

On Size Substitution and Its Role in Assortment and Inventory Planning

Yi-Chun Akchen

School of Management, University College London, London E14 5AB, United Kingdom.
yi-chun.akchen@ucl.ac.uk

Felipe Caro

Anderson School of Management, University of California, Los Angeles, California 90095, United States.
felipe.caro@anderson.ucla.edu

Problem definition: How should (apparel) retailers manage product sizes? For example, if most customers wearing a given shoe size, such as 9.5, are willing to accept a half-size up or down, is it necessary for a retailer to carry that size at all? Additionally, while identical products in different sizes are treated as distinct SKUs in inventory management, they are often aggregated for assortment and strategic planning, although there is no theoretical justification showing that such aggregation approach is sensible. In this paper, we address these fundamental questions about size management, which have remained largely unexplored in the operations literature. **Methodology/results:** We propose a choice model where each customer forms a consideration set based on the in-stock availability of products of her best-fit size and adjacent sizes. Using a real-world dataset from a large footwear retailer, we show that nearly 25% of the unmet demand caused by stockouts spills over to adjacent sizes. We further solve the assortment and inventory optimization problems under the proposed choice model. Our findings show that the optimal assortment remains unchanged regardless of the likelihood that customers might purchase adjacent sizes. We utilize this finding and further show that inventory policies that ignore size substitution can be (asymptotically) optimal when the demand rate is high or the selling horizon is long. We also propose a mixed-integer program to find inventory levels that account for size substitution and obtain higher profits in the low-demand settings. **Managerial implications:** We show that the prevalent size-aggregation approach adopted in apparel retail operations is sensible in high-demand settings such as e-commerce. In contrast, when the expected demand over the selling horizon is low, size substitution can be relevant and should be taken into account in stocking decisions.

Key words: retail operations; demand substitution; apparel product size; stockout; choice modeling; fashion; assortment and inventory optimization.

History: First version: November 22, 2023. This version: March 25, 2025.

1. Introduction

In recent years, firms and academia have witnessed the success of operational models in the apparel industry. Various analytical models have been proposed to improve operations efficiency and create value (Caro and Martínez-de Albéniz 2013). A cornerstone of these models is product demand estimation, which informs critical decisions such as inventory allocation (Caro and Gallien 2010, Caro et al. 2010), price markdown optimization (Caro and Gallien 2012), and initial shipments from warehouses to stores (Gallien et al. 2015).

The most common approach in the operations management literature to estimate demand and product substitution is the following. First, a “product” is viewed as the aggregation across sizes of stock keeping units (SKUs) of the same style. Second, demand is estimated based on the aggregated

units (Boada-Collado and Martínez-de Albéniz 2020). In this approach, demand substitution is restricted to happen only between product styles. Note that the style subsumes all the information about an apparel product, including its brand, design, and color, except its size. Put differently; the style includes all the fashion characteristics of the product. The aggregation approach is particularly sensible when considering a utility-based demand model like the multinomial choice model (MNL) in which a product’s utility is directly linked to its fashion design, not its size.

However, such size-aggregation approach can easily overlook product substitutions that are induced by the unavailability of certain sizes. It has been shown that the unavailability of sizes can cause the broken assortment effect (Smith and Achabal 1998, Caro and Gallien 2010, 2012), which refers to the empirical observation that a product’s sales rate decreases when the total inventory drops below a certain threshold, possibly because some sizes become unavailable. Furthermore, research in economics, marketing, and operations management has shown that failing to account for stockouts biases demand estimation (Campo et al. 2000, Che et al. 2012, Deng et al. 2022) and negatively impacts profitability (Musalem et al. 2010).

Most importantly, demand substitution can happen between sizes. When the desired product is out of stock, customers may consider products of adjacent sizes with the same fashion style, which we will call *size substitution* from here onward. Using a difference-in-differences approach and a dataset from one of the largest sports footwear retailers in China, Li et al. (2023) empirically show that 28.6% of the unmet demand of an out-of-stock footwear product spills over to the adjacent sizes of the same style. In principle, a demand model estimated by the size-aggregation approach would ignore size substitution and would not be able to assess how it affects stores’ profit and operational performance.

Given that product sizes play a vital role in apparel retail operations and size substitution has been observed in consumer choices, we have the following research question: *when does size substitution matter and when can it be put aside?* To illustrate this, imagine a retailer managing footwear inventory. If most customers who wear size 9.5 are willing to accept a half-size up or down, is it necessary to stock that size at all, or should the retailer allocate inventory to adjacent sizes instead, anticipating substitution? More broadly, how does size-based demand substitution, alongside the more commonly studied style-based substitution, influence downstream operational strategies? To address these questions, we take a prescriptive approach: we first propose a choice model, estimate it using real-world data, and analyze its implications for assortment and inventory optimization. Specifically, the paper makes the following contributions:

1. **A New Choice Model (Section 3):** We propose a novel choice model, called the *style-size* model, to model consumers’ decision-making process in purchasing apparel products. In this choice model, each customer is characterized by a tuple (s, σ, α) , where s is the customer’s

best-fit size, $\sigma \in \{+, -\}$ implies either the larger or the smaller adjacent size is the customer's second best-fit size, and α captures the customer's sensitivity to the lack of fit, i.e., the disutility for wearing a shoe in an adjacent size that does not fit perfectly. When facing a set of products, the customer (s, σ, α) first forms a *stock-induced consideration set* based on the products available in the best-fit size s ; if the best-fit size is unavailable, the customer would consider the adjacent size of the same style but penalizes them with a utility discount α . The customer then follows a multinomial logit (MNL) model to select a product from the consideration set.

2. **Model Estimation (Sections 4 and Appendix A):** We develop an expectation-maximization (EM) algorithm for model estimation, which iteratively solves a concave maximization problem while improving the likelihood at each step. Due to space constraints, we defer the details of the proposed EM algorithm to Appendix A. Using a dataset from a large footwear retailer, we estimate the style-size choice model and demonstrate that at least 24.9% of unmet demand due to stockouts spills over to adjacent sizes of the same style. Furthermore, we show that the proposed style-size choice model has strong representational power and outperforms benchmark models in out-of-sample prediction accuracy.
3. **Assortment and Inventory Optimization (Section 5):** We consider the assortment and inventory optimization problems under the proposed style-size choice model. We first show that the optimal assortment is invariant to customers' size sensitivity. That is, the optimal assortment is the same regardless of whether customers are likely to switch to adjacent sizes or less likely to do so. We then discuss the inventory optimization problem in which stockouts can trigger size substitution. Building on our result on the optimal assortment, we show that the size substitution effect is negligible when the planning horizon is long or customer demand is high, i.e., in the asymptotic regime. For the non-asymptotic regime, we first show that size substitution can affect profits and should be taken into account in stocking decisions. Then, we propose a mixed-integer program that captures the size substitution effect. In a numerical study we show that this policy performs well in the non-asymptotic regime, and subsequently we also prove that it is asymptotically optimal. All in all, our results provide guidance on when size substitution matters and when it does not.

Finally, we review the related literature in Section 2 and conclude the paper in Section 6. We relegate all proofs and additional numeric results to the appendix.

2. Literature Review

Fashion retail has attracted significant attention and contributions from the operations management community. For example, Caro and Gallien (2010) collaborate with one of the world's largest

fashion retailers, Zara, and show that their proposed network sales optimization model increases sales by 3-4% based on controlled experiments. Caro and Gallien (2012) also design a pricing process that increases Zara’s clearance revenue by 6%. The quick-response nature of fast fashion like Zara and H&M creates significant value as it can capture the latest trends and minimize production lead time (Cachon and Swinney 2011, Caro and Martínez-de Albéniz 2015).

Early works in apparel retail operations often ignore demand substitution, typically using single-product models. However, economics and market science have shown that demand substitution exists in consumer choice. A range of choice models has been developed to estimate demand substitution from data (Train 2009) and analyze its impact on operational decisions (Kök and Fisher 2007). Stockouts also influence demand, as customers may consider alternative products when their desired item is unavailable. Researchers in operations management and marketing science have proposed methodologies to estimate the impact of stockouts and show that ignoring them may lead to a biased estimation of product demand (Campo et al. 2000, Musalem et al. 2010, Che et al. 2012, Deng et al. 2022). Musalem et al. (2010) further propose a price promotion policy that can mitigate the negative economic impact of stockouts. Our model aligns with this research by examining stockout-driven size substitution in apparel products.

There is a growing interest in making effective inventory decisions under the stockout events. The seminal work of Mahajan and Van Ryzin (2001) first shows that the stockout-based inventory optimization problem, or *dynamic* inventory problem, is computationally challenging and proves that the revenue function is not even quasi-concave. Honhon et al. (2010), Honhon and Seshadri (2013) approximate the dynamic inventory problem with a continuous relaxation, discretize the time intervals according to the assortment change, and solve the inventory problem using a dynamic program, assuming that customers follow a ranking-based choice model to make decisions. Goyal et al. (2016) propose an FPTAS approximation under the assumption that the choice model only consists of nested rankings. Aouad et al. (2018) proposes an approximation algorithm with ratio 0.139 for the capacitated MNL inventory problem. Lee et al. (2016) discuss the stockout-based substitution and the inventory problem in the context of textbook retailing. Martínez-de Albéniz and Kunnumkal (2022) use a Markov chain to approximate the inventory problem under a fixed replenishment policy. Our work is related to a recent work by Liang et al. (2021), which considers an MNL-based demand and shows that the optimal inventory policy follows a gain-ordered structure under the fluid approximation (FA) of the dynamic problem. They prove that the rounded solution from the FA is asymptotically optimal in the exact problem with a nearly square-root convergence rate. For the MNL demand, a more recent work by Zhang et al. (2024) further improves the optimality gap by dropping the dependency on the number of products. Zhang et al. (2024) also provide an optimality gap for the FA under general substitutive choice models.

In apparel product demand substitution, most literature only focuses on the substitution between product styles and views a “product” as the aggregate of all sizes (Boada-Collado and Martínez-de Albéniz 2020). Such an approach ignores the stockouts induced by the missing sizes and overlooks the broken assortment effect (Smith and Achabal 1998, Caro and Gallien 2010, 2012). In a different direction, using a real-world footwear dataset and a difference-in-difference estimation procedure, Li et al. (2023) empirically show that a significant fraction of customers may consider adjacent sizes when the best-fit size is out of stock; see the discussion in Section 4.4. That is to say, demand substitution exists even between sizes. While it is clear that product size plays an essential role in fashion retailing, to the best of our knowledge, very few works have discussed the validity of the standard aggregation approach and addressed the difficulties of the operations when stockout-based size substitution happens. Our work aims to fill this gap in the literature.

3. Model

In this section, we propose a choice model that characterizes consumers’ apparel product choice based a two-step decision-making process.

3.1. Product, Style, Size, and Customers

We define an apparel product as a style-size pair. In particular, let $\mathcal{J}$ be the set of product styles and $\mathcal{K}$ be the set of product sizes. We consider a style-size pair (j, k) as an apparel product, where $j \in \mathcal{J}$ and $k \in \mathcal{K}$. The style contains all product information, including brand, design, and color, except its size. Put differently, if one views an SKU as a product, “style” summarizes all information of the SKU except the size. Notice that product sizes form a complete order, as we can always sort sizes in $\mathcal{K}$ as an increasing sequence. In addition, for a given size $k \in \mathcal{K}$, we use $\text{ADJ}_+(k)$ and $\text{ADJ}_-(k)$ to denote the larger and small-adjacent sizes of k , respectively. For example, consider a footwear universe of two styles $\mathcal{J} = \{\text{Nike Air Max White}, \text{Nike Air Force White}\}$ and nine sizes $\mathcal{K} = \{6, 6.5, 7, 7.5, \dots, 9.5, 10\}$. Then in this universe, there are $|\mathcal{J}| \times |\mathcal{K}| = 18$ products. The adjacent sizes follow immediately, and we have $\text{ADJ}_+(7) = 7.5$ and $\text{ADJ}_-(7) = 6.5$. Note that each middle size in $\mathcal{K}$ can have two adjacent sizes while the two boundary sizes can have only one adjacent size. To ease notation, we define $\mathcal{N} \equiv \{(j, k) \mid j \in \mathcal{J}, k \in \mathcal{K}\}$ as the set of products in the product universe. We also define $(0, 0)$ as the no-purchase option and $\mathcal{N}_+ = \mathcal{N} \cup \{(0, 0)\}$.

