This represents a significant improvement over the linear regret $O(T)$ of assortment-only algorithms. Importantly, the theorem takes into account the influence of position effects $\{\theta_i^*\}_{i\in[K]}$ through the two margins $\theta^*_{\max}$ and $\theta^*_{\min}$. Specifically, a larger value of $\theta^*_{\max}/\theta^*_{\min}$ leads to a larger upper bound on the regret. This observation is intuitive because in this case the position effects play a more significant role in determining the overall preference. This difference in the regret bound highlights a clear distinction between our approach, which considers position effects, and existing dynamic assortment selection methods that ignore position effects. It is worth noting that the regret may depend not only on $\{\theta^*_{\max}, \theta^*_{\min}\}$ but on all position effects $\{\theta_i^*\}_{i\in[K]}$. The precise dependency structure of the regret on the position effects can be intricate and is a subject for future research.

Because of the distinct problem formulations and UCB constructions, the proof of Theorem 1 differs significantly from existing dynamic assortment selection approaches (Agrawal et al. 2017, 2019; Chen et al. 2021b). It relies on two key properties: valid UCB construction (Section 4.1.1) and structural dominance (Section 4.1.2). The full details of the proof are provided in Online Appendix A.2, with the outline presented first.

- **Valid UCB construction:** With high probability, the constructed $\{v_{i,\ell}^{\mathrm{UCB}}\}_{i\in[N]}$ serve as valid upper confidence bounds for $\{v_i^*\}_{i\in[N]}$.
- **Structural dominance:** Consider two pairs of preferences and position effects $(\boldsymbol{v}_1, \boldsymbol{\theta}_1)$ and $(\boldsymbol{v}_2, \boldsymbol{\theta}_2)$. Let $(S_1, \sigma_1) = \arg\max_{|S|\leq K,\sigma} R(S, \sigma, \boldsymbol{v}_1, \boldsymbol{\theta}_1)$ and $(S_2, \sigma_2) = \arg\max_{|S|\leq K,\sigma} R(S, \sigma, \boldsymbol{v}_2, \boldsymbol{\theta}_2)$ be optimal assortment and positioning decisions with respect to them. If the preference matrix $\boldsymbol{v}_1\boldsymbol{\theta}_1^\top$ is entrywise greater than or equal to, that is, dominates, $\boldsymbol{v}_2\boldsymbol{\theta}_2^\top$, then $R(S_1, \sigma_1, \boldsymbol{v}_1, \boldsymbol{\theta}_1) \geq R(S_2, \sigma_2, \boldsymbol{v}_2, \boldsymbol{\theta}_2)$.

To establish the validity of our TLR-UCB construction (Proposition 2), we first prove a newly derived concentration result for the geometric linear bandit–type

feedback structure (Proposition 1), which is also of independent interest. Furthermore, we extend the preference vector in traditional dynamic assortment selection to a two-dimensional preference matrix that captures the attraction of each product placed at different positions, and then prove a new structural dominance lemma (Lemma 3), which indicates that selecting the optimal assortment and positioning decision with respect to larger preference matrix elements leads to higher revenue. By utilizing this property, the optimal revenue with respect to valid UCBs can be utilized as an appropriate intermediate term to control the regret. A detailed proof of the theorem is presented in Section A of the Online Appendix.

4.1.1. Valid UCB Construction. In this part, we first derive a valid confidence region for a general scenario. Then we apply the result on UCB construction in our proposed TLR-UCB policy.

In our UCB construction for the preference parameter v_i^* of product i, an adaptive sequence $\{\tilde{v}_{i,k}\}_{k\in\mathbb{N}^+}$ is utilized. In addition, there exists an associated filtration $\{\mathcal{G}_{i,k-1}\}_{k\in\mathbb{N}^+}$ such that $\mathbb{E}(\tilde{v}_{i,k}|\mathcal{G}_{i,k-1}) = \tilde{\theta}_{i,k}v_i^*$, $\tilde{\theta}_{i,k} \in \mathcal{G}_{i,k-1}$, and $\tilde{v}_{i,k} \in \mathcal{G}_{i,k}$. Given this dependency structure, one key observation arises that $\{\tilde{v}_{i,k}\}_{k\in\mathbb{N}^+}$ can be viewed as linear bandit–type feedbacks with unknown one-dimensional linear parameter v_i^* and action vectors $\{\tilde{\theta}_{i,k}\}_{k\in\mathbb{N}^+}$. In fact, $\{\tilde{\theta}_{i,k}\}_{k\in\mathbb{N}^+}, \{\tilde{v}_{i,k}\}_{k\in\mathbb{N}^+}, \{\mathcal{G}_{i,k}\}_{k\in\mathbb{N}}$ fit into a geometric linear bandit–type feedback structure with a formal definition in the following.

Definition 3 (Geometric Linear Bandit–Type Feedback Structure). An $\mathbb{R}^d$-valued stochastic process $\{\boldsymbol{x}_t\}_{t\in\mathbb{N}^+}$, a real-valued stochastic process $\{y_t\}_{t\in\mathbb{N}^+}$, a filtration $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$, and a vector $\boldsymbol{\mu} \in \mathbb{R}^d$ form a geometric linear bandit–type feedback structure if $y_t|\mathcal{F}_{t-1} \sim \text{Geo}(1/(1+\boldsymbol{x}_t^\top\boldsymbol{\mu}))$, $\boldsymbol{x}_t \in \mathcal{F}_{t-1}$ and $y_t \in \mathcal{F}_t$ for any $t \in \mathbb{N}^+$.

A significant distinction between this general feedback structure and the classic linear bandit structure is the presence of conditional geometric feedback instead of conditional sub-Gaussian feedback. Because of the absence of sub-Gaussian properties, the conventional method of constructing confidence regions for linear bandit feedback (Abbasi-Yadkori et al. 2011) cannot be applied. This method relies on deriving and utilizing the concentration of regression estimates around the unknown linear parameter. However, in a geometric linear bandit–type feedback structure, the geometric tails make it challenging to establish concentration for the standard regression estimate. To address this challenge, we employ the truncation approach used in heavy-tailed linear bandits (Medina and Yang 2016). By utilizing truncated linear responses in regression estimates, we can derive concentrations around their conditional expectations. Furthermore, we are able to control the biases introduced by truncation, that is, the discrepancies between the conditional expectations and the unknown linear parameter. In light of the geometric linear bandit–type feedback structure, we develop Proposition 1 to construct confidence regions that account for these properties.

Proposition 1. *Suppose* $\{\boldsymbol{x}_t\}_{t\in\mathbb{N}^+}, \{y_t\}_{t\in\mathbb{N}^+}, \{\mathcal{F}_t\}_{t\in\mathbb{N}}, \boldsymbol{\mu}$ *form a geometric linear bandit–type feedback structure with dimension* d*, that is,* $\boldsymbol{\mu}, \boldsymbol{x}_t \in \mathbb{R}^d$*, and there exists a constant* $\nu > 0$ *such that* $\boldsymbol{x}_t^\top\boldsymbol{\mu} \in [0,\nu]$, $\forall t \in \mathbb{N}^+$*. For any* $t \in \mathbb{N}^+$*, let* $\alpha_t = \log_{1+1/\nu}(t) = \log t/\log(1+1/\nu)$ *and* $\hat{y}_t = \min\{y_t, \alpha_t\}$ *be a truncation of* y_t*. Denote* $\boldsymbol{X}_t = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t]$, $\hat{\boldsymbol{Y}}_t = (\hat{y}_1, \ldots, \hat{y}_t)^\top$*, and* $\boldsymbol{V}_t = \lambda\boldsymbol{I} + \sum_{s=1}^t \boldsymbol{x}_s\boldsymbol{x}_s^\top = \lambda\boldsymbol{I} + \boldsymbol{X}_t\boldsymbol{X}_t^\top$*, where* $\lambda > 0$ *is a constant. Let* $\hat{\boldsymbol{\mu}}_t = \boldsymbol{V}_t^{-1}\boldsymbol{X}_t\hat{\boldsymbol{Y}}_t$ *be a ridge regression estimate of* $\boldsymbol{\mu}$ *by using the truncated values* $\{\hat{y}_s\}_{s\in[t]}$*. For any* $\delta > 0$ *and* $T \in \mathbb{N}^+$*, we have* $\mathbb{P}(\|\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}\|_{\boldsymbol{V}_t} \le 2\log T/\log(1+1/\nu) \times \sqrt{2\log(1/\delta) + \log(\det(\boldsymbol{V}_t)/\lambda^d)} + (\nu+1)\pi/\sqrt{6} + \lambda^{1/2}\|\boldsymbol{\mu}\|_2, \forall t \in [T]) \ge 1-\delta$*, where* $\|\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}\|_{\boldsymbol{V}_t} = \sqrt{(\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu})^\top\boldsymbol{V}_t(\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu})}$.

In Proposition 1, we utilize truncated linear responses to construct an estimate $\hat{\boldsymbol{\mu}}_t$ for the unknown linear parameter $\boldsymbol{\mu}$. The difference $\|\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}\|_{\boldsymbol{V}_t}$ can be decomposed into a variance term and a bias term, as formalized in the proof of Proposition 1. The variance term captures the difference between $\hat{\boldsymbol{\mu}}$ and its conditional expectation, while the bias term represents the discrepancy between this conditional expectation and $\boldsymbol{\mu}$. By employing truncation, we eliminate the geometric tail from the variance term, allowing us to bound it with high probability using theorem 1 in Abbasi-Yadkori et al. (2011), which is specifically designed for sub-Gaussian feedback. On the other hand, the bias term arises from the truncation process and can be bounded by analyzing the truncated geometric tails. Achieving a tight confidence region requires finding a balance between these two terms to minimize the overall upper bound rate. This necessitates a careful design of the truncation parameter sequence. For our case with geometric tails, we employ a novel logarithmic truncation parameter sequence to strike an optimal balance. It is important to note that in heavy-tailed linear bandits (Medina and Yang 2016), the truncation parameter sequence needs to be polynomial, and of a higher order, to further reduce the bias term. This is because of the heavier tail distribution considered in those scenarios, which differs from the geometric tails in our context.

In our proposed TLR-UCB policy, the upper confidence bounds for the preference parameters are constructed based on the truncated linear regression estimates $\overline{v}_{i,T_i(\ell-1)} = (\sum_{k=1}^{T_i(\ell-1)} \tilde{\theta}_{i,k}\hat{v}_{i,k})/(\lambda + \sum_{j=1}^{T_i(\ell-1)} \tilde{\theta}_{i,k}^2)$ in line 3 of Algorithm 2. By observing that $\{\tilde{\theta}_{i,k}\}_{k\in\mathbb{N}^+}$, $\{\tilde{v}_{i,k}\}_{k\in\mathbb{N}^+}$ and the associated filtration $\{\mathcal{G}_{i,k}\}_{k\in\mathbb{N}^+}$ form the

geometric linear bandit–type feedback structure, we develop Proposition 1 for concentration results on generic truncated linear regression estimates under a geometric linear bandit–type feedback structure. Therefore, we can apply Proposition 1 to derive the concentration for the specific truncated linear regression estimate $\bar{v}_{i,T_i(\ell-1)}$. This further leads to Proposition 2, which demonstrates that our UCB construction in Algorithm 2 is valid. Denote $L = \min_\ell\{\sum_{\tau=1}^{\ell} |\mathcal{E}_\tau| \geq T\}$ as the number of epochs in the finite horizon $[T]$.