We assume that each customer can be depicted by parameter (s, σ, α) , where $s \in \mathcal{K}$ represents the customers’ best-fit size in the size set $\mathcal{K}$, $\sigma \in \{+, -\}$ implies either the larger (+) or the smaller-adjacent size (-) of s is the customer’s second best-fit size, and $\alpha \geq 0$ characterizes her sensitivity toward size deviation. We further focus on the customer type $(s, +, \alpha)$ now, and type $(s, -, \alpha)$ will

follow as a symmetric case. A customer of type $\tau = (s, +, \alpha)$ has random utility u_{jk}^τ toward product $(j, k) \in \mathcal{N}$:

$$u_{jk}^\tau = v_{jk}^\tau + \epsilon_{jk}^\tau,$$

where ϵ_{jk}^τ follows an independent standard Gumbel distribution and the expected utility v_{jk}^τ is defined as

$$v_{jk}^\tau = \begin{cases} v_j, & \text{if } k = s, \\ v_j - \alpha, & \text{if } k = \text{ADJ}_+(s), \\ -\infty, & \text{otherwise.} \end{cases} \quad (1)$$

That is, if a product of style j is of the best-fit size to the customer $\tau = (s, +, \alpha)$, it has expected utility v_j ; if the product is of the larger-adjacent size, the product is still “acceptable” to the customer but has a discount α in its utility; if a product is of neither the best-fit size or a larger-adjacent size, it won’t be considered at all by the customer and has $-\infty$ utility. Following the convention, the no-purchase option has random utility ϵ_{00}^τ and its expected utility is zero.

We remark that the random utility u_{jk}^τ of product (j, k) for customer type $\tau = (s, -, \alpha)$ follows the same argument except that the corresponding expected utility $v_{jk}^\tau = v_j - \alpha$ when $k = \text{ADJ}_-(s)$. It implies that to customer $(s, -, \alpha)$, a product of the smaller adjacent size $\text{ADJ}_-(s)$ is still acceptable but has a discount α in its utility.

Finally, we remark that the expected utility defined in Equation (1) indicate that customers will only consider apparel that is of their best-fit size or of the adjacent size. In other words, cross-size substitution is restricted to a customer’s best-fit size and its immediate neighbors. This assumption is consistent with Assumption 1 in Li et al. (2023). In Section 5, we further discuss how our main results on assortment and inventory planning remain valid even if this assumption is relaxed.

3.2. Consider and Choose Based on In-Stock Size Availability

We assume customers follow a two-step process to make the purchase decision. When facing a set of available products (an *assortment*) $A \subseteq \mathcal{N}$, customers first forms a consideration set and then choose a product from the consideration set or leave without a purchase. The notion of the consideration set here is quite different from the one in the literature (Aouad et al. 2021, Jagabathula et al. 2024, Akchen and Mitrofanov 2023). We will come back to make a comparison in Section 3.5.

Now, let’s consider a customer of type $\tau = (s, +, \alpha)$.

Consider. The customer τ forms a consideration set as follows. For a given style $j \in \mathcal{J}$, the customer τ first considers the best-fit product, which is (j, k) for $k = s$, and checks whether it is

available. If it is not available, then the customer will consider the same style but of the larger-adjacent size, i.e., (j, k) for $k = \text{ADJ}_+(s)$. More formally, let $C_\tau(A) \subseteq A$ be the consideration set of customer $\tau = (s, +, \alpha)$. Then it is the disjoint union of two sets $C_\tau(A) \equiv C_\tau^1(A) \cup C_\tau^2(A)$, where

$$C_\tau^1(A) = \{(j, k) \in A \mid k = s, j \in \mathcal{J}\} \quad (2)$$

$$C_\tau^2(A) = \{(j, k) \in A \mid k = \text{ADJ}_+(s), (j, s) \notin A, j \in \mathcal{J}\}. \quad (3)$$

Here C_τ^1 is the collection of products in assortment A that are of the customer's best-fit size s , and C_τ^2 is the collection of products in A that are of the larger-adjacent size $\text{ADJ}_+(s)$ given that the best-fit size s of the same style is not available. A key observation is that for the same style, **an adjacent size is considered only if the customers' best-fit size is not available**. That is, the customer will not consider an adjacent size of a style if she can find her best size of that style in the assortment. We demonstrate the formation of the consideration set $C_\tau(A)$ in the following example.

EXAMPLE 1. (Consideration Set) Assume that a store provides three styles of shoes, $\mathcal{J} = \{X, Y, Z\}$. A customer whose best-fit foot size is 7 visits the store and she might consider the larger-adjacent size 7.5 as well. In other words, she is of the customer type $\tau = (7, +, \alpha)$ for some utility discount $\alpha \geq 0$. In the store, some products are out of stock and the set of in-stock products is

$$A = \{(X, 6.5), (X, 7), (X, 7.5), (Y, 7.5), (Z, 6.0), (Z, 6.5)\}.$$

When seeing the assortment S , the customer forms consideration set is $C_\tau(A) = \{(X, 7), (Y, 7.5)\}$, since $C_\tau^1(A) = \{(X, 7)\}$ and $C_\tau^2(A) = \{(Y, 7.5)\}$. Note that product $(X, 7.5)$ will not be considered since the best-fit product $(X, 7)$ of style X is available. On the other hand, since the best-fit size for style Y is not available, the customer is willing to consider the larger-adjacent size 7.5, although it is assigned with a lower utility. The two available products of style Z will not be considered since they are too small. $\square$

We remark that under this definition of consideration sets, each customer forms a *strict* preference hierarchy over apparel sizes as *best-fit size* $\succ$ *adjacent size* $\succ$ *other sizes*. As a result, an adjacent size will only be considered if the best-fit size is not available. This interpretation aligns with the standard definition of preference ordering (Block and Marschak 1959, Farias et al. 2013, van Ryzin and Vulcano 2014). The model also allow non-deterministic best-fit size behavior by mixing customer types (Section 3.3), which captures the scenarios where adjacent sizes may occasionally become the perceived best-fit size due to inherent variability in consumer choice.

Choose. Once the customer forms the consideration set $C_\tau(A)$, she selects a product from $C_\tau(A)$ based on an MNL model with the utility defined in Equation (1), or she leaves without a purchase.

Specifically, the probability of choosing product (j, k) from the assortment A for a customer of type $\tau = (s, +, \alpha)$ is

$$\mathbb{P}_\tau((j, k) | A) = \begin{cases} \frac{e^{v_j}}{1 + \sum_{(j', k') \in C_\tau^1(A)} e^{v_{j'}} + \sum_{(j', k') \in C_\tau^2(A)} e^{v_{j'} - \alpha}}, & \text{if } (j, k) \in C_\tau^1(A), \\ \frac{e^{v_j - \alpha}}{1 + \sum_{(j', k') \in C_\tau^1(A)} e^{v_{j'}} + \sum_{(j', k') \in C_\tau^2(A)} e^{v_{j'} - \alpha}}, & \text{if } (j, k) \in C_\tau^2(A), \\ 0, & \text{otherwise,} \end{cases} \quad (4)$$

with the no-purchase probability $\mathbb{P}_\tau((0, 0) | A) = 1 / \left(1 + \sum_{(j', k') \in C_\tau^1(A)} e^{v_{j'}} + \sum_{(j', k') \in C_\tau^2(A)} e^{v_{j'} - \alpha}\right)$. Finally, we remark that the choice probability $\mathbb{P}_\tau((j, k) | A)$ for a customer of type $\tau = (s, -, \alpha)$ follows the same expression as Equation (4) except that the sets $C_\tau^1(A)$ and $C_\tau^2(A)$ are redefined accordingly. Specifically, $C_\tau^2(A)$ follows as $C_\tau^2(A) = \{(j, k) \in A \mid k = \text{ADJ}_-(s), (j, s) \notin A, j \in \mathcal{J}\}$.

3.3. The Style-Size Choice Model: The General and Average Cases

Let $\Gamma = \{(s, \sigma, \alpha) \mid s \in \mathcal{K}, \sigma \in \{+, -\}, \alpha \geq 0\}$ be the collection of all customer types. We further use μ_τ to represent the density of customer type $\tau \in \Gamma$ in the market. Along with the utility parameters v_j of styles $j \in \mathcal{J}$, we define a **style-size choice model** as

$$[\text{General Model}]: \quad \mathbb{P}((j, k) | A) = \sum_{s \in \mathcal{K}, \sigma \in \{+, -\}} \int_0^\infty \mathbb{P}_{(s, \sigma, \alpha)}((j, k) | A) \cdot \mu_{(s, \sigma, \alpha)} d\alpha, \quad (5)$$

where the choice probability $\mathbb{P}_\tau((j, k) | A)$ is defined as in Equation (4).

In Expression (5), we seek a general representation of customers' experience on product sizes. In particular, the distribution μ_τ for $\tau = (s, \sigma, \alpha) \in \Gamma$ allows us to model a wide range of consumer decisions in the context of apparel product sizes. Let us use customer type $(s, +, \alpha)$ and men's footwear industry as an example. The range of shoe sizes is usually $\{7, 7.5, 8, 8.5, \dots, 12.5, 13\}$. On the other hand, customers' actual foot sizes are *continuously* distributed in the range between, let's say, 25 cm (corresponding to size 7) and 30 cm (corresponding to size 13). A customer whose foot size is exactly 27.5 cm (size 10) might feel uncomfortable when trying size 10.5, which can be too loose. In that case, the corresponding α is bigger. On the contrary, consider a customer whose best-fit size is 10 and actual foot size is slightly longer than 27.5 cm. When size 10 is out of stock, he is more flexible in choosing the adjacent size, 10.5. In that case, the corresponding utility discount α is smaller. The distribution of customer types over α reflects the fact that the standardized retail sizes are an approximation to each person's real foot size (or body size for clothes).

Later in Section 4, when we estimate the size substitution effect from a real-world dataset that involves the inventory information for nearly five hundreds apparel products over an eight-month horizon, we consider an *average* case of the style-size choice model (5). In this average model, we aim

to obtain a more succinct and interpretable representation of the general model (5). Specifically, we first use one parameter to represent the discomfort discount $\alpha = \alpha_0$ of all customers, which will be estimated from the dataset. This also helps us interpret the average cross-size demand spillover in this dataset and compare it with the empirical results by Li et al. (2023). Second, we assume that for each best-fit size $s \in \mathcal{K}$, customers are equal likely to be oversized (thus might consider the larger-adjacent size) or undersized (thus might consider the smaller-adjacent size) compared to s . That is, we assume $\mu_{(s,+, \alpha_0)} = \mu_{(s, -, \alpha_0)}$. With these reductions, we obtain a more compact style-size choice model.

$$\text{[Average Model]: } \mathbb{P}((j, k) \mid A; \alpha_0) = \sum_{s \in \mathcal{K}} \bar{\mu}_s \cdot \left(\frac{1}{2} \cdot \mathbb{P}_{(s,+, \alpha_0)}((j, k) \mid A) + \frac{1}{2} \cdot \mathbb{P}_{(s, -, \alpha_0)}((j, k) \mid A) \right), \quad (6)$$

where, with slight abuse of notation, we write $\bar{\mu}_s \equiv \mu_{(s,+, \alpha_0)} + \mu_{(s, -, \alpha_0)}$. In this average model, the parameter $\bar{\mu}_s$ represents the fraction of customers whose best-fit size is s . We remark that the average model (6) can be fully characterized by parameters α_0 and $(v_j, \bar{\mu}_s)_{j \in \mathcal{J}, s \in \mathcal{K}}$, leading to a total number of $|\mathcal{J}| + |\mathcal{K}| + 1$ parameters. This tight representation allows us to compare the performance of the style-size choice model with other benchmark approaches, which involve numbers of parameters of a similar scale; see Section 4.3.

3.4. Model Extension: Size Variation across Styles

Due to the diverse combinations of apparel styles, sizes, and customers' actual body measurements, it's unlikely that all consumer choices in apparel retail can be fully captured by the general style-size model (5). For instance, a *baggy fit* T-shirt is intentionally designed to be looser. A customer who normally wears size L in other styles might find that size M offers the best fit in this case. When size definitions for a particular style do not align with others, we can relabel sizes within $\mathcal{K}$ for that style to maintain consistency. These adjustments can be easily implemented during inventory management.

In more extreme cases where substantial size variation exists across apparel styles, we might define each customer type as a tuple $(\mathbf{s}, \boldsymbol{\sigma}, \boldsymbol{\alpha}) = (s_j, \sigma_j, \alpha_j)_{j \in \mathcal{J}}$, where for each style $j \in \mathcal{J}$, size $s_j \in \mathcal{K}$ is the customer's best-fit size, $\sigma_j \in \{+, -\}$ indicates which an adjacent size would be considered, and α_j represents the utility discount associated with choosing that adjacent size. The customer type (s, σ, α) defined in Section 3.1 is a special case of this tuple, where s_j , σ_j , and α_j are fixed across all styles $j \in \mathcal{J}$. While we will not address this extension in depth due to the added notational complexity, we will later show in Section 5 that our main results on assortment and inventory planning – Theorem 1 and Proposition 1 – still apply under this extended model.

3.5. Comparison to Other Choice Models in the Literature

Now we compare the style-size choice model in Expression (5) with other existing choice models in the literature. At first glance, the style-size choice model resembles the mixed-MNL model (Train 2009), which assumes that there are several customer types in the market and each customer type makes decisions according to a distinct MNL model. The style-size choice model also allows customer heterogeneity in Expression (5), but it differentiates itself from the mixed-MNL model by incorporating the notion of consideration set in the decision-making process. This consideration set structure allows us to model the strict hierarchy between sizes given that there exists the most suitable size, the adjacent size, and the unacceptable sizes for each customer. In contrast, in the mixed-MNL model, it is not possible to construct a hierarchy between sizes as long as each has non-zero choice probability, and a customer may still buy a much larger or a much smaller size of a given style even if the best-fit size is offered.

The style-size choice model contributes to the growing literature of the consideration set-based choice models. In particular, Aouad et al. (2021) and Jagabathula et al. (2024) consider a consider-then-choose (CTC) model, which is defined as a distribution over the product space of subsets and rankings. In the CTC model, a customer type is characterized by a subset-ranking pair (C, σ) . When an assortment A is offered, a customer of type (C, σ) will choose $\arg \min_{i \in C \cap A} [\sigma(i)]$, i.e., choose the product with the highest rank in the intersection of the consideration set C and the offered assortment A . Our style-size choice model differs from the CTC model in several aspects. First, in the “choose” step, our model follows an MNL model while the CTC model follows a ranking preference. Second, the consideration set in the style-size choice model is *stock-based*, i.e., a function of stock, while the consideration set in the CTC model is *independent* of the set of available products. In the following example, we show that the consideration set $C_\tau(A)$ defined in Section 3.2 cannot be represented by the CTC model as an intersection of assortment A and a fixed subset C of products.

EXAMPLE 2. (Stock-based Consideration Set) Consider a universe of one style $\mathcal{J} = \{X\}$ and two sizes $\mathcal{K} = \{7, 7.5\}$, where the two sizes are adjacent to each other. We also consider a customer of type $\tau = (7, +, \alpha)$. For $A_1 = \{(X, 7), (X, 7.5)\}$, we have $C_\tau(A_1) = \{(X, 7)\}$; for $A_2 = \{(X, 7)\}$, we have $C_\tau(A_2) = \{(X, 7)\}$; for $A_3 = \{(X, 7.5)\}$, we have $C_\tau(A_3) = \{(X, 7.5)\}$. Assume that $C_\tau(A)$ can be presented as $C \cap A$ for a subset $C \subseteq \mathcal{N}$. By $C_\tau(A_2) = \{(X, 7)\}$, we know that $(X, 7) \in C$. By $C_\tau(A_3) = \{(X, 7.5)\}$, we know that $(X, 7.5) \in C$. However, this implies that $C \cap A_1 = \{(X, 7), (X, 7.5)\} \neq C_\tau(A_1)$, a contradiction. $\square$

The style-size choice model is analogous to a context-dependent (i.e., assortment-dependent) choice (Tversky and Simonson 1993) in which customers make decision based on the context of

products and their comparisons to each other in the offered assortment. In the style-size choice model, a customer sees the set of available products and decides not to consider the adjacent sizes if the best-fit size of the same style is available in the assortment. One can also view the style-size choice model as cue-triggered consumer behavior (Pennesi 2021) in which consumers’ decisions are driven by a stimulus from the environment. In the style-size choice model, the unavailability of the best-fit size in the assortment triggers customers to consider the adjacent sizes of the same style.

Finally, while the style-size choice model is analogous to the context-dependent choice models, it still satisfies the *substitutability* property (or also called the *stochastic rationalizability* property; see Jagabathula and Rusmevichientong (2019), Chen and Mišić (2022), Zhang et al. (2024)). The property is a widely used axiom in the economics and decision theory literature (Rieskamp et al. 2006). It is satisfied by several popular choice models, including the mixed-MNL and ranking-based models, and defined as follows.

DEFINITION 1. A choice model $\mathbb{P}$ over choices in $\mathcal{N}_+$ satisfies the *substitutability* property if $\mathbb{P}(m \mid A \cup \{n\}) \leq \mathbb{P}(m \mid A)$ for all assortments A and choices m and n such that $n \in \mathcal{N} \setminus A$.

The property implies that the probability of choosing any product will not increase if we enlarge an assortment. The substitutability property is refereed as the least restricted form of rational choice and thus sometimes dubbed as “weak rationality.” However, it can still be violated if a choice model involves context-dependent or assortment-dependent phenomena. One example is the decoy effect, a marketing phenomenon where adding an inferior “decoy” product to an assortment increases the appeal of a superior “target” product, making consumers more likely to choose it (Huber et al. 1982). When a choice model violates the substitutability property, it usually leads computationally expensive methodologies for the downstream applications (Akchen and Mišić 2021). Although the style-size choice model can be considered context-dependent due to the size-triggered consideration set, it still satisfies the substitutability property as long as the utility discount is nonnegative. We have the following lemma.

LEMMA 1. The choice probability $\mathbb{P}_{(s,\sigma,\alpha)}$ satisfies the substitutability property if and only if $\alpha \geq 0$.

We relegate the proof to Appendix B.1. Later in Section 5.4, we will see that Lemma 1 leads to an intuitive inventory policy that is asymptotically optimal.

4. The Dataset and The Estimation Outcomes

In this section, we apply our model to real-world inventory and sales data and estimate how customers may consider adjacent sizes when stockouts happen.

	Mean	Median	Std. Dev.	Min	Max
Visitors	3,857.5	3,981.5	593.6	2,865.6	4,880.7
Units stocked	453.8	449.1	115.8	234.7	604.4
Products available	271.2	284.9	63.5	143.0	359.2
Units sold	30.8	30.5	18.5	6.5	77.3
Sizes offered	6.4	6.4	1.5	3.9	8.4

Table 1 Weekly summary statistics (averaged across stores)

4.1. Data

We obtained the dataset from a large footwear retailer. The company operates hundreds of stores and also owns an e-commerce website. We focus on the data collected from brick-and-mortar stores. Notice that the style-size combination is the most disaggregate product level observed in the dataset. We again follow Section 3 to define such style-size combination as a product or SKU.

The data spans 33 weeks in the 2019-2020 season, starting on July 28, 2019, and is for 51 styles of woman casual booties, which is a midsize category among 50+ categories overall. There are nine shoe sizes ranging from size 6 to size 10 with half sizes in between. The dataset includes the following information from each store $m \in \mathcal{M}$ and week $t \in \mathcal{T}$:

N^{mt} : the number of visitors to store m during week t . The data were collected by the traffic counter at the entrance of each store. On average, there were around four thousand visitors to each store in each week.

$Q_{(j,k)}^{mt}$: the number of sold units of product (j,k) at store m during week t . On average, 30.8 units were sold at a given store in a week. Hence, roughly 99% of the customers either bought a product outside $\mathcal{N}$ or did not make a purchase.

$I_{(j,k)}^{mt}$: the number of stocked units of product (j,k) for store m and week t . We remark that we also know the replenished units. We find that the values of the stocked units, the units sold, and replenished units are quite consistent, and it shows that the inventory records are reliable. On average, a store stocked 453.8 units during a week.

A^{mt} : the set of available products for store m and week t . Recall that A^{mt} is defined as $A^{mt} = \{(j,k) \in \mathcal{N} \mid I_{(j,k)}^{mt} \geq 1\}$. For simplicity, we assume that A^{mt} is the same throughout the week, i.e., each customer who visited the store during the week saw the same set of available products. This is a reasonable assumption, as we observe that only a small fraction of products were sold in a week and thus the set of available products A^{mt} would not change significantly during the same week. On average, there were 271.2 products available, out of a total of 459 ($= 51 \times 9$). For the styles offered, on average there were 6.4 sizes in stock (out of 9).

We summarize the aforementioned information in Table 1. Specifically, we report the weekly visitors N^{mt} , units stocked $\sum_{(j,k) \in \mathcal{N}} I_{(j,k)}^{mt}$, products available $|A^{mt}|$, units sold $\sum_{(j,k) \in \mathcal{N}} Q_{(j,k)}^{mt}$, and

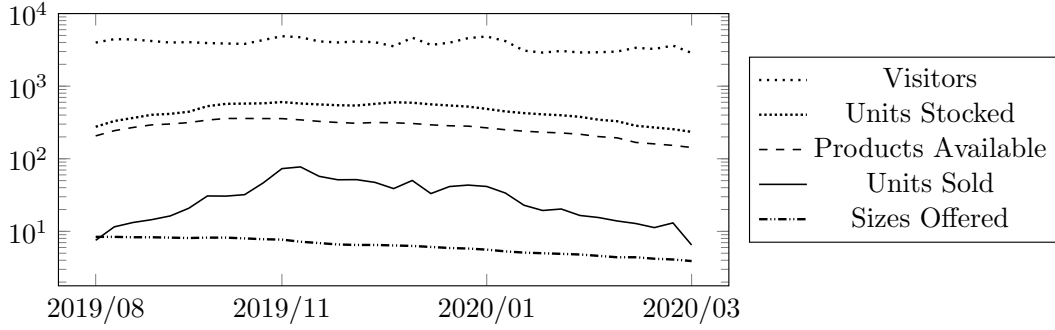


Figure 1 Evolution of visitors, units stocked, products available, units sold, and sizes offered from 2019 Fall to 2020 Spring averaged across stores

sizes offered, all averaged across stores $m \in \mathcal{M}$.¹ We also show the evolution of these quantities in Figure 1. From the figure it can be seen that the number of visitors decreased slowly in the period considered. Similarly, the number of sizes offered decreased almost monotonically from 8.4 sizes (out of 9) to 3.9. In contrast, the stocked units and the number of available products peaked in mid-October 2019, and the number of sold units peaked in November 2019, a few weeks behind the peak of the stocked units and right before the holiday season.

4.2. Estimation Method: The EM Algorithm

We propose an estimation method for the average style-size choice model (6) based on the expectation-maximization (EM) algorithm. Given space constraints, we defer the technical details to Appendix A and present a high-level summary below.