Proposition 2. *Under the TLR-UCB policy, with probability at least* $1-\frac{1}{T}$, *for any epoch* ℓ *that satisfies* $\tilde{L} \leq \ell \leq L-1$ *and* $i \in [N]$, $v_i^* \leq v_{i,\ell}^{\mathrm{UCB}} \leq v_i^* + 2\cdot(2\log T/\log(1+1/\nu_0)\sqrt{2\log(NT)+\log(1+T_i(\ell)(\theta^*_{\max})^2/\lambda)} + (\nu_0+1)\pi/\sqrt{6}+\sqrt{\lambda}\nu_0)/(\theta^*_{\min}\sqrt{T_i(\ell)})$.

4.1.2. Structural Dominance. In TLR-UCB, $(\dot{S}_\ell, \dot{\sigma}_\ell) = \arg\max_{|S|\leq K,\sigma} R(S,\sigma,\boldsymbol{v}_{\ell-1}^{\mathrm{UCB}},\boldsymbol{\theta}^*)$ is repeatedly offered in epoch ℓ. The difference $R(S^*,\sigma^*,\boldsymbol{v}^*,\boldsymbol{\theta}^*) - R(\dot{S}_\ell,\dot{\sigma}_\ell,\boldsymbol{v}^*,\boldsymbol{\theta}^*)$ plays a key role in the regret analysis. To utilize the fact that $(\dot{S}_\ell,\dot{\sigma}_\ell)$ is optimal with respect to $(\boldsymbol{v}_{\ell-1}^{\mathrm{UCB}},\boldsymbol{\theta}^*)$, we consider an intermediate term $R(\dot{S}_\ell,\dot{\sigma}_\ell,\boldsymbol{v}_{\ell-1}^{\mathrm{UCB}},\boldsymbol{\theta}^*)$. This term turns out to dominate $R(S^*,\sigma^*,\boldsymbol{v}^*,\boldsymbol{\theta}^*)$ when $\boldsymbol{v}_{\ell-1}^{\mathrm{UCB}}$ truly upper bounds $\boldsymbol{v}^*$. Specifically, this claim is implied by the structural dominance Lemma 3 that leverages the entrywise dominance relationship between $\boldsymbol{v}_{\ell-1}^{\mathrm{UCB}}(\boldsymbol{\theta}^*)^\top$ and $\boldsymbol{v}^*(\boldsymbol{\theta}^*)^\top$. With the introduction of position effects in the DAP problem, the notion of preference vector $\boldsymbol{v}^*$ in typical dynamic assortment selection problems is extended to an overall preference matrix $\boldsymbol{v}^*(\boldsymbol{\theta}^*)^\top$. The (i,k)th element of this overall preference matrix represents the attraction of a product i when placed on position j. It captures the overall preferences of customers for different combinations of products and positions. In fact, we developed a more general result in Lemma 3 that goes beyond the scope of rank 1 preference matrices like $\boldsymbol{v}^*(\boldsymbol{\theta}^*)^\top$. This lemma provides insights into the relationship between overall preference matrices and revenue optimization, allowing for a broader understanding of how different elements of the preference matrix impact the revenue.

Lemma 3 (Structural Dominance Lemma). *Let* $\boldsymbol{M}^*, \overline{\boldsymbol{M}} \in \mathbb{R}^{N\times K}$ *be* $N\times K$ *matrices with nonnegative entries. Denote* $R_0((S,\sigma),\boldsymbol{M}) = \sum_{i\in S} r_i M_{i,\sigma(i)}/(1+\sum_{j\in S} M_{j,\sigma(j)})$. *Further denote* $(S^*,\sigma^*) = \arg\max_{|S|\leq K,\sigma} R_0((S,\sigma),\boldsymbol{M}^*)$ *and* $(\overline{S},\overline{\sigma}) = \arg\max_{|S|\leq K,\sigma} R_0((S,\sigma),\overline{\boldsymbol{M}})$. *Suppose* $\overline{M}_{ij} \geq M^*_{ij} > 0$ *for any* $(i,j)\in[N]\times[K]$, *then we have* $R_0((\overline{S},\overline{\sigma}),\overline{\boldsymbol{M}}) \geq R_0((S^*,\sigma^*),\boldsymbol{M}^*)$.

Lemma 3 suggests that if $\overline{\boldsymbol{M}}$ upper bounds $\boldsymbol{M}^*$ in all entries, the optimal revenue for $\overline{\boldsymbol{M}}$, given by $R_0((\overline{S},\overline{\sigma}),\overline{\boldsymbol{M}})$, also upper bounds the optimal revenue $R_0((S^*,\sigma^*),\boldsymbol{M}^*)$ for $\boldsymbol{M}^*$. By setting $\overline{\boldsymbol{M}} = \boldsymbol{v}_1\boldsymbol{\theta}_1^\top$ and $\boldsymbol{M}^* = \boldsymbol{v}_2\boldsymbol{\theta}_2^\top$, we obtain our previously claimed structural dominance property that if $\boldsymbol{v}_1\boldsymbol{\theta}_1^\top$ dominates $\boldsymbol{v}_2\boldsymbol{\theta}_2^\top$, then $R(S_1,\sigma_1,\boldsymbol{v}_1,\boldsymbol{\theta}_1) = R_0((S_1,\sigma_1),\boldsymbol{v}_1\boldsymbol{\theta}_1^\top) \geq R_0((S_2,\sigma_2),\boldsymbol{v}_2\boldsymbol{\theta}_2^\top) = R(S_2,\sigma_2,\boldsymbol{v}_2,\boldsymbol{\theta}_2)$. Here $(S_1,\sigma_1) = \arg\max_{|S|\leq K,\sigma} R(S,\sigma,\boldsymbol{v}_1,\boldsymbol{\theta}_1) = \arg\max_{|S|\leq K,\sigma} R_0((S,\sigma),\boldsymbol{v}_1\boldsymbol{\theta}_1^\top)$ and $(S_2,\sigma_2) = \arg\max_{|S|\leq K,\sigma} R(S,\sigma,\boldsymbol{v}_2,\boldsymbol{\theta}_2) = \arg\max_{|S|\leq K,\sigma} R_0((S,\sigma),\boldsymbol{v}_2\boldsymbol{\theta}_2^\top)$.

By the structural dominance property, we obtain $R(\dot{S}_\ell,\dot{\sigma}_\ell,\boldsymbol{v}_{\ell-1}^{\mathrm{UCB}},\boldsymbol{\theta}^*) \geq R(S^*,\sigma^*,\boldsymbol{v}^*,\boldsymbol{\theta}^*)$ when $\boldsymbol{v}_{\ell-1}^{\mathrm{UCB}}$ upper bounds $\boldsymbol{v}^*$. This enables us to bound the difference $R(S^*,\sigma^*,\boldsymbol{v}^*,\boldsymbol{\theta}^*) - R(\dot{S}_\ell,\dot{\sigma}_\ell,\boldsymbol{v}^*,\boldsymbol{\theta}^*)$ by $R(\dot{S}_\ell,\dot{\sigma}_\ell,\boldsymbol{v}_{\ell-1}^{\mathrm{UCB}},\boldsymbol{\theta}^*) - R(\dot{S}_\ell,\dot{\sigma}_\ell,\boldsymbol{v}^*,\boldsymbol{\theta}^*)$.

4.2. Regret Lower Bound

Theorem 2 establishes the regret lower bound for the DAP problem.

Theorem 2. *Consider the DAP problem with* $K \leq N/4$ *and* $\nu_0 \geq 1$. *Let* $\tilde{V} = \{\boldsymbol{v}^* \in \mathbb{R}^N : v_i^* \in (0,\nu_0]\}$, $\tilde{R} = \{\boldsymbol{r}\in\mathbb{R}^N : r_i \in (0,1]\}$. *There exists a universal constant* C *such that*

$$\inf_{\pi} \sup_{\boldsymbol{v}^*\in\tilde{V},\boldsymbol{r}\in\tilde{R}} \mathrm{Reg}_\pi(T) \geq C\theta^*_{\min}\min\{\sqrt{NT},T\}.$$

The lower bound in Theorem 2 contains two components. First, when the time horizon is relatively short, such that $T \leq N$, a simple policy that selects a fixed assortment without considering any information can achieve a regret of $\Theta(T)$ and matches the optimal rate implied by Theorem 2. In the more common and interesting scenario where $N \leq T$, the lower bound is $\Omega(\sqrt{NT})$. This lower bound demonstrates the inherent difficulty of the DAP problem and signifies that achieving a regret lower than $\sqrt{NT}$ is not feasible. Note that the regret upper bound of our TLR-UCB policy is $\tilde{O}(\sqrt{NKT})$, as demonstrated in Theorem 1. This confirms the optimality of our TLR-UCB policy in terms of the horizon length T and number of products N. However, there is a gap of $\sqrt{K}$ between our proved regret upper and lower bounds. In practice, the gap of $\sqrt{K}$ is quite minor because the number of positions, especially those that could catch customer's attention, is far less than the number of products N and the horizon length T. In addition, the number of positions K is usually fixed, not like N and T, which can increase naturally as new products arrive and new sales reason begins.

To prove the lower bound in Theorem 2, we construct a set of problem instances that render no policy capable of performing well on all of them. Compared with the lower bound proof in Chen and Wang (2018) for the dynamic assortment selection problem without position effects, our proof of Theorem 2 is similar in several high-level ideas and points. However, our

proof differs significantly in detailed approaches to achieve these ideas and points due to the presence of position effects and the additional positioning decision in the DAP problem. A complete presentation of the proof is given in Online Appendix A.3, starting with an outline.

5. Extension to Unknown Position Effects

In our TLR-UCB policy, we assume known position effects $\boldsymbol{\theta}^*$. The assumption is reasonable, for example, in online retailing or advertising scenarios where positions are fixed and historical information for position effects is available. However, there are certain cases in which the position effects are unknown. In these scenarios, the joint learning of unknown position effects and unknown preferences is necessary and should be integrated into the goal of revenue maximization.