The EM algorithm is a widely used framework for maximum likelihood estimation in models with latent variables. It alternates between two steps: an expectation (E) step, where the expected values of the missing or unobserved variables are computed given the observed data and current parameter estimates, and a maximization (M) step, where these expectations are used to optimize the model parameters. In our setting, customer types τ are unobserved in data, making them natural latent variables for an EM approach. In the E step, we compute the conditional expectation of customer-type assignments using Bayes' rule, based on the current model parameters and the observed sales data $(N^{mt}, \{Q_{(j,k)}^{mt}\}_{(j,k) \in A^{mt}})_{m \in \mathcal{M}, t \in \mathcal{T}}$. In the M step, we maximize the expected complete-data log-likelihood with respect to model parameters. This step further decomposes into two independent optimization problems under the style-size choice model: one for estimating the distribution over customer types, which has a closed-form solution, and the other for estimating style utilities and size sensitivity, which involves a concave maximization problem that can be

¹ The sizes offered equals the ratio between $|A^{mt}|$ and the number of styles in the assortment.

solved efficiently. In Appendix A, we derive the complete-data log-likelihood based on the style-size choice model (Section A.1) and then develop the E and M steps in detail (Section A.2).

The EM algorithm has been used for estimating choice models from data. Examples include the estimation of the LC-MNL model (Train 2009), the general attraction model (Gallego et al. 2015), and the ranking-based model (van Ryzin and Vulcano 2014). Our EM approach provides both flexibility and computational efficiency while allowing for the incorporation of seasonality effects. By introducing time-fixed effects in the utility function, it accounts for variations in consumer preferences across different sales periods. This is particularly important given that our dataset spans thirty-three weeks, as illustrated in Figure 1. In Appendix A.3, we further highlight the advantages of our EM algorithm by comparing its simplicity and scalability to the estimation procedures of other choice models.

4.3. Estimation Outcomes

We present our estimation outcomes in Table 2, which compares the performance of four models, the size aggregation model, the nested logit model, the granular model, and the style-size model, under three metrics. The style-size model is the proposed model in this paper. As discussed at the beginning of Section 3.3, due to the large size of the dataset, we consider estimating the average style-size choice model (6).

The first benchmark, the size aggregation model (Size-Agg), refers to the traditional approach described in the introduction (Section 1). Specifically, in this approach, one aggregates all sizes (all SKUs) under the same style to create a “product” and only considers it out of stock if all sizes are not available. Following this approach, we estimate the utility v_j^{agg} of each style by first creating the aggregated products from the data and then estimating $(v_j^{\text{agg}})_{j \in \mathcal{J}}$ via maximum likelihood estimation. The choice probability for an apparel product (j, k) in the assortment A under the size aggregation model is simply $\mathbb{P}((j, k) | A) = \exp(v_j^{\text{agg}}) / \left(1 + \sum_{j' | \exists (j', k') \in A} \exp(v_{j'}^{\text{agg}})\right) \cdot \hat{\mu}_k$, where $\hat{\mu}_k$ is the fraction of sales of size k . In other words, under the size aggregation model, we assume that the demand of (j, k) is simply the demand of the style j times the fraction of demand of size k .

The second benchmark is the nested logit model (Train 2009), which has a natural structure that incorporates the apparel styles and sizes. Specifically, we consider a two-level nested logit model, where the first and second levels encode the apparel sizes and styles, respectively. For simplicity, we only present one variant of the nested logit model. Another variant, in which styles are encoded first, is discussed in Appendix C, with Figure 4 illustrating both variants. In the same section, we discuss how the style-size choice model proposed in this paper differ from these two variants of the nested logit model. Note that the two variants have similar performance in terms of out-of-sample prediction in our numeric setting.

Model	Size-Agg	Nested Logit	Style-Size	Granular
Number of Parameters	60	61	61	459
Size Sensitivity (α_0)	-	-	1.39***	-
KL Divergence (10^{-2})	1.88	1.72	1.66	1.67
Mean Absolute Error (10^{-3})	2.64	2.55	2.50	2.54
KL on No-Purchase (10^{-4})	8.25	6.97	6.87	6.96

*** Significant at the 0.1% level

Table 2 Estimation Results for the footwear products in the dataset

Lastly, we refer to the granular model as the MNL model at the most disaggregate level. In particular, the model assumes each product (j, k) has expected utility v_{jk} and customers make purchase decisions according to the MNL model $\mathbb{P}((j, k) | S) = \exp(v_{jk}) / (1 + \sum_{(j, k) \in S} \exp(v_{jk}))$. We call it the granular model because it assigns model parameters at the most disaggregate/granular level, i.e., assigns a parameter to each style-size pair. Notice that the granular model has $|\mathcal{J}||\mathcal{K}| = 459$ parameters, while the style-size model with average size sensitivity parameter considered in this section only has $|\mathcal{J}| + |\mathcal{K}| + 1 = 61$ parameters. Similarly, the nested logit model has $|\mathcal{J}| + |\mathcal{K}| + 1 = 61$ parameters while the size aggregation model has $|\mathcal{J}| + |\mathcal{K}| = 60$. Therefore, among all the models we consider in this numerical study, the granular model has the largest number of parameters. If the problem instance grows bigger, the granular model can be more disadvantage for practitioners for the purpose of interpreting consumer choice and designing business strategies.

In a sense, the granular model is neither practical nor compact, as it assumes that customers may substitute shoes of a very large size for shoes of a small size. To this end, we view the granular model as the benchmark that captures the consumer choice at the most disaggregate level. While other stronger choice models exist, such as the LC-MNL model, the number of parameters in those models would further increase and make the comparison with the style-size choice model less informative. For example, a ten-class LC-MNL model would have 4590 parameters in contrast to 61 in the style-size choice model with average size sensitivity. When the model complexities differ up to eighty times, one can expect that the more complex model can fit the data better while being intractable and harder to implement in practice and with the risk of overfitting. In fact, in our experience, it is computationally intractable to estimate the LC-MNL model for the current dataset.

We present the estimation outcome of each model in Table 2. The first row presents the number of parameters in each model. The second row reports the estimated average size sensitivity parameter is $\alpha_0 = 1.39$. The estimation outcome also passes the likelihood ratio test with very small p -value against the style-size choice model of zero size substitution effect. We will come back to provide insights on this value of size sensitivity parameter and connect it to the spillover effect reported by Li et al. (2023).

The third to last row report the predictive performance of each model, under three different metrics. Specifically, here we consider the out-of-sample performance. For simplicity, in each trial of experiment, we uniformly at random assign each store to be either in the training group $\mathcal{M}^{\text{train}}$ or in the testing group $\mathcal{M}^{\text{test}}$. We then use sales data from the stores in the training group $\mathcal{M}^{\text{train}}$ to learn the choice models, and examine the performance of each model based on the sales data from the testing group $\mathcal{M}^{\text{test}}$. We run the experiment forty times and report the average performance. We use three different metrics to measure the performance of each model. The first two metrics, the KL divergence and the Mean Absolute Error (MAE), are the standard metrics used in the literature. We define them as follows. Let $\tilde{p}_{(j,k)}^{mt} = \mathbb{P}((j,k) | A^{mt})$ and $\hat{p}_{(j,k)}^{mt} = Q_{(j,k)}^{mt}/N^{mt}$ be the predicted and empirical choice probability of product (j,k) in week t at store m . We write $A_+ \equiv A \cup \{0,0\}$ for any assortment A . The KL divergence is defined as

$$\text{KL} = - \left(\sum_{m \in \mathcal{M}^{\text{test}}} \sum_{t \in \mathcal{T}} N^{mt} \sum_{(j,k) \in A_+^{mt}} \hat{p}_{(j,k)}^{mt} \cdot \log(\tilde{p}_{(j,k)}^{mt} / \hat{p}_{(j,k)}^{mt}) \right) / \left(\sum_{m \in \mathcal{M}^{\text{test}}} \sum_{t \in \mathcal{T}} N^{mt} \right). \quad (7)$$

We further let $\tilde{Q}_{(j,k)}^{mt}$ be the predicted sales of product (j,k) in week t at store m . Then the MAE is defined and rewritten as

$$\text{MAE} = \frac{\sum_{m \in \mathcal{M}^{\text{test}}} \sum_{t \in \mathcal{T}} \sum_{(j,k) \in A_+^{mt}} |\tilde{Q}_{(j,k)}^{mt} - Q_{(j,k)}^{mt}|}{\sum_{m \in \mathcal{M}^{\text{test}}} \sum_{t \in \mathcal{T}} N^{mt}} = \frac{\sum_{m \in \mathcal{M}^{\text{test}}} \sum_{t \in \mathcal{T}} N^{mt} \sum_{(j,k) \in A_+^{mt}} |\tilde{p}_{(j,k)}^{mt} - \hat{p}_{(j,k)}^{mt}|}{\sum_{m \in \mathcal{M}^{\text{test}}} \sum_{t \in \mathcal{T}} N^{mt}}$$

For both metrics, a smaller value implies better predictive performance.

Table 2 shows that the performance of the size aggregation model is significantly worse than that of other models. In particular, since the model overlooks the broken assortment effect caused by the size stockouts, it *underestimates* the style utility - when a customer couldn't find her best size of a style, the model would misinterpret that as the style is not "attractive" enough and thus undervalue it. This numerical finding highlights the peril of aggregating sizes in demand estimation, especially in a setting shown in Table 1 where sizes are not always complete.

According to Table 5, among the three remaining models, the proposed style-size model has the best performance. Notably, it outperforms the nested logit model, which has the same number of parameters, in both metrics. When compared to the granular model, which has nearly eight times more parameters, the style-size model demonstrates a clear advantage in predictive performance measured by the MAE score. In terms of KL divergence, the style-size and granular models perform comparably. This is surprising, as we initially expected the granular model to perform better due to its higher number of parameters. To further investigate this result, we define a third metric "KL on No-Purchase" as

$$- \left(\sum_{m \in \mathcal{M}^{\text{test}}} \sum_{t \in \mathcal{T}} N^{mt} \left(\hat{p}_{(0,0)}^{mt} \cdot \log \left(\frac{\tilde{p}_{(0,0)}^{mt}}{\hat{p}_{(0,0)}^{mt}} \right) + (1 - \hat{p}_{(0,0)}^{mt}) \cdot \log \left(\frac{1 - \tilde{p}_{(0,0)}^{mt}}{1 - \hat{p}_{(0,0)}^{mt}} \right) \right) \right) / \left(\sum_{m \in \mathcal{M}^{\text{test}}} \sum_{t \in \mathcal{T}} N^{mt} \right),$$

which measures how accurately a choice model can predict whether a customer would make a purchase or not. Particularly, the “KL on No-purchase” measures the information loss over purchase/no-purchase decisions, $\hat{p}_{(0,0)}^{mt} \cdot \log \left(\tilde{p}_{(0,0)}^{mt} / \hat{p}_{(0,0)}^{mt} \right) + (1 - \hat{p}_{(0,0)}^{mt}) \cdot \log \left((1 - \tilde{p}_{(0,0)}^{mt}) / (1 - \hat{p}_{(0,0)}^{mt}) \right)$, instead of that over all purchase decisions (including no-purchase option) in A_+^{mt} , i.e., $\sum_{(j,k) \in A_+^{mt}} \hat{p}_{(j,k)}^{mt} \cdot \log \left(\tilde{p}_{(j,k)}^{mt} / \hat{p}_{(j,k)}^{mt} \right)$, compared to Equation (7).

In Table 5, we observe that the style-size model predicts more accurately than both the nested logit and granular models on whether customers make purchases. Additionally, while the granular model significantly outperforms the nested logit model in terms of KL divergence for all purchase decisions, this outperformance is not observed in the KL divergence for purchase/no-purchase decisions. This suggests that the additional parameters in the granular model improve its fit for consumer choices when purchases are made but do not effectively capture when and whether customers choose not to purchase. We attribute this to *model misspecification*. In both the granular and nested logit models, customers may substitute shoes of very distant sizes, leading to an underestimation of the no-purchase probability. In contrast, the style-size model assumes that customers only substitute adjacent sizes, resulting in a more accurate prediction of the no-purchase option.