In this section, we investigate this interesting and important case of unknown position effects in the DAP problem. Note that the reverse scaling of the preferences $\boldsymbol{v}^*$ and position effects $\boldsymbol{\theta}^*$ leads to exactly the same problem instance. Thus, to resolve the identifiability issue under unknown position effects, we force $\theta_1^* = 1$. In the following, we first propose a joint learning procedure of preferences $\boldsymbol{v}^*$ and position effects $\boldsymbol{\theta}^*$. It consists of a newly defined loss function and an AM algorithm with convergence guarantees. Then we propose the EI-TLR policy, which first conducts the joint learning and then implements a new generalized TLR-UCB procedure using estimated position effects. Finally, we establish theoretical guarantees for EI-TLR. Extensive simulation studies and the real data analysis for EI-TLR are conducted in Sections 6 and 7 to evaluate its practical performance.

5.1. Joint Learning of Preferences and Position Effects

In the scenario of unknown position effects, the truncated linear regression estimate and the related UCB construction for the case of known position effects no longer apply. Thus, new estimation techniques need to be developed to jointly learn the unknown preferences and position effects. In the following, we first illustrate a potential type of purchase data we may encounter. Then we motivate a new loss function for the joint estimation by specifying how purchase data depend on unknown preferences and position effects. Finally, to compute the estimate which is defined as the minimizer of the loss function, we propose an AM algorithm and prove its convergence to the estimate.

Suppose we are faced with the purchase data generated from pure explorations in J_0 epochs. In each epoch $\ell \le J_0$, we repeatedly offer the assortment and positioning (S_ℓ, σ_ℓ), which satisfies $|S_\ell| = K$, until a no-purchase outcome occurs. At the end of each epoch ℓ, we compute $\hat{v}_{\ell,k} = \sum_{t\in\mathcal{E}_\ell} \mathbb{1}_{\{c_t = \sigma_\ell^{-1}(k)\}}$, $\forall k \in [K]$, that is, the number of purchases of product $\sigma_\ell^{-1}(k)$ at position $k \in [K]$.

To learn preferences and position effects from the above type of purchase data, our first step is to specify how preferences and position effects are involved in the distribution of the purchase data. For any subset $H \subseteq [K]$, define $\hat{v}_{\ell,H} = \sum_{k\in H} \hat{v}_{\ell,k}$ as the sum of numbers of purchases for products at positions in H. By Lemma S5 in Section B of the Online Appendix, which is indeed an extension of Lemma 2, $\hat{v}_{\ell,H}$ follows a geometric distribution $\mathrm{Geo}(1/(1+\sum_{k\in H} v^*_{\sigma_\ell^{-1}(k)}\theta_k^*))$. Thus, we have $\hat{v}_{\ell,k} \sim \mathrm{Geo}(1/(1+v^*_{\sigma_\ell^{-1}(k)}\theta_k^*))$ and hence $\mathbb{E}(\hat{v}_{\ell,k}) = v^*_{\sigma_\ell^{-1}(k)}\theta_k^*$, $\forall k \in [K]$. Upon this observation, $\forall i \in [N]$, $k \in [K]$, we can aggregate, across the J_0 epochs, a set of purchase data $\{\hat{v}_{\ell,k}\}_{\{\ell\le J_0 : \sigma_\ell^{-1}(k) = i\}}$ with the same expectation of $v_i^*\theta_k^*$.

Now, we define $\boldsymbol{Q}$ as an $N \times K$ matrix whose (i,k)th element is $(\sum_{\tau=1}^{J_0} \mathbb{1}\{\sigma_\tau^{-1}(k) = i\} \cdot \hat{v}_{\tau,k})/(\sum_{\tau=1}^{J_0} \mathbb{1}\{\sigma_\tau^{-1}(k) = i\})$, that is, the average of the above set of purchase data. Because $(\sum_{\tau=1}^{J_0} \mathbb{1}\{\sigma_\tau^{-1}(k) = i\} \cdot \hat{v}_{\tau,k})/(\sum_{\tau=1}^{J_0} \mathbb{1}\{\sigma_\tau^{-1}(k) = i\})$ is a surrogate for $v_i^*\theta_k^*$, the matrix $\boldsymbol{Q}$ serves as an approximation for $\boldsymbol{Q}^* = \boldsymbol{v}^*(\boldsymbol{\theta}^*)^\top$. This motivates us to design a loss function

$$L(\boldsymbol{v}, \boldsymbol{\theta}; \boldsymbol{Q}) = \sum_{i\in[N]} (Q_{i,1} - v_i)^2 + \sum_{i\in[N]} \sum_{k\in[K]\setminus\{1\}} (Q_{i,k} - v_i\theta_k)^2.$$

Note that here we use $\boldsymbol{\theta} = (\theta_2, \ldots, \theta_K)^\top \in \mathbb{R}^{K-1}$ because the first element of the true position effects $\boldsymbol{\theta}^*$ satisfies $\theta_1^* = 1$ due to the identifiability issue as illustrated before and does not need to be estimated. In addition, to align with such $\boldsymbol{\theta}$, we denote $\boldsymbol{\theta}^*_{-1} = (\theta_2^*, \ldots, \theta_K^*)$ as the part of $\boldsymbol{\theta}^*$ that needs estimation. This loss function enjoys a key property that the true preferences and position effects $(\boldsymbol{v}^*, \boldsymbol{\theta}^*_{-1})$ turn out to be the unique minimizer of $L(\boldsymbol{v}, \boldsymbol{\theta}; \boldsymbol{Q}^*)$. Therefore, as $\boldsymbol{Q}$ serves as an approximation for $\boldsymbol{Q}^*$, minimizing over $L(\boldsymbol{v}, \boldsymbol{\theta}; \boldsymbol{Q})$ leads to a reasonable estimate for the true preferences and position effects $(\boldsymbol{v}^*, \boldsymbol{\theta}^*_{-1})$.

In Section B of the Online Appendix, we demonstrate in Lemma S6 the existence of the minimizer. Specifically, as long as the input matrix $\boldsymbol{Q}$ has all positive entries, a minimizer $(\dot{\boldsymbol{v}}, \dot{\boldsymbol{\theta}})$ for the loss function $L(\boldsymbol{v}, \boldsymbol{\theta}; \boldsymbol{Q})$ exists in the region of $(\mathbb{R}^+ \cup \{0\})^{N+K-1}$. Note that each entry of $\boldsymbol{Q}$ is an average of numbers of purchases, and hence the above condition is easily satisfied with a high probability. Thus, we can use $(\dot{\boldsymbol{v}}, \dot{\boldsymbol{\theta}})$ as a joint estimate for the true preferences and position effects $(\boldsymbol{v}^*, \boldsymbol{\theta}^*_{-1})$. Note that because of the noncompact minimization domain and nonconvexity of the loss function, the proof for the existence of a minimizer involves several challenges. For instance, we need to eliminate the case that a divergent sequence $\{(\boldsymbol{v}_n, \boldsymbol{\theta}_n)\}_{n\in\mathbb{N}^+}$ to infinity has loss function values

approaching $\inf_{v,\boldsymbol{\theta}} L(v,\boldsymbol{\theta};\boldsymbol{Q})$, that is, the infimum of the loss function.

Now we propose an algorithm to compute the joint estimate $(\dot{v},\dot{\boldsymbol{\theta}})$. Namely, we need to compute a minimizer for a specific family of loss functions $L(v,\boldsymbol{\theta};\boldsymbol{Q})$ without a closed form. A first challenge for this optimization problem arises from the nonconvexity of $L(v,\boldsymbol{\theta};\boldsymbol{Q})$ in $(v,\boldsymbol{\theta})$. This avoids us from applying the fruitful methods for convex optimization (Boyd and Vandenberghe 2004). Nevertheless, by exploring the biconvex structure of $L(v,\boldsymbol{\theta};\boldsymbol{Q})$, we find that by fixing $v \neq \mathbf{0}$ (or $\boldsymbol{\theta}$), the loss function would be quadratic and convex in the other variable $\boldsymbol{\theta}$ (or v). Thus, it would be straightforward to minimize $L(v,\boldsymbol{\theta};\boldsymbol{Q})$ over $\boldsymbol{\theta}$ (or v) for fixed v (or $\boldsymbol{\theta}$). Specifically, these two marginal minimizations admit the forms of $H_1(v) \triangleq \operatorname{argmin}_{\boldsymbol{\theta}} L(v,\boldsymbol{\theta};\boldsymbol{Q}) = ((\sum_{i=1}^N Q_{i,2}v_i)/(\sum_{i=1}^N v_i^2),\ldots,(\sum_{i=1}^N Q_{i,k}v_i)/(\sum_{i=1}^N v_i^2),\ldots,(\sum_{i=1}^N Q_{i,K}v_i)/(\sum_{i=1}^N v_i^2))$ and $H_2(\boldsymbol{\theta}) \triangleq \operatorname{argmin}_{v} L(v,\boldsymbol{\theta};\boldsymbol{Q}) = ((Q_{1,1}+\sum_{k=2}^K Q_{1,k}\theta_k)/(1+\sum_{k=2}^K \theta_k^2),\ldots,(Q_{i,1}+\sum_{k=2}^K Q_{i,k}\theta_k)/(1+\sum_{k=2}^K \theta_k^2),\ldots,(Q_{N,1}+\sum_{k=2}^K Q_{N,k}\theta_k)/(1+\sum_{k=2}^K \theta_k^2))$. Such a property motivates Algorithm 3.

Algorithm 3 (AM)

1: **Input:** $\boldsymbol{Q} = \{Q_{i,k}\}_{i\in[N],k\in[K]} \in \mathbb{R}^{N\times K}$
2: **Initialization:** $v^{(1)} \in (\mathbb{R}^+)^N$
3: Compute $\boldsymbol{\theta}^{(1)} = H_1(v^{(1)}) = \arg\min_{\boldsymbol{\theta}} L(v^{(1)},\boldsymbol{\theta};\boldsymbol{Q})$.
4: **For** $n = 2,3,\ldots,$ **do**
5: Compute $v^{(n)} = H_2(\boldsymbol{\theta}^{(n-1)}) = \arg\min_{v} L(v,\boldsymbol{\theta}^{(n-1)};\boldsymbol{Q})$.
6: Compute $\boldsymbol{\theta}^{(n)} = H_1(v^{(n)}) = \arg\min_{\boldsymbol{\theta}} L(v^{(n)},\boldsymbol{\theta};\boldsymbol{Q})$.
7: **End for**

As shown in Algorithm 3, the AM algorithm takes a matrix $\boldsymbol{Q} \in \mathbb{R}^{N\times K}$ as the input and initializes a $v^{(1)} \in (\mathbb{R}^+)^N$. Then it conducts an AM procedure that alternately applies $H_1(\cdot)$ and $H_2(\cdot)$ to minimize $L(v,\boldsymbol{\theta};\boldsymbol{Q})$ over $\boldsymbol{\theta}$ (or v) for fixed v (or $\boldsymbol{\theta}$). With an initialization $v^{(1)} \in (\mathbb{R}^+)^N$, this procedure generates a well-defined infinite sequence of $(v^{(n)},\boldsymbol{\theta}^{(n)})_{n\in\mathbb{N}^+}$. Proposition 3 proves that the AM sequence $(v^{(n)},\boldsymbol{\theta}^{(n)})_{n\in\mathbb{N}^+}$ converges to the minimizer of $L(v,\boldsymbol{\theta};\boldsymbol{Q})$, that is, the joint estimate for preferences and position effects. Thus, our proposed Algorithm 3 is a valid computation method.