We also remark that one can possibly design an even more advanced version of the style-size choice model, by allowing each apparel product (j, k) to have its utility parameter $v_{(j,k)}$, along with the structure of the consideration sets and customer types. We expect such model would possibly further improve the predictive accuracy, as it has the best of both worlds - the additional parameters can help to predict the apparel product demand better if the customer eventually make purchases, as in the granular model, while the structure of the consideration sets leads to a more reasonable way to model how customers faces size stockouts, as in the vanilla style-size choice model. However, we do not pursue such approach in this paper, as we do not intend to propose a new model that surpasses all choice models in the literature in terms of prediction accuracy. Instead, we propose a parsimonious model to study the demand substitution under the synergy of apparel styles and sizes and provide operational insights, as we will soon see in Section 5.

Lastly, Figure 2 presents the uncensored distribution $\bar{\mu}_k \equiv \mu_{(k,+, \alpha_0)} + \mu_{(k,-, \alpha_0)}$ of customers’ best-fit sizes (blue bars) in the estimated style-size choice model, and compares it with the censored distribution (yellow bars), which is the fraction of units sold in each size $\hat{\mu}_k \propto \sum_{mtj} Q_{(j,k)}^{mt}$. We observe that the distribution $\bar{\mu}_k$ is smoother than the fraction of sales $\hat{\mu}_k$. Indeed, the censored distribution $\hat{\mu}_k$ overestimates the probability mass of the “major” sizes in the middle, i.e., $k \in \{7.5, 8, 8.5\}$, at the expense of the less popular “minor” sizes at the extremes – namely, $k \in \{6, 6.5, 9.5, 10\}$. The stocking decisions censor the demand of the minor sizes, but it is reestablished by the EM algorithm. We also observe that both $\bar{\mu}_k$ and $\hat{\mu}_k$ are not unimodal over k . For example, $\bar{\mu}_{6.5}$ and $\bar{\mu}_{9.5}$ are slightly smaller than $\bar{\mu}_6$ and $\bar{\mu}_{10}$, respectively. This is a truncation effect since sizes 6 and 10

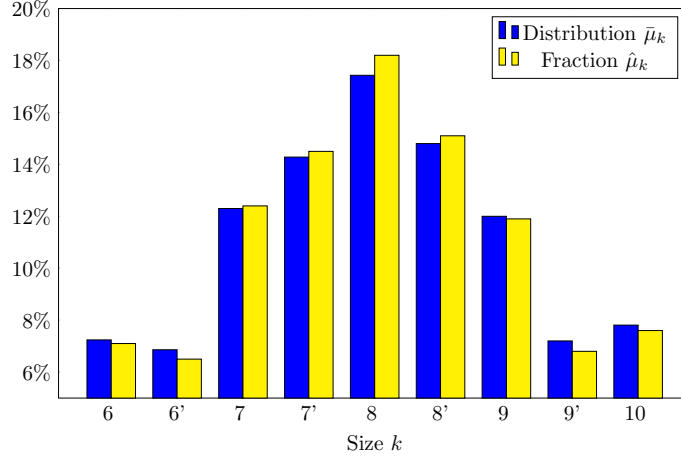


Figure 2 The uncensored best-fit distribution $\bar{\mu}_k$ and the observed fraction of sales $\hat{\mu}_k$.

receive spillover demand from consumers that have a shoe size slightly smaller than 6 or slightly larger than 10, respectively.

4.4. Discussion

In Table 2, we show that the estimated size sensitivity parameter is $\alpha_0 = 1.39$. In this section, we will further relate it to the spillover effect of unmet demand over the adjacent sizes reported by Li et al. (2023). Consider the style-size choice model we estimated from the dataset in Section 4.3 with average size sensitivity parameter α_0 and fix a style $j \in \mathcal{J}$. Assume a customer of type $\tau = (k, +, \alpha_0)$ visits a store. Let us define two assortments A_1 and A_2 , where $A_1 = A_0 \cup \{(j, k) \mid k \in \mathcal{K}\}$, $A_2 = A_1 \setminus (j, k)$, and A_0 is any assortment composed of products of styles other than j . We can interpret A_2 as the scenario where the best-fit size k of style j is out of stock. In the first assortment A_1 , the choice probability of product (j, k) is

$$\mathbb{P}_\tau((j, k) \mid A_1) = \frac{e^{v_j}}{1 + e^{v_j} + \sum_{(j', k') \in C_\tau^1(A_0)} e^{v_{j'}} + \sum_{(j', k') \in C_\tau^2(A_0)} e^{v_{j'} - \alpha_0}}$$

and the choice probability of (j, k') is simply zero for $k' = \text{ADJ}_+(k)$. In the second assortment A_2 , the choice probability of (j, k) is zero, as it is out of stock, and the choice probability of the adjacent-larger size (j, k') for $k' = \text{ADJ}_+(k)$ is

$$\mathbb{P}_\tau((j, k') \mid A_2) = \frac{e^{v_j - \alpha_0}}{1 + e^{v_j - \alpha_0} + \sum_{(j', k') \in C_\tau^1(A_0)} e^{v_{j'}} + \sum_{(j', k') \in C_\tau^2(A_0)} e^{v_{j'} - \alpha_0}}.$$

It follows that:

$$\frac{\mathbb{P}_\tau((j, k') \mid A_2)}{\mathbb{P}_\tau((j, k) \mid A_1)} \geq \exp(-\alpha_0) = 24.9\%. \quad (8)$$

The inequality in Equation (8) holds for any style j , any best-fit size k , and any customer type $\tau = (k, \sigma, \alpha_0)$ for $\sigma \in \{+, -\}$. Therefore, it implies that with probability at least 24.9% that a customer will switch to an adjacent size of the same style when the best-fit size is out of stock. If we adopt the classic interpretation of choice probability as the demand rate, Equation (8) suggests that, on average, at least 24.9% of the unmet demand for an apparel product due to stockouts may substitute to the adjacent sizes of the same style². By the symmetry in the average style-size choice model, this substitution is evenly split: approximately 12.5% spills over to the larger adjacent size, while the remaining 12.5% shifts to the smaller adjacent size.

Finally, we remark that the paper by Li et al. (2023) investigates similar consumer behavior in size substitution under stockouts. The authors work with one of China's largest sportswear fashion retailers and develop a difference-in-differences (DID) framework to estimate the stockout-based spillover effect on men's sports footwear over a two-year period. They show that, when adjacent sizes are always in stock, 25.1% and 26.6% of the unmet demand of an out-of-stock SKU spill to the adjacent-larger and the adjacent-smaller sizes, respectively. They term this the *theoretical* cross-size demand spillover. The authors also consider the *actual* cross-size spillover, in which they take into account the availability of adjacent sizes when each stockout happened (Section 6.3 of Li et al. (2023)), and show that 16.7% and 11.9% of the unmet demand of an out-of-stock SKU spill to the adjacent-larger and the adjacent-smaller sizes, respectively.

It is remarkable that the findings in Li et al. (2023), especially the actual cross-size demand spillover, are quite comparable to ours in the sense that both papers report an appreciable amount of unmet demand spilling to adjacent sizes due to stockouts, although the empirical approaches differ (DID vs. choice modeling), and the product categories are different (men's sports shoes vs. women casual booties). In addition, with the proposed prescriptive approach, we will be able to discuss how the size substitution effect may impact the operational decisions on assortment and inventory planning.

5. Assortment and Inventory Optimization

In this section, we investigate how the size substitution effect may impact the operational decisions in assortment and inventory optimization problems. All proofs are relegated to Appendix (Section B).

5.1. Assortment Optimization

We first consider the assortment optimization problem under the proposed style-size choice model. We assume that each product $(j, k) \in \mathcal{N}$ has an unit revenue r_j , i.e., the unit revenue is independent

² The ratio in Equation (8) is less than $e^{-\alpha_0} \cdot \max_j \{(1 + e^{v_j}) / (1 + e^{v_j - \alpha_0})\}$, which in our dataset is 24.92%.

of product size. This is a reasonable assumption, as stores usually do not charge different prices for products of the same style. Without loss of generality, we write that $\mathcal{J} \equiv \{1, 2, \dots, J\}$, and $r_1 \geq r_2 \geq \dots \geq r_J \geq 0$. Then the assortment optimization problem is defined as

$$\underset{A \subseteq \mathcal{N}}{\text{maximize}} \left\{ R(A) \equiv \sum_{(j,k) \in A} r_j \cdot \mathbb{P}((j,k) | A) = \sum_{s \in \mathcal{K}, \sigma \in \{+, -\}} \int_0^\infty \mu_{(s,\sigma,\alpha)} \cdot R_{(s,\sigma,\alpha)}(A) d\alpha \right\}, \quad (9)$$

where $R(A)$ is the expected revenue of assortment A and $R_\tau(A) \equiv \sum_{(j,k) \in A} r_j \cdot \mathbb{P}_\tau((j,k) | A)$ is the expected revenue collected from customer type $\tau = (s, \sigma, \alpha)$, with $\mathbb{P}_\tau$ defined in Equation (4). We further write $w_j \equiv e^{v_j}$ as the attraction parameter of style j and thus $R_\tau(A)$ follows

$$R_\tau(A) = \frac{\sum_{(j,k) \in C_\tau^1(A)} r_j w_j + \sum_{(j,k) \in C_\tau^2(A)} e^{-\alpha} r_j w_j}{1 + \sum_{(j,k) \in C_\tau^1(A)} w_j + \sum_{(j,k) \in C_\tau^2(A)} e^{-\alpha} w_j}.$$

Recall that in Section 4, we showed that size substitution happens. Remarkably, in the following theorem, we show that the size substitution effect does not have an impact on the assortment decision. Additionally, the optimal decision of Problem (9) has a revenue-ordered structure in product styles.

THEOREM 1. *Let $\{1, 2, \dots, j^*\}$ be the optimal assortment under the style-only MNL choice model:*

$$\{1, 2, \dots, j^*\} = \arg \max_{A_{\text{style}} \subseteq \mathcal{J}} \left\{ \frac{\sum_{j \in A_{\text{style}}} r_j w_j}{1 + \sum_{j \in A_{\text{style}}} w_j} \right\}. \quad (10)$$

Then there exists an optimal solution $A^ \subseteq \mathcal{N}$ to the assortment problem (9) that takes the form*

$$A^* = \{(1, k), (2, k), \dots, (j^*, k) \mid k \in \mathcal{K}\}. \quad (11)$$

That is, for style $j \in \{1, \dots, j^\}$, it is optimal to offer all sizes. For other styles, do not offer any size at all.*

Theorem 1 reveals a simplification in assortment planning under the style-size choice model. Although demand substitution can occur both across apparel styles and sizes, which are inherently “two-dimensional,” the optimal assortment follows a one-dimensional structure. Specifically, product sizes and size-substitution effects can be ignored, and the optimal decision can be made solely at the style level, mirroring the classic MNL assortment optimization problem (Talluri and Van Ryzin 2004) in Problem (10) and rendering the revenue-ordered structure of Equation (11). Moreover, *a priori*, an apparel retailer may consider skipping some sizes for less popular styles. That approach would contravene Theorem 1, which states that if it is optimal to include a style in the assortment, then all sizes should be included, regardless of the style’s popularity.

Theorem 1 also provides theoretical justification for size aggregation in assortment optimization, a common approach in the operations management literature (Boada-Collado and Martínez-de

Albéniz 2020). Notably, the optimal assortment (11) remains unchanged regardless of the distribution μ_τ over customer types $\tau = (s, \sigma, \alpha)$. In other words, the optimal assortment decision is independent of whether customers are more flexible with size variations ($\mu_{(s, \sigma, \alpha)}$ concentrated at a low α) or more sensitive to them ($\mu_{(s, \sigma, \alpha)}$ concentrated at a high α). Moreover, this result aligns with industry practices, where retailers typically focus on style selection rather than size differentiation when designing catalogs or arranging store displays. Later, we show that Theorem 1 also leads to an asymptotically optimal inventory policy that remains invariant to size substitution effects.