Proposition 3. *Let $\{(v^{(n)},\boldsymbol{\theta}^{(n)})\}_{n\in\mathbb{N}^+}$ be the infinite sequence generated by Algorithm 3 with the input $\boldsymbol{Q} \in (\mathbb{R}^+)^{N\times K}$ and initialization $v^{(1)} \in \mathbb{N}^+$. If there exists a unique stationary point of $L(v,\boldsymbol{\theta};\boldsymbol{Q})$ within the region $(\mathbb{R}^+ \cup \{0\})^{N+K-1}$, then the sequence $\{(v^{(n)},\boldsymbol{\theta}^{(n)})\}_{n\in\mathbb{N}^+}$ converges to the unique minimizer $(\dot{v},\dot{\boldsymbol{\theta}})$ of $L(v,\boldsymbol{\theta};\boldsymbol{Q})$ within the region $(\mathbb{R}^+ \cup \{0\})^{N+K-1}$.*

The proof for Proposition 3 is faced with several challenges. Firstly, the alternate minimization type of algorithms are not guaranteed to converge. Its convergence property must depend on the loss functions, whereas our adopted $L(v,\boldsymbol{\theta};\boldsymbol{Q})$ is nonconvex and has a quite specific form. In fact, Algorithm 3 may not converge if there exist zero entries in $\boldsymbol{Q}$. Secondly, the two marginal minimization functions $H_1(\cdot)$ and $H_2(\cdot)$, although with closed forms, are rather complicated and hence difficult to analyze. Thirdly, the minimization is conducted on a noncompact set, which avoids the application of many convergence results developed on a compact set. Note that the AM sequence is not easily guaranteed to be bounded, because the $v_i\theta_k$ terms in $L(v,\boldsymbol{\theta};\boldsymbol{Q})$ make near-minimal values of $L(v,\boldsymbol{\theta};\boldsymbol{Q})$ possible for large v or $\boldsymbol{\theta}$.

To overcome the above difficulties and prove Proposition 3, an essential step is to prove that the AM sequence $\{(v^{(n)},\boldsymbol{\theta}^{(n)})\}_{n\in\mathbb{N}^+}$ is indeed bounded. Note that $v^{(n+1)} = H_2(H_1(v^{(n)}))$. Therefore, to prove the boundedness, we fully exploit the algebraic structures of the marginal minimization functions $H_1(\cdot)$ and $H_2(\cdot)$. Specifically, we find that the inner product $\langle v, H_2(H_1(v))\rangle$ of v and $H_2(H_1(v))$, that is, the "next" v, can be formulated as a weighted average of $\langle v,v\rangle$ and $\langle v,\boldsymbol{p}\rangle$, where $\boldsymbol{p}$ denotes the first column of $\boldsymbol{Q}$. Note that this weighted average argument also enjoys a geometric significance that we further exploit. By inequalities such as Cauchy-Schwarz, we establish critical relationships between $\langle H_2(H_1(v)), H_2(H_1(v))\rangle$ and $\langle v,v\rangle$. Then we further utilize this relationship between two consecutive elements in the AM sequence to derive its boundedness.

5.2. EI-TLR Algorithm

In this part, we propose the EI-TLR policy to tackle unknown position effects in DAP. The EI-TLR policy first applies the joint learning procedure established above through an exploration phase. It then implements a generalized TLR-UCB procedure, which is driven by the estimated position effects for preference learning, and incorporates the exploration phase data for acceleration. In the following, we present EI-TLR in Algorithm 4 and provide comprehensive introductions for its components. For a clearer presentation, detailed steps of the generalized TLR-UCB procedure used in Algorithm 4 are presented in Algorithm 5.

Algorithm 4 (EI-TLR)

1: **Input:** Minimum number of exploration epochs $J_0 = J_0(T)$; assortment and positioning sequence $\{(S_\ell,\sigma_\ell)\}_{\ell\in\mathbb{N}^+}$; truncation parameter sequence $\{\alpha_k = \log_{(1+1/\nu_0)}(k)\}_{k\in\mathbb{N}^+}$; regularization parameter $\lambda \in \mathbb{R}^+$; horizon length T
2: **Initialization:** Time period $t = 1$, epoch $\ell = 1$; mean-purchase matrix $\boldsymbol{Q} = \mathbf{0}_{N\times K} \in (\mathbb{R}^+ \cup \{0\})^{N\times K}$, number-of-epoch matrix $\boldsymbol{O} \in \mathbb{N}^{N\times K}$
3: **While** $\ell \le J_0$ or $\exists i \in [N], k \in [K]$ s.t. $Q_{i,k} = 0$ **do** (Exploration phase)

4: **Implement epoch-based offering with** (S_ℓ, σ_ℓ).
5: **For** $k = 1, 2, \ldots, K$ **do** (Update data matrices)
6: Compute $\hat{v}_{\ell,k} = \sum_{t\in\mathcal{E}_\ell} \mathbb{1}_{\{c_t=\sigma_\ell^{-1}(k)\}}$, that is, number of purchases of product $\sigma_\ell^{-1}(k)$ at position k.
7: Update $Q_{\sigma_\ell^{-1}(k),k} = (O_{\sigma_\ell^{-1}(k),k} * Q_{\sigma_\ell^{-1}(k),k} + \hat{v}_{\ell,k})/(O_{\sigma_\ell^{-1}(k),k} + 1) = (\sum_{\tau=1}^{\ell} \mathbb{1}\{\sigma_\tau^{-1}(k) = \sigma_\ell^{-1}(k)\} \cdot \hat{v}_{\tau,k})/(\sum_{\tau=1}^{\ell} \mathbb{1}\{\sigma_\tau^{-1}(k) = \sigma_\ell^{-1}(k)\})$.
8: Update $O_{\sigma_\ell^{-1}(k),k} = O_{\sigma_\ell^{-1}(k),k} + 1 = \sum_{\tau=1}^{\ell} \mathbb{1}\{\sigma_\tau^{-1}(k) = \sigma_\ell^{-1}(k)\}$.
9: **End for**
10: Denote $\hat{v}'_{i,\ell} = \sum_{t\in\mathcal{E}_\ell} \mathbb{1}_{\{c_t=i\}}$, $\forall i \in S_\ell$. (Prepare purchase data for generalized TLR-UCB procedure)
11: $\ell = \ell + 1$.
12: **End while**
13: Apply the AM Algorithm 3 to compute (Joint estimation)

$$(\hat{v}, \hat{\boldsymbol{\theta}}) \in \arg\min_{\boldsymbol{v},\boldsymbol{\theta}} L(\boldsymbol{v}, \boldsymbol{\theta}; \boldsymbol{Q}) \cap (\mathbb{R}^+ \cup \{0\})^{N+K-1},$$

where $L(\boldsymbol{v}, \boldsymbol{\theta}; \boldsymbol{Q}) = \sum_{i\in[N]} (Q_{i,1} - v_i)^2 + \sum_{i\in[N]} \sum_{k\in[K]\setminus\{1\}} (Q_{i,k} - v_i\theta_k)^2$.
14: Denote $J = \ell - 1$ as the actual number of epochs in the exploration phase.
15: Implement the generalized TLR-UCB procedure in Algorithm 5 with inputs
16: $\{\hat{\boldsymbol{\theta}}, \{S_\ell, \sigma_\ell, \{\hat{v}'_{i,\ell}\}_{i\in S_\ell}\}_{\ell\in[J]}, \{\alpha_k\}_{k\in\mathbb{N}^+}, \lambda, T\}$.

5.2.1. Input (Line 1). The EI-TLR policy takes the horizon length T as an input. In addition, it specifies a minimum number of epochs $J_0 = J_0(T)$ for the exploration phase. The third input is a sequence of assortment and positioning pairs $\{(S_\ell, \sigma_\ell)\}_{\ell\in\mathbb{N}^+}$, in which the ℓth pair would be repeatedly offered in the ℓth epoch of the exploration phase.

To guarantee the consistency of the joint estimates for preference parameters and position effects, we would require the number of epochs $J_0(T) \to +\infty$ as the horizon length $T \to +\infty$. In addition, the sequence $\{(S_\ell, \sigma_\ell)\}_{\ell\in\mathbb{N}^+}$ has to place any product $i \in [N]$ at any position $k \in [K]$ for infinite epochs $\ell \in \mathbb{N}^+$. A standard choice would be a periodic-type sequence.

5.2.2. Initialization (Line 2). Besides the time period $t = 1$ and the epoch $\ell = 1$, EI-TLR also initializes two $N \times K$ matrices $\boldsymbol{Q}$ and $\boldsymbol{O}$ to accumulate the purchase data for the joint learning procedure. Specifically, the (i,k)th entry of $\boldsymbol{O} \in \mathbb{N}^{N\times K}$ records the number of epochs that place product i on position k, whereas the (i,k)th entry of $\boldsymbol{Q} \in \mathbb{N}^{N\times K}$ records the average number of purchases for product i over all epochs that place it on position k. We call $\boldsymbol{O}$ the number-of-epoch matrix and $\boldsymbol{Q}$ the mean-purchase matrix.

Under the joint learning procedure developed in Section 5.1, the mean-purchase matrix $\boldsymbol{Q}$ is the key data input for the loss function $L(\boldsymbol{v}, \boldsymbol{\theta}; \boldsymbol{Q})$ to jointly estimate preferences and position effects.

5.2.3. Exploration Phase (Lines 3–12). The EI-TLR policy conducts an exploration phase to generate the purchase data required for the joint learning procedure. In each epoch ℓ, EI-TLR implements the same epoch-based offering strategy as in TLR-UCB (line 4), that is, (S_ℓ, σ_ℓ) is repeatedly offered until a no-purchase outcome occurs. In lines 5–9, we pack the purchase data into $\{\hat{v}_{\ell,k}\}_{k\in[K]}$, where $\hat{v}_{\ell,k}$ denotes the number of purchases of product $\sigma_\ell^{-1}(k)$ at position k. Then we update the two data matrices $\boldsymbol{Q}$ and $\boldsymbol{O}$ accordingly.

By the ending criteria in line 3, EI-TLR may continue explore after J_0 epochs to avoid zero entries in the mean-purchase matrix $\boldsymbol{Q}$. Therefore, the actual number of epochs in the exploration phase, denoted as J, may exceed J_0. Such a design is to ensure the following two properties.

- **Existence**: the joint estimate for preference parameters and position effects, which is defined as the minimizer of the loss function with $\boldsymbol{Q}$ as an input (line 13), does exist;
- **Convergence**: the AM Algorithm 3, which is applied in line 13, does converge.

Moreover, in line 10, we pack the purchase data into a more regular and better shape to incorporate them into the followed generalized TLR-UCB procedure.

5.2.4. Joint Estimate and Its Computation (Line 13). Under unknown position effects, the joint estimation for unknown preferences and position effects is a key challenge. To tackle this challenge, we apply the joint learning procedure established in Section 5.1. Specifically, the joint estimate $(\hat{v}, \hat{\boldsymbol{\theta}})$ is defined as the minimizer of the loss function $L(\boldsymbol{v}, \boldsymbol{\theta}; \boldsymbol{Q})$, in which the input $\boldsymbol{Q}$ is the mean-purchase matrix constructed from the exploration phase. Furthermore, to compute the estimate, we use Algorithm 3 developed in Section 5.1.

By the ending criteria of the exploration phase, all entries of the mean-matrix $\boldsymbol{Q}$ are positive. Therefore, by Lemma S6 in Section B of the Online Appendix, the minimizer and joint estimate $(\hat{v}, \hat{\boldsymbol{\theta}})$ exists and thus is well defined in EI-TLR. Moreover, under conditions in Proposition 3, our adopted AM algorithm converges to the joint estimate $(\hat{v}, \hat{\boldsymbol{\theta}})$, and thus its computation is feasible.

5.2.5. Generalized TLR-UCB Procedure (Lines 14 and 15 and Algorithm 5). With both estimated position effects $\hat{\boldsymbol{\theta}}$ and packed purchase data of exploration

phase $\{S_\ell, \sigma_\ell, \{\hat{v}'_{i,\ell}\}_{i\in S_\ell}\}_{\ell\in[J]}$ available, we are able to implement the generalized TLR-UCB procedure detailed in Algorithm 5. In the following, we elaborate on two key features of the generalized TLR-UCB procedure that differ from the original TLR-UCB policy in Algorithm 1, with references to specific parts in Algorithm 5.

- The generalized TLR-UCB procedure is designed for the case of unknown position effects. It uses the input estimated position effects $\hat{\boldsymbol{\theta}}$ (line 1) to replace the role of the true position effects $\boldsymbol{\theta}^*$ in TLR-UCB. Specifically, the position effects $\{\tilde{\theta}_{i,k}\}_{k\in[T_i(\ell-1)]}$ used for UCB construction (line 9) are constructed from the estimated position effects $\hat{\boldsymbol{\theta}}$, as shown in lines 5 and 15.
- Extra purchase data, as the algorithm's input (line 1), can be incorporated into the generalized TLR-UCB procedure. In fact, the input purchase data $\{S_\ell, \sigma_\ell, \{\hat{v}'_{i,\ell}\}_{i\in S_\ell}\}_{\ell\in[J]}$ from the exploration phase formulate the foundation of all data used in the generalized TLR-UCB procedure. Namely, these extra data work as if they are the purchase data from the very first J epochs, ahead of all truly implemented epochs, and thus are used for any UCB construction in the generalized TLR-UCB.

Compared with TLR-UCB, the adoption of estimated position effects in the generalized TLR-UCB procedure may introduce bias and weaken its performance, whereas the incorporation of extra purchase data may accelerate the preference learning and improve its performance. Interestingly, the generalized TLR-UCB procedure in EI-TLR, although implemented after the exploration phase, actually starts at the beginning of the horizon in terms of the data used. In this sense, the exploration phase can be viewed as a part of the generalized TLR-UCB procedure, which is why we call this algorithm EI-TLR-UCB.

Algorithm 5 (Generalized TLR-UCB Procedure)

1: **Input:** Estimated position effects $\hat{\boldsymbol{\theta}}$; extra purchase data $\{S_\ell, \sigma_\ell, \{\hat{v}'_{i,\ell}\}_{i\in S_\ell}\}_{\ell\in[J]}$; truncation parameter sequence $\{\alpha_k = \log_{(1+1/\nu_0)}(k)\}_{k\in\mathbb{N}^+}$; regularization parameter $\lambda \in \mathbb{R}^+$; horizon length T
2: **Initialization:** Time period $t = 1$, epoch $\ell = 1$; Offered times $T_i(0) = 0$, $\forall i \in [N]$
3: **For** $\ell = 1, 2, \ldots, J$ **do** (Incorporate extra purchase data as the foundation in TLR-UCB)
4: Update $T_i(\ell) = T_i(\ell-1) + 1$, $\forall i \in S_\ell$, and $T_i(\ell) = T_i(\ell-1)$, $\forall i \notin S_\ell$.
5: Denote $\tilde{v}_{i,T_i(\ell)} = \hat{v}'_{i,\ell}$, $\tilde{\theta}_{i,T_i(\ell)} = \hat{\theta}_{\sigma_\ell(i)}$, $\forall i \in S_\ell$. (Reindexing extra purchase data)
6: **End for**
7: **For** $\ell = J+1, J+2, \ldots,$ **do** (TLR-UCB procedure driven by estimated position effects $\hat{\boldsymbol{\theta}}$)
8: **For** $i = 1, 2, \ldots, N$ **do** (UCB construction)
9: Apply the UCB Construction Algorithm 2 with inputs $\{\{\tilde{v}_{i,k}\}_{k\in[T_i(\ell-1)]}, \{\tilde{\theta}_{i,k}\}_{k\in[T_i(\ell-1)]}, \{\alpha_k\}_{k\in\mathbb{N}^+}, \lambda\}$; obtain its output $v^{\mathrm{UCB}}_{i,\ell-1}$.
10:
11: **End for**
12: Set the assortment and positioning $(S_\ell, \sigma_\ell) = \arg\max_{|S|\le K,\sigma} R(S, \sigma, \boldsymbol{v}^{\mathrm{UCB}}_{\ell-1}, \hat{\boldsymbol{\theta}})$,
13: where $\boldsymbol{v}^{\mathrm{UCB}}_{\ell-1} = (v^{\mathrm{UCB}}_{1,\ell-1}, v^{\mathrm{UCB}}_{2,\ell-1}, \ldots, v^{\mathrm{UCB}}_{N,\ell-1})^\top$.
14: **Implement epoch-based offering with** (S_ℓ, σ_ℓ).
15: Compute the number of purchases of product i as $\hat{v}_{i,\ell} = \sum_{t\in\mathcal{E}_\ell} \mathbb{1}_{\{c_t=i\}}$, $\forall i \in S_\ell$.
16: Update $T_i(\ell) = T_i(\ell-1) + 1$, $\forall i \in S_\ell$, and $T_i(\ell) = T_i(\ell-1)$, $\forall i \notin S_\ell$.
17: Denote $\tilde{v}_{i,T_i(\ell)} = \hat{v}_{i,\ell}$, $\tilde{\theta}_{i,T_i(\ell)} = \hat{\theta}_{\sigma_\ell(i)}$, $\forall i \in S_\ell$; update $\ell = \ell + 1$. (Reindexing)
18: **End for**

5.3. Consistency of the Joint Estimate in the EI-TLR Policy

Following the exploration phase of EI-TLR, the estimated position effects are used to drive the generalized TLR-UCB procedure. Thus, the accuracy of estimated position effects is crucial for reducing the regret from the generalized TLR-UCB procedure. For this reason, it is essential to investigate the consistency of the joint estimate, that is, whether it converges to the true preferences and position effects as $T \to +\infty$.

Denote $\boldsymbol{Q}(\ell)$ as the updated mean-purchase matrix $\boldsymbol{Q}$ at the end of epoch ℓ. To investigate the consistency, we need to evaluate the mean-purchase matrix $\boldsymbol{Q}(J(T))$ because it serves as an input for the loss function and helps derive the joint estimate. In fact, $\forall i \in [N], k \in [K]$, the (i,k)th entry of $\boldsymbol{Q}(J_0(T))$, is the average of i.i.d. random variables following the distribution of Geo$(1/(1+v_i^*\theta_k^*))$. Therefore, as long as $\boldsymbol{O}(J_0(T))_{i,k}$, that is, the number of these i.i.d. random variables, goes to infinity $\forall i \in [N], k \in [K]$ as $T \to +\infty$, $\boldsymbol{Q}(J_0(T))$ would converge to $\boldsymbol{Q}^*$ in probability. In addition, $\boldsymbol{Q}(J(T))$ would also converge to $\boldsymbol{Q}^*$ because $J(T)$ would coincide with $J_0(T)$ with probability converging to one as $J_0(T) \to +\infty$. Note that $(\hat{\boldsymbol{v}}, \hat{\boldsymbol{\theta}})$ minimizes $L(\boldsymbol{v}, \boldsymbol{\theta}; \boldsymbol{Q}(J(T)))$ and $(\boldsymbol{v}^*, \boldsymbol{\theta}^*_{-1})$ minimizes $L(\boldsymbol{v}, \boldsymbol{\theta}; \boldsymbol{Q}^*)$. Therefore, to evaluate whether $(\hat{\boldsymbol{v}}, \hat{\boldsymbol{\theta}})$ could approach $(\boldsymbol{v}^*, \boldsymbol{\theta}^*_{-1})$, we develop the following Proposition 4, which bounds the difference between $(\boldsymbol{v}^*, \boldsymbol{\theta}^*_{-1})$ and the minimizer of $L(\boldsymbol{v}, \boldsymbol{\theta}; \boldsymbol{Q})$ when $\boldsymbol{Q}$ is close to $\boldsymbol{Q}^*$.

Proposition 4. *Assume* $\boldsymbol{v}^* \in (\mathbb{R}^+)^N, \boldsymbol{\theta}^*_{-1} \in (\mathbb{R}^+)^{K-1}$. *Denote* $\boldsymbol{Q}^* = \boldsymbol{v}^*(\boldsymbol{\theta}^*)^\top \in (\mathbb{R}^+)^{N\times K}$. *Then,* $\forall \epsilon > 0$, *there exists* $\delta > 0$ *such that* $\forall \boldsymbol{Q} \in \mathbb{R}^{N\times K}$ *that satisfies* $\|\boldsymbol{Q} - \boldsymbol{Q}^*\|_2 \le \delta$, *we have* $\|(\dot{\boldsymbol{v}}, \dot{\boldsymbol{\theta}}) - (\boldsymbol{v}^*, \boldsymbol{\theta}^*_{-1})\|_2 \le \epsilon$ *for any* $(\dot{\boldsymbol{v}}, \dot{\boldsymbol{\theta}}) \in \arg\min_{v,\boldsymbol{\theta}} L(\boldsymbol{v}, \boldsymbol{\theta}; \boldsymbol{Q})$.