We utilize the following three facts in the proof of Theorem 1: (i) The unit revenue or net profit of a product only depends on its style, not its size. (ii) The utility of a product only depends on its style and not on its size, as long as the product is of the correct size. (iii) A product has a lower utility to customers if it is of an adjacent size. Note that the second fact also relates to the formation of the consideration sets (Section 3.2). As long as we design the offered assortment to satisfy every customer's first-best choice (here, the best-fit size), customers would behave according to a standard MNL at the style level. Hence, Theorem 1 actually holds for a more general setting of the style-size choice model. First, the theorem applies to the model extension described in Section 3.4, as the assortment A^* defined Equation (11) remains optimal for each possible generalized customer type $(\mathbf{s}, \boldsymbol{\sigma}, \boldsymbol{\alpha})$ by satisfying their best-fit size under each offered style. Second, the theorem also applies to the case that there might exist the third or fourth best-fit size for a customer; see the discussion in the last paragraph of Section 3.1. Given that the first-best choice under each offered style is satisfied by A^* , the second, the third, and the fourth choices do not matter. The above two examples highlight the key strength of Theorem 1 – despite variations in customer demand patterns due to the combinatorial nature of style-size pairs in apparel products, the assortment decision can still be intuitively made.

Finally, Theorem 1 also contributes to the assortment optimization literature. Recall that the style-size choice model resembles the mixed-MNL model, as it is a mixture of consider-then-choose models for many customer types in which the choice step follows an MNL. It is well-known that the optimal assortment of the mixed-MNL model in general does not have the revenue-ordered structure, and finding the optimal assortment is NP-hard (Bront et al. 2009, Rusmevichientong et al. 2014). On the other hand, the optimal assortment of the MNL model does have a revenue-ordered structure (Talluri and Van Ryzin 2004). Thus, we can view the style-size model as an interesting middle point between the MNL and mixed-MNL models.

5.2. Inventory Optimization

We further consider a stockout-based inventory optimization problem under the proposed style-size choice model. For convention, we write $\mathbb{N} \equiv \{1, 2, \dots\}$ as the set of positive integers and $\mathbb{N}_+ \equiv$

$\mathbb{N} \cup \{0\}$. We specify the inventory model as follows. Let $\ell = 1, 2, \dots$ be a sequence of customers. Each customer visits the store at time t^ℓ and makes a purchase decision $D^\ell \in \mathcal{N}_+$. We make two assumptions about the customers. First, we assume that the arrival time of customers $(t^\ell)_{\ell \in \mathbb{N}_+}$ follow a homogeneous Poisson process of rate $\lambda > 0$. For simplicity, we ignore seasonality. Second, we assume that customers' decisions D^ℓ follow the distribution $D^\ell \sim \mathbb{P}(\cdot | A^\ell)$, where A^ℓ is the set of available products when customer ℓ visits and $\mathbb{P}(\cdot | \cdot)$ is the proposed style-size choice model (5).

Let $I_{jk}^\ell \in \mathbb{N}_+$ be the remaining stock of product $(j, k) \in \mathcal{N}$ at time t^ℓ , i.e., at the time that ℓ th customer visits. Then the set of available products is thus defined as $A^\ell = \{(j, k) \in \mathcal{N} \mid I_{j,k}^\ell > 0\}$. The stock $\mathbf{I}^\ell = (I_{jk}^\ell)_{(j,k) \in \mathcal{N}}$ follows the recursive equation: $I_{jk}^{\ell+1} = I_{jk}^\ell - 1$ if $D^\ell = (j, k) \in A^\ell$; and $I_{jk}^{\ell+1} = I_{jk}^\ell$ otherwise. That is, if a customer chooses to buy a product of style j and size k , then the corresponding stock level decreases by one. Notice that $\mathbf{I}^{\ell+1} \geq \mathbf{0}$ for all $\ell \in \mathbb{N}_+$, as $\mathbb{P}((j, k) \mid A^\ell) = 0$ whenever $I_{jk}^\ell = 0$.

The store will make an inventory decision $\mathbf{I} \in \mathbb{N}_+^{|\mathcal{N}|}$ (non-negative integers) for the initial inventory depth, i.e., deciding $\mathbf{I} = \mathbf{I}^1$. Associated with the decision, the store pays unit procurement cost c_j to order each unit of product (j, k) and charges unit price p_j for each sale of (j, k) , which are assumed to be independent of the size k . We also write $p_0 = 0$ and $c_0 = 0$ for the no-purchase option. The goal of the store is to maximize the expected profit up to a given time T . That is, the store maximizes

$$P_{\text{inv}} := \underset{\mathbf{I} \in \mathbb{N}_+^{JK}}{\text{maximize}} \left[\Pi(\mathbf{I}) := \mathbb{E} \left[\sum_{\ell=1}^{\infty} p_{D^\ell} \cdot \mathbb{I}[t_\ell \leq T] \right] - \sum_{(j,k) \in \mathcal{N}} c_j I_{jk} \right]. \quad (12)$$

The objective function $\Pi(\mathbf{I})$, which is the expected profit, consists of two terms, the expected revenue and the total cost. Notice that the revenue $\sum_{\ell=1}^{\infty} p_{D^\ell} \cdot \mathbb{I}[t_\ell \leq T]$ is a random variable, as both customer arrival times and customers' decisions are random. We can also rewrite the expected revenue as follows. Let L be the number of customers that arrive during $[0, T]$. Then L is a Poisson random variable with parameter $T\lambda$ and thus $\mathbb{E}[\sum_{\ell=1}^{\infty} p_{D^\ell} \cdot \mathbb{I}[t_\ell \leq T]] = \mathbb{E}[\sum_{\ell=1}^L p_{D^\ell}]$. We use $\mathbf{I}^*$ and Π_{inv}^* to denote the optimal solution and the optimal objective value of the inventory problem P_{inv} , respectively. Without loss of generality, in this section, we write $\mathcal{J} \equiv \{1, 2, \dots, J\}$ and $\mathcal{K} \equiv \{1, 2, \dots, K\}$, where two sizes k and k' are adjacent if $|k - k'| = 1$. We also label product styles in a way that $r_1 \equiv p_1 - c_1 \geq r_2 \equiv p_2 - c_2 \geq \dots \geq r_J \equiv p_J - c_J \geq 0$, i.e., styles are ordered in a decreasing order of their unit profits. Throughout the section, we write $w_j \equiv \exp(v_j)$ as the attraction parameter for style $j \in \mathcal{J}$.

Before presenting our approach to solve Problem (12), we pause to comment on the computational and theoretical challenges behind it. Notice that the *stockout-based* inventory optimization problem like Problem (12) is notoriously hard (Mahajan and Van Ryzin 2001). In fact, as Aouad

et al. (2018) point out, given an initial inventory vector, even the efficient evaluation of the expected revenue $\mathbb{E} \left[\sum_{\ell=1}^L p_{D^\ell} \right]$ is an open question for most choice models of interest, including the standard MNL model, due to the existence of stockout-based substitution. That is, customers may substitute another product in A^ℓ for their most desired product when it is out of stock, and the substitution follows the choice model $\mathbb{P}(\cdot | A^\ell)$. In addition, the set of available products A^ℓ varies according to each product's in-stock availability. That is why Problem (12) is also referred as the dynamic inventory problem with stockout-based substitution. In contrast, demand substitution in Problem (9) is *assortment-based*, or static, because it assumes that demand is completely determined by the products offered in the assortment regardless of whether they are in stock at any particular point in time. Recall that a product is a style-size pair. However, Theorem 1 shows that assortment planning can be done at the style level under assortment-based substitution.

To illustrate the previous point, consider the revenue collected under two scenarios: (i) customers do not consider adjacent sizes; and (ii) customers do consider adjacent sizes when the best-fit size is not available. We show that ignoring the size substitution effect underestimates a product's expected revenue.

EXAMPLE 3. (Size Substitution Effect in a Stockout-based Setting) Consider a market of one style of a T-shirt $\mathcal{J} = \{1\}$ and two sizes $\mathcal{K} = \{\text{Medium } (M), \text{ Large } (L)\}$. The style has attraction $w_1 = 3$ and unit price $p_1 = 1$. Let all customers in the market have the same size substitution parameter α_0 and each of customer type with $\tau = (s, \sigma, \alpha_0)$ has weight 0.25 for $s \in \{M, L\}$ and $\sigma \in \{+, -\}$. We assume that only the M -size is currently available and the L -size is out of stock, i.e., $A_\ell = \{(1, M)\}$. If we assume that the coming customer ℓ will not consider adjacent sizes, i.e., $\beta_0 := \exp(-\alpha_0) = 0$, then the expected revenue collected from this customer is $p_1 \times (0.25 + 0.25) \times (w_1 / (1 + w_1)) = 0.375$. If we assume that the customer will consider adjacent size with parameter $\beta_0 = 2/3$, then the expected revenue is $0.375 + p_1 \cdot 0.25 \cdot \beta w_1 / (1 + \beta w_1) = 0.525$. The difference in the expected revenue comes from that the fact that in the former case, if the customer happens to be the type $(L, -, \alpha_0)$ and if we assume that she won't consider the M -size T-shirt at all, then the demand of $(1, M)$ is underestimated, leading to an underestimation of the expected revenue. $\square$

From the example, if we do not consider size substitution, we would mistakenly assume that a product of size k can only attract customers whose best-fit size is k and overlook the possibility that the product may be appealing to customers of adjacent sizes. Such underestimation of a product's expected revenue may also lead to suboptimal inventory decisions, as the firm would not stock the product at all if the product's cost c_1 is greater than 0.375 in Example 3. This hints that size substitution should be taken into account in inventory planning, at least when the expected demand over the selling horizon is low, as we show next.

5.3. An IP-Based Inventory Policy

Due to the computational challenges in stockout-based substitution, we first consider solving a lower bound of Problem (12):

$$P_{LB} : \quad \underset{\mathbf{I} \in \mathbb{N}_+^{JK}}{\text{maximize}} \quad \left[\sum_{(j,k) \in \mathcal{N}} p_j \cdot \min \{ T\lambda \cdot \pi_{jk}(\mathbf{I}), I_{jk} \} - \sum_{(j,k) \in \mathcal{N}} c_j \cdot I_{jk} \right], \quad (13)$$

where $\pi_{jk}(\mathbf{I}) = \mathbb{P}((j, k) \mid A(\mathbf{I}))$ is the choice probability of product (j, k) based on the set of available products. The objective function in P_{LB} is indeed a lower bound to the objective Π of Problem (12). It first assumes that customers arrive in a deterministic manner and then approximates a product's demand based on its choice probability given the initial set of available products. Such inventory problems have been widely considered in the literature (Ryzin and Mahajan 1999, Topaloglu 2013) due to their simplicity and tractability compared to the stockout-based substitution problems. In the context of the style-size choice model, the lower bound model (13) utilizes the size substitution effect through the initial assortment. We further approximate the lower bound model (13) by assuming that the style-size choice model has an average size sensitivity parameter α_0 and solve the corresponding inventory problem by a linear mixed-integer program. Therefore, the collection of customer types is $\Gamma = \{(s, \sigma, \alpha_0) \mid s \in \mathcal{K}, \sigma \in \{+, -\}\}$. One can easily relax this assumption by expanding Γ , as discussed in Section 3.4, though it comes at the expense of introducing additional variables.

We define variables as follows. Let $\mathbf{I} \in \mathbb{N}_+^{JK}$ be the inventory decision for stocking I_{jk} units for product (j, k) and variable $\boldsymbol{\xi} \in \mathbb{R}_+^{JK}$ be the sales of each product (j, k) . A key step of solving Problem (13) is to connect the choice probability $\boldsymbol{\pi} = (\pi_{jk})_{j \in \mathcal{J}, k \in \mathcal{K}}$ with the choice model. Specifically, we use $\mathbf{x} \in \{0, 1\}^{JK}$ to indicate whether each product (j, k) is available at time $t = 0$. We also define variables $\mathbf{y} = (y_{j,\tau})_{j \in \mathcal{J}, \tau \in \Gamma}$ for the construction of the consideration sets described in Section 3.2. Variable $y_{j,\tau}$ indicates whether a customer of type $\tau \in \Gamma$ will consider her adjacent size for style j . Consequently, when we have the following constraints for customer type $\tau = (s, \sigma, \alpha_0) \in \Gamma$:

$$y_{j,\tau} \leq x_{j, \text{ADJ}_\sigma(s)}, \quad y_{j,\tau} \leq 1 - x_{j,s}, \quad x_{j, \text{ADJ}_\sigma(s)} - x_{j,s} \leq y_{j,\tau}. \quad (14)$$

This constraint enforces that customer $\tau = (s, \sigma, \alpha_0)$ will not consider the adjacent size $\text{ADJ}_\sigma(s)$ unless the best size s of style j is not available. Next, to represent the choice probability (4) of each customer type, which is a linear-fractional form, we use the classic linearization technique (Charnes and Cooper 1962). For each customer type τ , we use h_τ to denote its no-purchase probability and further use $\theta_{j,\tau}$ and $\phi_{j,\tau}$ to denote the products $x_{j,s}h_\tau$ and $y_{j,\tau}h_\tau$, respectively. We thus have the following constraint system that linearizes h_τ , $\theta_{j,\tau}$ and $\phi_{j,\tau}$:

$$h_\tau + \sum_{j \in \mathcal{J}} w_j \theta_{j,\tau} + \sum_{j \in \mathcal{J}} \beta_0 w_j \phi_{j,\tau} = 1, \quad (15)$$

$$\theta_{j,\tau} \leq h_\tau, \quad \theta_{j,\tau} \leq x_{j,s}, \quad h_\tau \leq 1 + \theta_{j,\tau} - x_{j,s}, \quad (16)$$

$$\phi_{j,\tau} \leq h_\tau, \quad \phi_{j,\tau} \leq y_{j,\tau}, \quad h_\tau \leq 1 + \phi_{j,\tau} - y_{j,\tau}. \quad (17)$$

Finally, as the demand π_{jk} of product (j, k) comes from customers whose best fit size is k and from customers of adjacent sizes, we have

$$\pi_{jk}/w_j = \sum_{\tau \in \{(k, +, \alpha_0), (k, -, \alpha_0)\}} \mu_\tau \theta_{j,\tau} + \sum_{\tau \in \{(k-1, +, \alpha_0), (k+1, -, \alpha_0)\}} \beta_0 \mu_\tau \phi_{j,\tau}, \quad (18)$$

where the second sum characterizes the size substitution from customers of adjacent sizes. With the defined variables and the constraints, we write down the following mixed-integer linear program to solve the lower bound (13), which we call the IP-based inventory policy.

$$\begin{aligned} P_{\text{LB-IP}} := & \quad \text{maximize} \quad \sum_{k \in \mathcal{K}} \sum_{j \in \mathcal{J}} (p_j \cdot \xi_{j,k} - c_j \cdot I_{jk}) \\ & \text{subject to} \quad \xi_{jk} \leq T\lambda \cdot \pi_{jk}, \quad \xi_{jk} \leq I_{jk}, \quad x_{jk} \leq I_{jk} \leq M \cdot x_{jk} \quad \forall k \in \mathcal{K}, j \in \mathcal{J}, \\ & \quad \text{Constraints (14)-(18)} \\ & \quad I_{jk}, \xi_{jk} \in \mathbf{N}_+, \quad x_{jk} \in \{0, 1\}, \quad \pi_{jk}, y_{j,\tau}, \theta_{j,\tau}, \phi_{j,\tau} \in [0, 1]. \end{aligned} \quad (19)$$

Here M is a large constant in the big- M notation. For the boundary cases of sizes, we simply set $x_{j,k-1} = 0$ for $k = 1$ and $x_{j,k+1} = 0$ for $k = K$.

In the following numerical study, we examine the performance of the IP-based inventory policy and highlight its advantages when the expected demand over the selling horizon is low, which is in contrast to the asymptotic regime to be introduced in Section 5.4. We calibrate the choice model parameters using the real-world dataset discussed in Section 4, including the utility v_j for each style $j \in \mathcal{J}$ and the fraction μ_τ of customer type τ . The dataset also provides the price p_j for each style $j \in \mathcal{J}$, while the cost c_j of the product is not available. To address this, we make an assumption that the firm implements a 120% markup pricing scheme. This assumption aligns with insights from practitioners (Farra 2019, Claypoole 2019) which suggest that firms usually markup products with a 120% to 150% margin. We vary the expected number of customers $\bar{L} = T\lambda$ to evaluate the performance of the policies in the non-asymptotic regime. From Section 4.1 we know that each store receives approximately $W = 4000$ visitors per week on average. Consequently, we examine scenarios ranging from one month (roughly four weeks) to eight months (roughly thirty-two weeks) by setting $\bar{L} \in \{4W, 8W, 12W, 16W, 20W, 24W, 32W\}$, consistent with the scale we observed in Section 4.

We conduct a comparison between the IP-based inventory policy and two benchmark inventory policies: the newsvendor policy and the fluid approximation (Zhang et al. 2024). Specifically, the newsvendor policy is given by the standard quantile policy in which the demand of each product (j, k) is treated independently. The fluid approximation stocks I_{jk} units of product (j, k) as $I_{jk}^{\text{FA}} =$

$\lceil T\lambda \cdot \mathbb{P}((j, k) \mid A^*) \rceil$, where A^* is the optimal assortment to Problem (9) with $r_j = p_j - c_j$. We remark that both the newsvendor policy and the fluid approximation are *size-substitution-invariant*. That is, the stocking decisions under both policies ignore the value of α_0 . For the newsvendor policy, such property is obviously true as the policy views each product's demand independently. For the fluid approximation, since $\mathbb{P}(\cdot \mid A^*)$ is invariant under α_0 according to Theorem 1, we know that the resulting stocking decision $\lceil T\lambda \cdot \mathbb{P}(\cdot \mid A^*) \rceil$ is also invariant.

In what follows, we assess the performance of each inventory policy by evaluating the expected profit generated by the corresponding inventory vector. Specifically, let $\mathbf{I}^{\text{IP}}$, $\mathbf{I}^{\text{FA}}$, and $\mathbf{I}^{\text{NV}}$ be the inventory vector returned by the IP-based, fluid approximation, and newsvendor policies, respectively. To evaluate the profit $\Pi(\cdot)$ associated with each inventory vector, we employ Monte Carlo simulation based on the stochastic process outlined in Section 5.2, along with the common random number technique for variance reduction. We consider two values for $\beta_0 = \exp(-\alpha_0) \in \{24.9\%, 100.0\%\}$. The former corresponds to the estimated value $\alpha_0 = 1.39$ obtained from the dataset, whereas the latter represents the maximum value that β_0 can take, which happens when $\alpha = 0$, as stated in Lemma 1. It corresponds to the scenario that customers are highly tolerant to the adjacent sizes and the demand loss due to the stock-out of the best-fit size can be completely compensated by the adjacent sizes. By examining these two values, we can explore a range of scenarios and assess the sensitivity of the results.

$\beta_0 = 24.9\%$						$\beta_0 = 100.0\%$					
$\bar{L}$	$\Pi_{\text{BT}}^{\text{per}}$			N_{Size}	N_{Prod}	$\bar{L}$	$\Pi_{\text{BT}}^{\text{per}}$			N_{Size}	N_{Prod}
	$\mathbf{I}^{\text{NV}}$	$\mathbf{I}^{\text{FA}}$	$\mathbf{I}^{\text{IP}}$	$\mathbf{I}^{\text{IP}}$	$\mathbf{I}^{\text{IP}}$		$\mathbf{I}^{\text{NV}}$	$\mathbf{I}^{\text{FA}}$	$\mathbf{I}^{\text{IP}}$	$\mathbf{I}^{\text{IP}}$	$\mathbf{I}^{\text{IP}}$
4W	-18.88	-18.65	0.11	3.0	3	4W	-18.88	-18.65	0.29	2.0	12
8W	-4.45	-3.89	0.79	2.7	30	8W	-3.07	-2.39	2.07	2.5	86
12W	-0.16	0.53	1.72	2.6	90	12W	1.57	2.34	3.74	3.2	161
16W	1.30	2.09	2.49	3.7	162	16W	2.50	3.23	4.13	3.9	201
20W	2.71	3.28	3.42	4.5	230	20W	3.51	4.16	4.69	4.6	235
24W	3.02	3.53	3.59	5.4	275	24W	3.93	4.68	5.05	4.9	250
32W	3.71	4.31	4.34	7.1	362	32W	4.98	5.54	5.67	5.6	286

Table 3 Expected profit per customer $\Pi_{\text{BT}}^{\text{per}}$, sizes offered N_{Size} , and products available N_{Prod} for varying demand $\bar{L}$ with $\beta_0 \in \{24.9\%, 100.0\%\}$. The newsvendor and fluid approximation offer all products (and sizes).

Table 3 displays $\Pi_{\text{BT}}^{\text{per}}(\cdot)$, the expected profit per customer visit to the casual booties category, which is defined as $\Pi_{\text{BT}}^{\text{per}}(\cdot) = \Pi(\cdot)/\bar{L}_{\text{BT}}$. Note that we do not have the exact customer traffic for casual booties in the dataset. We thus approximate $\bar{L}_{\text{BT}}$ by multiplying the total customer visits $\bar{L}$ by the fraction of sales of the casual booties category (roughly 2.8%), i.e., $\bar{L}_{\text{BT}} = 0.028\bar{L}$. The table also presents the number of sizes offered N_{Size} and the number of products available N_{Prod}

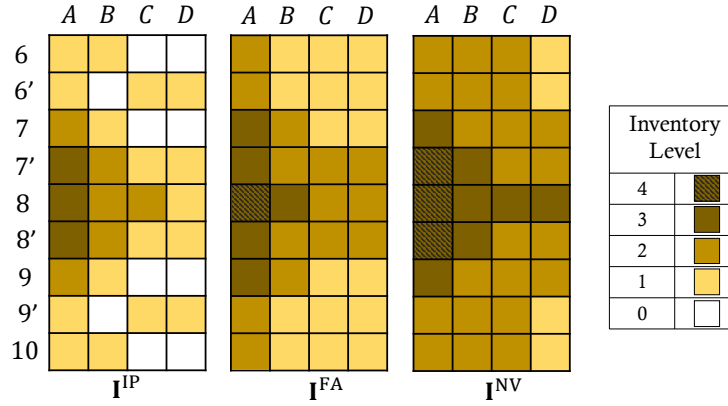


Figure 3 Inventory profile of the three policies for the four most popular styles (of the fifty-one) and nine sizes, with $\bar{L} = 12W$ and $\beta_0 = 24.9\%$ (where the prime sign represents half sizes)

under the IP-based policy. The newsvendor and fluid approximation offer all styles in all sizes, so the number of products available under those policies is $51 \times 9 = 459$.

In Table 3, we observe that all three inventory policies exhibit superior performance when $\bar{L}$ is large, which can be attributed to the decreased demand volatility. However, when $\bar{L}$ is small, both the newsvendor and fluid approximation perform poorly regardless of the level of size substitution given by the parameter β_0 . The reason is that these two policies stock too much – at least one unit for each size of each style – so substitution does not occur, in which case β_0 is irrelevant. In contrast, the IP-based policy incorporates size substitution and strategically offers a smaller set of sizes and styles to satisfy the demand, resulting in positive profits. Figure 3 visualizes the stocking decisions made by the three inventory policies for the four most popular styles out of 51 in the dataset when $\bar{L} = 12W$ and $\beta_0 = 24.9\%$. Note that Style *A* is also the most expensive style. The figure shows that, in contrast to the newsvendor and fluid approximation, the IP-based policy does not offer the complete range of sizes for all styles. Instead, it leverages the size substitution effect to fulfill unmet demand. For instance, it does not offer sizes 6 and 10 of Styles *C* and *D*, as the demand of these products can be covered by sizes 6.5 and 10.5 of the same style, respectively. The IP-based policy also holds less inventory: 1.5 units per product in Figure 3, whereas the fluid approximation and newsvendor hold 1.8 and 2.3 units per product, respectively.