Proposition 4 resembles a continuity argument for the mapping from $\boldsymbol{Q}$ to the minimizer of $L(\boldsymbol{v},\boldsymbol{\theta};\boldsymbol{Q})$. It is in the same essence of the maximum theorem (Berge 1963) that concerns the continuity of maximizers of a function with respect to its parameters. However, the maximum theorem requires compact maximization domains and concludes on upper hemicontinuity. Thus, Proposition 4 displays significant differences. Its proof is challenging and significantly exploits the properties of the loss function. The proof is roughly decomposed into two parts. The first part demonstrates that $L(\boldsymbol{v},\boldsymbol{\theta};\boldsymbol{Q}^*)$ can only approach $L(\boldsymbol{v}^*,\boldsymbol{\theta}^*_{-1};\boldsymbol{Q}^*)$ when $(\boldsymbol{v},\boldsymbol{\theta})$ is close to $(\boldsymbol{v}^*,\boldsymbol{\theta}^*_{-1})$; the second part partitions the noncompact minimization domain into two parts $\mathcal{D}$ and $\mathcal{D}^c$. It then proves, for $\boldsymbol{Q}$ in a neighborhood of $\boldsymbol{Q}^*$, the uniform incompetence of $\mathcal{D}$ in minimizing $L(\boldsymbol{v},\boldsymbol{\theta};\boldsymbol{Q})$, and uniform closeness of the two functions $L(\boldsymbol{v},\boldsymbol{\theta};\boldsymbol{Q})$ and $L(\boldsymbol{v},\boldsymbol{\theta};\boldsymbol{Q}^*)$ on $\mathcal{D}^c$. Combining these two parts, we can conclude that for $\boldsymbol{Q}$ close to $\boldsymbol{Q}^*$, only those $(\boldsymbol{v},\boldsymbol{\theta})$ that are close to $(\boldsymbol{v}^*,\boldsymbol{\theta}^*_{-1})$ could survive as candidates to minimize $L(\boldsymbol{v},\boldsymbol{\theta};\boldsymbol{Q})$.

By applying Proposition 4, we prove the consistency of the joint estimate in Theorem 3.

Theorem 3. *Let $\{(S_\ell,\sigma_\ell)_{\ell\in\mathbb{N}^+}\}$ be an assortment and positioning sequence, and $\{\boldsymbol{O}(\ell)\in\mathbb{N}^{N\times K}\}_{\ell\in\mathbb{N}^+}$ be the number-of-epoch matrices such that $O(\ell)_{i,k}=\sum_{\tau=1}^{\ell}\mathbb{1}\{\sigma_\tau^{-1}(k)=i\}$. If the number of exploration epochs $J_0(T)$ satisfies $\lim_{T\to+\infty}J_0(T)=+\infty$, and $\lim_{\ell\to+\infty}O(\ell)_{i,k}=+\infty$, $\forall i\in[N],k\in[K]$, the estimated preferences and position effects $(\hat{\boldsymbol{v}},\hat{\boldsymbol{\theta}})$ in the EI-TLR policy is consistent.*

By Theorem 3, for long enough horizon, the joint estimate $(\hat{\boldsymbol{v}},\hat{\boldsymbol{\theta}})$ can be as accurate as possible. Thus, the bias introduced by using estimated position effects in the generalized TLR-UCB procedure would vanish, leading to low additional per-period regret compared with an oracle version that adopts true position effects. In addition, the conditions in Theorem 3 on $J_0(T)$ and the sequence $\{(S_\ell,\sigma_\ell)_{\ell\in\mathbb{N}^+}\}$ are easy to fulfill. For instance, $J_0(T)=c\cdot T^\alpha$ for any $\alpha\in(0,1)$, and any periodic choice of $\{(S_\ell,\sigma_\ell)_{\ell\in\mathbb{N}^+}\}$ would suffice. Furthermore, in Section 6, we conduct extensive simulations on EI-TLR and illustrate an estimation error rate around $J_0^{-1/2}$ for the joint estimate and show sublinear regret rates for EI-TLR.

Remark 3 (Challenges in Deriving Theoretical Error Bounds). The consistency of joint estimates in the EI-TLR policy helps control the additional per-period regret in the generalized TLR-UCB procedure that attributes to errors in its adopted estimated position effects. Further deriving a regret upper bound for EI-TLR requires a nonasymptotic error bound for the joint estimate. This involves tackling several new challenges. Firstly, the observed numbers of purchases for different products, which correspond to noisy matrix elements, are dependent. Note that for any $H\subseteq[K]$ with $|H|\geq 2$, both the sum $\sum_{k\in H}\hat{v}_{\ell,k}$ and the elements $\{\hat{v}_{\ell,k}\}_{k\in H}$ follow geometric distributions by Lemma S5 in Section B of the Online Appendix. Therefore, it is easy to illustrate that $\{\hat{v}_{\ell,k}\}_{k\in H}$, which contributes to different entries of $\boldsymbol{Q}$, are dependent. Secondly, in each epoch of the exploration phase, we are indeed sampling a special group of entries $\{(i,\sigma(i)):i\in S\}$ to estimate $\boldsymbol{v}^*(\boldsymbol{\theta}^*)^\top$ when offering (S,σ). Because there is a one-to-one matching between the products and positions, the sampled entries are also dependent. Thirdly, the noisy matrix elements are geometric and hence not sub-Gaussian. This largely affects the probabilistic derivations for error bounds. Fourthly, to develop online learning algorithms, we need to establish more refined entrywise error bounds instead of the Frobenius norm.

6. Simulation Study

In this section, we perform numerical studies on synthetic data sets to evaluate the practical performance of our proposed TLR-UCB and EI-TLR policies. In particular, we conduct two separate comparisons under known and unknown position effects. Moreover, through simulations, we further investigate the rates of estimation errors and cumulative regrets for our proposed EI-TLR policy.

The first comparison is under known position effects, including TLR-UCB and two benchmark policies LR-UCB and A-UCB-V that also make use of known position effects. In the following, we briefly introduce linear regression upper confidence bound **(**LR-UCB) and A-UCB variant (A-UCB-V). More details can be found in Section E of the Online Appendix.

- **LR-UCB**, a variant of TLR-UCB, which avoids the truncating steps and replaces the UCB bonus in TLR-UCB with a standard one.
- **A-UCB-V**, a variant of the UCB algorithm in Agrawal et al. (2019), which leverages known position effects in the preference estimation and static assortment optimization.

The second comparison is under unknown position effects, including EI-TLR and two assortment-only algorithms S-ETC and A-UCB that ignore position effects.

- **S-ETC**, the ETC-type algorithm 3 in Sauré and Zeevi (2013), which ignores position effects.
- **A-UCB**, the UCB algorithm in Agrawal et al. (2019), which ignores position effects.

Then, we introduce the three simulation settings. From Examples 1–3, we gradually increase the number of products N and the number of positions K. In addition, we also vary the position effects, preference vectors, and revenue vectors.

- Example 1: $N=3$ products with $K=2$ positions; Position effect vector $\boldsymbol{\theta}^*=(1,1/2)^\top$; Preference vector $\boldsymbol{v}^*=(1/4,2/5,4/5)^\top$; Revenue vector $r=(4/5,3/4,1/2)^\top$.

- Example 2: $N = 5$ products with $K = 3$ positions; Position effect vector $\boldsymbol{\theta}^* = (1, 1/2, 1/3)^\top$; Preference vector $\boldsymbol{v}^* = (1, 0.8, 0.6, 0.4, 0.2)^\top$; Revenue vector $r = (0.4, 0.2, 0.8, 0.6, 1)^\top$.
- Example 3: $N = 10$ products with $K = 5$ positions; Position effect vector $\boldsymbol{\theta}^* = (1, 1/2, 1/3, 1/4, 1/5)^\top$; Preference vector $\boldsymbol{v}^* = (1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1)^\top$; Revenue vector $r = (0.3, 0.2, 0.7, 0.4, 0.5, 0.6, 0.1, 0.9, 1, 0.8)^\top$.

6.1. Comparison Under Known Position Effects

In this part, we conduct the comparison among TLR-UCB, LR-UCB, and A-UCB-V, all using known position effects. For our proposed TLR-UCB policy, we use $\nu_0 = 1$ and $\lambda = 1$ under all three settings. We perform 500 replications and average the results for each method.

Figure 1 shows the cumulative regret curves of TLR-UCB, LR-UCB, and A-UCB-V under Examples 1–3. As we can see, the performance of LR-UCB on all three simulation settings are much worse than TLR-UCB. This demonstrates the importance of the special UCB bonus used in TLR-UCB. Namely, it is essential to design a UCB bonus tailored to our truncated linear regression estimate, so that we are able to tackle the random and adaptive position effects, and the geometric tails. On the other hand, A-UCB-V performs worse than TLR-UCB on the first two simulation settings in a horizon of length $T = 200{,}000$. On the third setting, because of its linearly increasing cumulative regret, A-UCB-V is also gradually outperformed by TLR-UCB in a longer horizon of length $T = 1{,}000{,}000$. The originally constructed UCB bonus is no longer valid for the new preference estimates due to the scaling of $\hat{v}_{i,\ell}$ with random and adaptive position effects $\theta_{\sigma_\ell(i)}$. This demonstrates the essence of our strategy in using the truncated linear regression estimate and constructing new UCB bonus by deriving its concentration in TLR-UCB.

6.2. Comparison Under Unknown Position Effects

In this part, we conduct the comparison among EI-TLR, S-ETC, and A-UCB. All three methods do not know the position effects. However, our proposed EI-TLR policy is aware of position effects and acts accordingly, whereas S-ETC and A-UCB are not. For EI-TLR, we set the minimum number of exploration epochs as $J_0(T) = \lceil c \cdot T^{2/3} \rceil$ with $c = 0.1$. Note that such a choice of the exploration phase is quite common in bandit algorithms (Lattimore and Szepesvári 2020), and we fix the same constant c across Examples 1–3. In addition, the sequence of assortment and positioning $\{(S_\ell, \sigma_\ell)\}_{\ell \in \mathbb{N}^+}$ satisfies $S_\ell = \{([(\ell-1)+i-1] \bmod N)+1 | i \in [K]\}$ and $\sigma_\ell[([(\ell-1)\cdot\kappa + i - 1] \bmod N) + 1] = i$, $\forall i \in [K]$. Lastly, we use the same $\nu_0 = 1$ and $\lambda = 1$ as TLR-UCB for all three settings.

Figure 2 shows the cumulative regret curves of EI-TLR, S-ETC, and A-UCB under Examples 1–3, averaged over 500 replications. As we can see, EI-TLR substantially outperforms the position-unaware algorithms S-ETC and A-UCB. Such a significant advantage is mainly attributed to the quite flat regret curves and low per-period regrets in the latter part of the generalized TLR-UCB procedure. This demonstrates the benefits of preference learning in the generalized TLR-UCB procedure, despite being driven by estimated position effects. Moreover, the incorporation of exploration phase data further reduces the regrets from the generalized TLR-UCB procedure by accelerating the preference learning and cutting off a portion of the surging exploration costs at the beginning of this procedure. In contrast, both S-ETC and A-UCB display a clear linearly increasing pattern in their cumulative regret curves. This matches our theoretical claims in Lemma 1 that the assortment-only algorithms perform poorly under position effects.

6.3. Further Investigations on EI-TLR

For EI-TLR, we conduct two more experiments to investigate the rate of estimation errors and cumulative regrets respectively. Because of space limitations, we only display the results for Example 1 in Figure 3. Similar results for Examples 2 and 3 can be found in Section E of the Online Appendix.

Figure 1. (Color online) Regret Comparison of Our Proposed TLR-UCB Policy with LR-UCB and A-UCB-V Under Known Position Effects

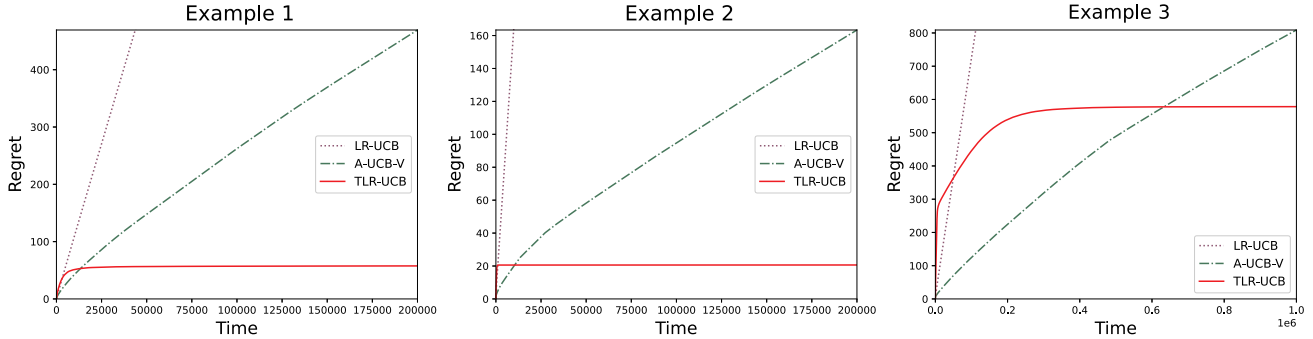
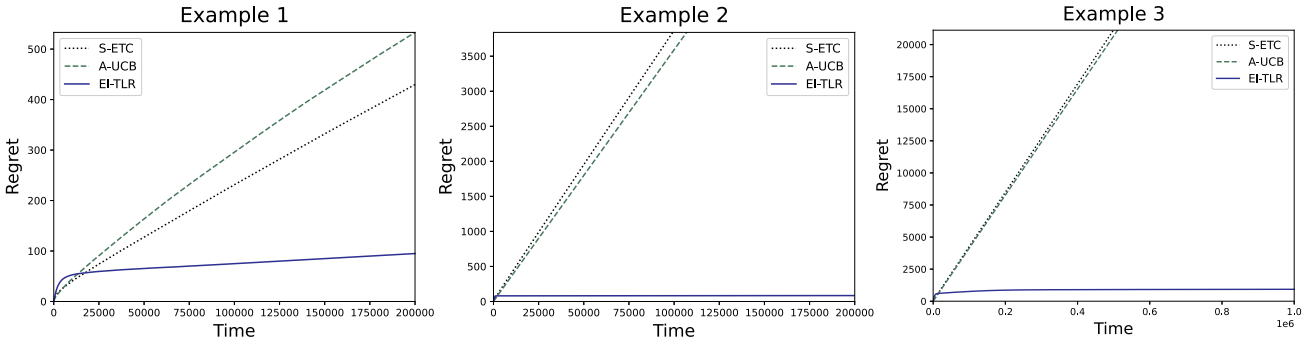


Figure 2. (Color online) Regret Comparison of Our Proposed EI-TLR Policy with S-ETC and A-UCB Under Unknown Position Effects



To illustrate how estimation errors vary with the epoch number J_0, we design a set of 20 J_0's ranging from 1,000 to 20,000 in which the ith J_0 equals $1{,}000 \times 20^{(i-1)/19}$. In the left and middle subplots of Figure 3, we plot the logarithms of estimation errors $\log(\|\hat{v} - v^*\|_2)$ and $\log(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*_{-1}\|_2)$ versus $\log(J_0)$, which display a strong linear trend. The linear regression fits display a clear $-1/2$ slope, which demonstrates that the estimation error rates are around $J_0^{-1/2}$. On the other hand, to illustrate the regret rates, we design a set of 20 horizon lengths T ranging from 200,000 to 2,000,000 in which the ith T equals $200{,}000 \times 10^{(i-1)/19}$. We set the same number of epochs $J_0(T)$ and the sequence of assortment and positioning $\{(S_\ell, \sigma_\ell)\}_{\ell \in \mathbb{N}^+}$ as the regret comparison part. In the right subplot of Figure 3, we show the regret rates by plotting the logarithmic cumulative regrets over logarithmic horizon lengths. As we can see, the plot illustrates a clear sublinearity of cumulative regrets over horizon lengths for our proposed EI-TLR policy.

7. Real Data Analysis

In this section, we present a numerical study using the "UCI Car Evaluation Database" as our real data set, which has also been explored in Agrawal et al. (2019). It consists of $N = 1{,}728$ cars, each accompanied by consumer ratings and various other attributes. Specifically, each car is rated at one of the four levels: unacceptable, acceptable, good, and very good. In addition, each car has six categorical attributes: price, maintenance costs, number of doors, passenger capacity, luggage capacity, and safety perception. Their detailed values are presented in Table 1.

To encode these categorical attributes, we convert them into 21 new attributes with binary values by adding dummy attributes. For example, "price very high" and "price high" form into two different attributes with values of one or zero. In addition, we add an intercept for each car. Thus, each car is associated with an attribute vector $\boldsymbol{x}_i \in \{0,1\}^{22}, i \in [1,728]$.

We next estimate the ground truth preference parameters for these cars. We assume the true preference parameter v_i^* of a car is modeled as the exponential of a linear function of its attributes $\boldsymbol{x}_i$. Namely, there exists a fixed linear parameter $\boldsymbol{\zeta} \in \mathbb{R}^{22}$ such that $v_i^* = e^{\boldsymbol{x}_i^\top \boldsymbol{\zeta}}$, $\forall i \in [N]$. Similar to Agrawal et al. (2019), we assume a logistic probabilistic model on the consumer ratings given the car's preference parameter. Specifically, the consumer ratings of acceptable, good, and very good are considered as intention to buy and encoded as $y_i = 1$. The consumer rating of unacceptable is viewed as no intention to buy and encoded as $y_i = 0$. Then we have $\mathbb{P}(y_i = 1 | \boldsymbol{\zeta}, \boldsymbol{x}_i) = e^{\boldsymbol{x}_i^\top \boldsymbol{\zeta}}/(1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\zeta}})$, $\mathbb{P}(y_i = 0 | \boldsymbol{\zeta}) = 1/(1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\zeta}})$. Finally, we obtain the ℓ_2-regularized

Figure 3. (Color online) Rates of Estimation Errors and Cumulative Regrets for Our Proposed EI-TLR Policy in Example 1

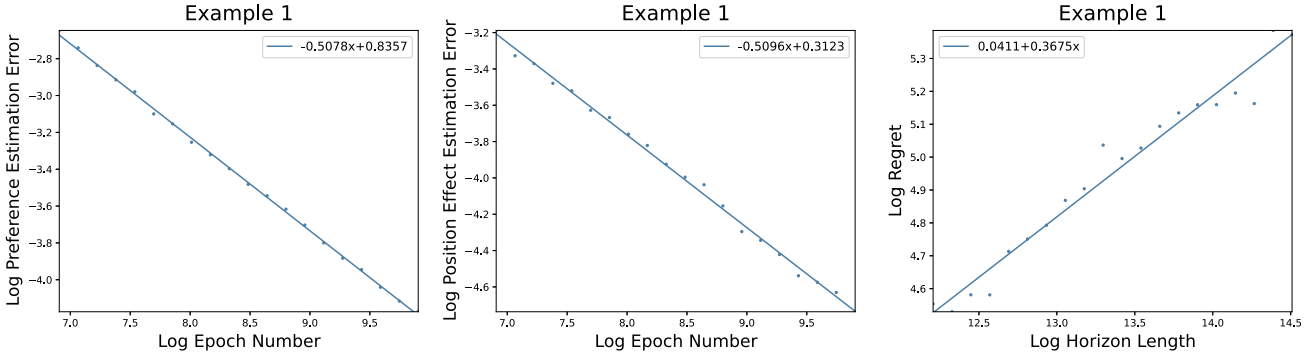


Table 1. Attribute Information of Cars for the UCI Car Evaluation Database

Attribute	Attribute values
Price	Very high, high, medium, and low
Maintenance costs	Very high, high, medium, and low
No. of doors	Two, three, four, and more
Passenger capacity	Two, four, and more
Luggage capacity	Small, medium, and big
Safety perception	Low, medium, and high

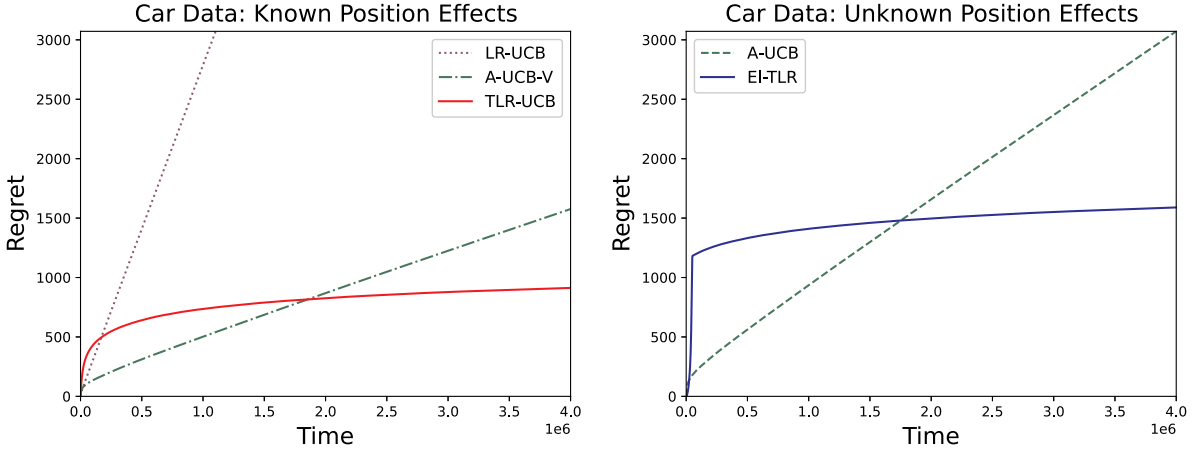
maximum likelihood estimate $\boldsymbol{\zeta}_{\text{MLE}}$ as $\boldsymbol{\zeta}_{\text{MLE}} = \arg\min_{\boldsymbol{\zeta}} -C\sum_{i=1}^{N}(y_i \log \mathbb{P}(y_i = 1|\boldsymbol{\zeta}, \boldsymbol{x}_i) + (1-y_i)\log \mathbb{P}(y_i = 0|\boldsymbol{\zeta}, \boldsymbol{x}_i)) + \frac{1}{2}\|\boldsymbol{\zeta}\|_2^2$, where the regularization constant C is set as 0.01. We then use this estimate $\boldsymbol{\zeta}_{\text{MLE}}$ to calculate the preference parameter of each car by using the formula $v_i^* = e^{\boldsymbol{x}_i^\top \boldsymbol{\zeta}}$, $\forall i \in [N]$. Similar to Agrawal et al. (2019), we assume that the retailer can only display at most 100 cars to the customer at each time period. Namely, there are $K = 100$ positions in our formulated DAP problem. We consider the exponentially decaying position effects $\theta_i^* = 2^{-2(i-1)/(K-1)}$, $\forall i \in [K]$. Namely, we have the first and maximal position effect $\theta_1^* = \theta_{\max}^* = 1$, and the last and minimal position effect $\theta_K^* = \theta_{\min}^* = \frac{1}{4}$. These position effects together with the estimated preference parameters are used as the ground truth to compare the algorithm performances.

Similar to simulation settings, we conduct comparisons for known and unknown position effects, respectively. In the case of known position effects, we compare our proposed TLR-UCB policy with LR-UCB and A-UCB-V; under unknown position effects, we compare our proposed EI-TLR policy with the assortment-only algorithm A-UCB in Agrawal et al. (2019). We do not include a comparison with S-ETC in Sauré and Zeevi (2013) for two reasons. Firstly, S-ETC is computationally intensive because it requires solving relatively large-scale static optimization problems frequently throughout the horizon. Secondly, S-ETC has comparable performance to A-UCB in our simulation studies and is outperformed by A-UCB on this car data set as reported in Agrawal et al. (2019).

For TLR-UCB, similar to simulations, we specify the inputs as $v_0 = \max_{i\in[N]} v_i^*$ and $\lambda = 1$. For EI-TLR, we use similar inputs in the simulations as well. Namely, we set the number of epochs $J_0(T) = \lceil c \cdot T^{2/3} \rceil$; the sequence of assortment and positioning $\{(S_\ell, \sigma_\ell)\}_{\ell\in\mathbb{N}^+}$ satisfies $S_\ell = \{([(\ell-1)+i-1] \bmod N)+1 | i \in [K]\}$ and $\sigma_\ell[([(\ell-1)\cdot\kappa + i - 1] \bmod N)+1] = i$, $\forall i \in [K]$. In addition, we adopt the same $v_0 = \max_{i\in[N]} v_i^*$ and $\lambda = 1$ as those used in TLR-UCB.

In Figure 4, we plot the cumulative regrets for all policies by averaging over 50 replications. Under both known and unknown position effects, our proposed TLR-UCB and EI-TLR policy outperform other benchmark policies. Additionally, we align the range of regrets for these two plots to compare the performance between TLR-UCB and EI-TLR. Despite being designed for different cases, these two policies share similar TLR-UCB elements and thus several interesting points arise from their comparison. Firstly, as expected, TLR-UCB accumulates lower regrets than EI-TLR as it tackles an easier setting of known position effects. Secondly, the difference of per-period regrets in latter periods of these two policies is quite minor. Namely, when sufficient preference learning is conducted, the additional per-period regret of the generalized TLR-UCB procedure due to its utilization of estimated position effects can be low. Thirdly, because of the incorporation of exploration phase data, the regret curve of the generalized TLR-UCB procedure increases more slowly at the beginning than that of TLR-UCB.

Figure 4. (Color online) Regret Comparison of Our Proposed TLR-UCB and EI-TLR Policy with LR-UCB, A-UCB-V, and A-UCB on the UCI Car Evaluation Database



Acknowledgments

The authors thank the editor-in-chief (Amy R. Ward) and area editor (Xi Chen) for guidance and oversight throughout the review process and the associate editor and anonymous reviewers for insightful comments and constructive suggestions. The code and data to support the numerical experiments in this paper can be found at https://github.com/yiyun851/Assortment-Positioning/blob/main/Code_Data.zip.

References

Abbasi-Yadkori Y, Pál D, Szepesvári C (2011) Improved algorithms for linear stochastic bandits. Shawe-Taylor J, Zemel RS, Bartlett PL, Pereira F, Weinberger KQ, eds. *Adv. Neural Inform. Processing Systems*, vol. 24 (Curran Associates, Red Hook, NY), 2312–2320.

Abeliuk A, Berbeglia G, Cebrian M, Van Hentenryck P (2016) Assortment optimization under a multinomial logit model with position bias and social influence. *4OR* 14(1):57–75.

Agrawal S, Avadhanula V, Goyal V, Zeevi A (2017) Thompson sampling for the MNL-bandit. Kale S, Shamir O, eds. *Proc. 30th Conf. Learn. Theory* (PMLR, New York), 76–78.

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

Aouad A, Segev D (2021) Display optimization for vertically differentiated locations under multinomial logit preferences. *Management Sci.* 67(6):3519–3550.

Aouad A, Farias V, Levi R (2021) Assortment optimization under consider-then-choose choice models. *Management Sci.* 67(6): 3368–3386.

Aznag A, Goyal V, Perivier N (2021) MNL-bandit with knapsacks: A near optimal algorithm. Preprint, submitted June 2, https://arxiv.org/abs/2106.01135.

Berge C (1963) *Topological Spaces* (Oliver and Boyd, Edinburgh, UK).

Boyd S, Vandenberghe L (2004) *Convex Optimization* (Cambridge University Press, Cambridge, UK).

Bubeck S, Cesa-Bianchi N, Lugosi G (2013) Bandits with heavy tail. *IEEE Trans. Inform. Theory* 59(11):7711–7717.

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

Chen X, Wang Y (2018) A note on a tight lower bound for capacitated mnl-bandit assortment selection models. *Oper. Res. Lett.* 46(5):534–537.

Chen X, Wang Y, Zhou Y (2021a) Optimal policy for dynamic assortment planning under multinomial logit models. *Math. Oper. Res.* 46(4):1639–1657.

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

Chen J, Dong H, Wang X, Feng F, Wang M, He X (2023) Bias and debias in recommender system: A survey and future directions. *ACM Trans. Inform. Systems* 41(3):1–39.

Cheung WC, Tan V, Zhong Z (2019) A Thompson sampling algorithm for cascading bandits. Chaudhuri K, Sugiyama M, eds. *Proc. 22nd Internat. Conf. Artificial Intelligence Statist.*, vol. 89 (PMLR, New York), 438–447.

Craswell N, Zoeter O, Taylor M, Ramsey B (2008) An experimental comparison of click position-bias models. Najork M, Broder AZ, Chakrabarti S, eds. *Proc. 2008 Internat. Conf. Web Search Data Mining (Palo Alto, California)*, 87–94.

Feldman J, Segev D (2022) The multinomial logit model with sequential offerings: Algorithmic frameworks for product recommendation displays. *Oper. Res.* 70(4):2162–2184.

Foussoul A, Goyal V, Gupta V (2023) MNL-bandit in non-stationary environments. Preprint, submitted March 4, https://arxiv.org/abs/2303.02504.

Gallego G, Li A, Truong VA, Wang X (2020) Approximation algorithms for product framing and pricing. *Oper. Res.* 68(1):134–160.

Ke C, Wang R, Zhao Z (2023) Discrete choice models with piecewise linear utility: Modeling, estimation and pricing. Preprint, submitted March 20, http://dx.doi.org/10.2139/ssrn.4394213.

Kveton B, Szepesvari C, Wen Z, Ashkan A (2015) Cascading bandits: Learning to rank in the cascade model. Bach F, Blei D, eds. *Proc. 32nd Internat. Conf. Machine Learn.*, vol. 36 (PMLR, New York), 767–776.

Lattimore T, Szepesvári C (2020) *Bandit Algorithms* (Cambridge University Press, Cambridge, UK).

Li S, Luo Q, Huang Z, Shi C (2025) Online learning for constrained assortment optimization under Markov chain choice model. *Oper. Res.* 73(1):109–138.

Li S, Wang B, Zhang S, Chen W (2016) Contextual combinatorial cascading bandits. Balcan MF, Weinberger KQ, eds. *Proc. 33rd Internat. Conf. Machine Learn.*, vol. 48 (PMLR, New York), 1245–1253.

Medina AM, Yang S (2016) No-regret algorithms for heavy-tailed linear bandits. Balcan MF, Weinberger KQ, eds. *Proc. 33rd Internat. Conf. Machine Learn.*, vol. 48 (PMLR, New York), 1642–1650.

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

Rusmevichientong P, Shen ZJM, Shmoys DB (2010) Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Oper. Res.* 58(6):1666–1680.

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

Shen S, Chen X, Fang E, Lu J (2023) Combinatorial inference on the optimal assortment in multinomial logit models. Preprint, submitted January 28, https://arxiv.org/abs/2301.12254.

Wang R, Zhao Z, Ke C (2022) Modeling consumer choice and optimizing assortment under the threshold multinomial logit model. Preprint, submitted August 8, https://doi.org/10.2139/ssrn.4184044.

Zhalechian M, Keyvanshokooh E, Shi C, Van Oyen MP (2022) Online resource allocation with personalized learning. *Oper. Res.* 70(4): 2138–2161.

Yiyun Luo is an assistant professor in the School of Statistics and Data Science at Shanghai University of Finance and Economics. His research focuses on the design and application of bandit algorithms to improve online decision making, particularly in dynamic pricing and assortment optimization.

Will Wei Sun is an associate professor of quantitative methods at the Daniels School of Business, Purdue University, and is also affiliated with the Department of Statistics. His research centers on statistical foundations of large language models, trustworthy reinforcement learning, and online decision making in two-sided markets. His research has been partially supported by grants from the National Science Foundation.

Yufeng Liu is a professor in the Department of Statistics and Operations Research, Department of Biostatistics, and Department of Genetics at University of North Carolina at Chapel Hill. His current research interests include statistical machine learning, high-dimensional data analysis, bioinformatics, individualized decision making, and e-commerce. His research has been partially supported by grants from the National Science Foundation.