We also observe that the profitability of the IP-based policy is higher as size substitution becomes more prevalent. In the left panel of Table 3, we can see that the expected profit per customer of the IP-based policy is 19% higher compared to the fluid approximation when $\bar{L} = 16W$ and $\beta_0 = 24.9\%$. This advantage increases to 28% when $\beta_0 = 100.0\%$, as shown in the right panel. Similarly, while $\mathbf{I}^{\text{IP}}$ and $\mathbf{I}^{\text{FA}}$ statistically have the similar performance when $\bar{L} = 32W$ and $\beta_0 = 24.9\%$, the former is strictly better than the latter for the same $\bar{L}$ when $\beta_0 = 100.0\%$. This highlights the importance of

incorporating size substitution when customers show a strong tendency to explore adjacent sizes. However, the advantage of the IP-based policy diminishes as $\bar{L}$ increases. In the left panel where $\beta_0 = 24.9\%$, the advantage of $\mathbf{I}^{\text{IP}}$ over $\mathbf{I}^{\text{FA}}$ shrinks from 19% to 1% as $\bar{L}$ increases from $16W$ to $32W$. Moreover, we will show that the IP-based policy and the fluid approximation have the same asymptotic limit (see Section 5.4). Since the fluid approximation is size-substitution-invariant, the convergence of both policies suggests that the effect of size substitution shrinks as overall demand increases. We will revisit this discussion from a theoretical standpoint in Section 5.4.

It is hard to compare the IP-based policy to the (average) store performance reported in Table 1 because the latter includes inventory replenishment and the styles were introduced in a staggered manner. However, it is worth noting that the maximum number of products available was 359.2 over an horizon of 33 weeks. This contrasts with the IP-policy that suggests carrying 362 products when $\bar{L} = 32W$ and $\beta_0 = 24.9\%$. In terms of sizes offered, the average store in Table 1 started with 8.4 sizes whereas the IP-policy suggests 7.1. In other words, the IP-policy offers slightly less sizes but they are distributed across a wider selection of styles. Indeed, all styles are initially available under the IP-policy since $362/7.1 = 51$, whereas a back-of-the-envelope calculation based on Table 1 gives $359.2/8.4 = 42.8$.³

We end the discussion with two additional remarks. First, the IP-based policy is not computationally too expensive. For all instances in Table 3, the mixed-integer linear program (19) was optimally solved within five minutes; see Appendix D.1 for more details. Given that these instances involve nearly five hundred products, this runtime highlights the compactness of the style-size choice model. Second, the IP-based policy is flexible in accommodating other business constraints. It offers the convenience of incorporating capacity limitations into product inventory, which can be based on factors such as style or size. For instance, one can enforce a distinction between major and minor sizes and ensure that minor sizes are not offered unless all major sizes are provided. This type of policy has already found successful implementation in the fashion industry (Caro and Gallien 2010). In our study, we have incorporated such constraints into the IP-based policy and present its performance in Appendix D. Additionally, the IP-based policy allows for easy inclusion of initial stock or remaining stock from the previous period in the integer program. Combining these features with its favorable performance for short planning horizons, the IP-based policy can be an effective tool for making replenishment decisions during the sales season.

³ More precisely, the 359.2 products in the average store came from 44.75 styles and 8.03 sizes.

5.4. Asymptotically, Size Substitution Does Not Matter

In this section we study the asymptotic regime, i.e., when the expected customer volume $\bar{L} = T\lambda$ grows to infinity. Recall that by Theorem 1, the fluid approximation can be expressed as $I_{jk}^{\text{FA}} = \lceil \bar{L} \cdot \mathbb{P}((j, k) \mid A^*) \rceil \equiv \lceil \bar{L} \rho_j \bar{\mu}_k \rceil$, where

$$\rho_j = \frac{w_j \cdot \mathbb{I}_{j \leq j^*}}{1 + \sum_{j \leq j^*} w_j} \quad \text{and} \quad \bar{\mu}_k = \int_0^\infty \sum_{\sigma \in \{+, -\}} \mu_{(k, \sigma, \alpha)} d\alpha. \quad (20)$$

Here j^* is defined as in the style-only assortment problem (10) with margin $r_j = p_j - c_j$. As mentioned, the fluid approximation is size-substitution-invariant because the quantities it prescribes are independent of the size sensitivity parameter α and its distribution. One can interpret the fluid approximation as follows. The firm first solves the style-only MNL assortment optimization problem (10) to decide which styles to offer. For each offered style $j \in \{1, 2, \dots, j^*\}$, the store will stock in total $\bar{L} \rho_j$ units, which is based on the style-only MNL model. Furthermore, among these $\bar{L} \rho_j$ units of style j , the store allocates a fraction $\bar{\mu}_k$ of it to size k , i.e., it stocks $\bar{L} \rho_j \bar{\mu}_k$ units for product (j, k) , where $\bar{\mu}_k$ is the fraction of customers whose best-fit size is k . The fluid approximation actually is an aggregation-disaggregation approach, as the firm first aggregates all products across sizes when deciding which styles to offer, and then disaggregates or “splits” the demand of each offered style among sizes. In the following proposition we show that such aggregation-disaggregation approach is asymptotically optimal.

PROPOSITION 1. *Assume that the maximal product price $p_{\max} = \max_j p_j$ and the maximal product cost $c_{\max} = \max_j c_j$ are independent of both the horizon T and the customer arrival rate λ . For the stockout-based inventory optimization problem (12), the fluid approximation policy $\mathbf{I}^{\text{FA}}$ has optimality gap $O\left(\sqrt{JK \cdot T\lambda}\right)$ and it is asymptotically optimal.*

Note that the asymptotic performance is defined as the approximation ratio of an inventory policy relative to the optimal solution as $T\lambda \rightarrow \infty$. In Section B.3 where we prove the proposition, we show that the approximation ratio converges to one under the fluid approximation policy, implying the asymptotic optimality. Alternatively, Proposition 1 shows that as the customer volume increases, the profit loss *per customer* eventually reaches to zero. This follows from the fact that while the optimality gap grows at a rate of $\sqrt{T\lambda}$, the expected number of customer scales as $T\lambda$.

Proposition 1 has an intuitive interpretation: as customer volume increases, the stochasticity of the problem diminishes because the standard deviation of demand grows at a slower rate, so just stocking the mean becomes a sufficiently good strategy, which is akin to ignoring size substitution as in Theorem 1. Formally, our proof follows the performance guarantee of the fluid approximation in the inventory problem under choice models that satisfy the substitutability property (Zhang

et al. 2024). Per Lemma 1, the result in Zhang et al. (2024) is applicable to our inventory problem, though a modification is required to consider a random number of customer arrivals L , as Zhang et al. (2024) assume that the number of customer visits is deterministic and known in advance.

We highlight that Proposition 1 supports the common practice of ignoring size substitution for stocking purposes. However, ignoring both style and size substitution like in the newsvendor model could lead to a poor performance. We demonstrate this observation in Appendix D.3. Another important observation is given in the following proposition. It shows that the performance of the IP-based solution $\mathbf{I}^{\text{IP}}$ introduced in Section 5.3 and the fluid approximation $\mathbf{I}^{\text{FA}}$ becomes indistinguishable when the expected demand $\bar{L}$ is sufficiently large.

PROPOSITION 2. *The IP-based policy and the fluid approximation have the same asymptotic performance.*

Proposition 2 gives an edge to the IP-based policy because it matches the asymptotic performance of the fluid approximation, and per section 5.3, it has a better performance in the non-asymptotic regime. Put differently, in the asymptotic regime a “wide-net” approach that stocks all sizes works well, whereas in the non-asymptotic regime a more targeted approach is better. One can think that the former is more applicable to online settings, whereas the latter could make more sense for brick-and-mortar stores. Finally, to complement Propositions 1 and 2, in Appendix E, we further explore the asymptotic performance of a fluid-like policy under general choice model environment that may not follow the substitutability property of Lemma 1.

6. Conclusion and Future Directions

We introduced the style-size choice model to capture size substitution effects and demonstrated, using real-world data, that unmet demand due to stockouts shifts to adjacent sizes of the same style. We then analyzed assortment and inventory optimization under this model, showing that firms can disregard size substitution in static (assortment-based) settings and in dynamic (stockout-based) settings when the demand is high. In the low-demand regime, we proposed an IP-based solution to leverage size substitution in a computationally tractable manner. Our work opens several directions for future research, such as allowing for inventory replenishment or incorporating a goodwill cost when customers like a style but cannot find a suitable size. The latter could lead to excessive leftover inventory, adding an environmental dimension to the problem. Additionally, it would be valuable to investigate whether other choice models, once the apparel style and size structure are incorporated, would also yield a simplified optimal assortment structure.

Acknowledgments

We sincerely appreciate the insightful feedback from the department editor Feryal Erhun, the associate editor, and two anonymous referees, which has significantly enhanced this work.

References

- Yi-Chun Akchen and Velibor V Mišić. Assortment optimization under the decision forest model. *arXiv preprint*, 2021.
- Yi-Chun Akchen and Dmitry Mitrofanov. Consider or choose? the role and power of consideration sets. *arXiv preprint*, 2023.
- Ali Aouad, Retsef Levi, and Danny Segev. Greedy-like algorithms for dynamic assortment planning under multinomial logit preferences. *Operations Research*, 66(5):1321–1345, 2018.
- Ali Aouad, Vivek Farias, and Retsef Levi. Assortment optimization under consider-then-choose choice models. *Management Science*, 67(6):3368–3386, 2021.
- Henry David Block and Jacob Marschak. Random orderings and stochastic theories of response. Technical report, Cowles Foundation for Research in Economics, Yale University, 1959.
- Pol Boada-Collado and Victor Martínez-de Albéniz. Estimating and optimizing the impact of inventory on consumer choices in a fashion retail setting. *Manufacturing & service operations management*, 22(3):582–597, 2020.
- Juan José Miranda Bront, Isabel Méndez-Díaz, and Gustavo Vulcano. A column generation algorithm for choice-based network revenue management. *Operations research*, 57(3):769–784, 2009.
- Gérard P Cachon and Robert Swinney. The value of fast fashion: Quick response, enhanced design, and strategic consumer behavior. *Management science*, 57(4):778–795, 2011.
- Katia Campo, Els Gijsbrechts, and Patricia Nisol. Towards understanding consumer response to stock-outs. *Journal of Retailing*, 76(2):219–242, 2000.
- Felipe Caro and Jérémie Gallien. Inventory management of a fast-fashion retail network. *Operations research*, 58(2):257–273, 2010.
- Felipe Caro and Jérémie Gallien. Clearance pricing optimization for a fast-fashion retailer. *Operations research*, 60(6):1404–1422, 2012.
- Felipe Caro and Victor Martínez-de Albéniz. Operations management in apparel retailing: Processes, frameworks and optimization. *BEIO, Boletín de Estadística e Investigación Operativa*, 29(2):103–116, 2013.
- Felipe Caro and Victor Martínez-de Albéniz. Fast fashion: Business model overview and research opportunities. *Retail supply chain management: Quantitative models and empirical studies*, 2015.
- Felipe Caro, Jérémie Gallien, Miguel Díaz, Javier García, José Manuel Corredoira, Marcos Montes, José Antonio Ramos, and Juan Correa. Zara uses operations research to reengineer its global distribution process. *Interfaces*, 40(1):71–84, 2010.
- Abraham Charnes and William W Cooper. Programming with linear fractional functionals. *Naval Research logistics quarterly*, 9(3-4):181–186, 1962.
- Hai Che, Xinlei Chen, and Yuxin Chen. Investigating effects of out-of-stock on consumer stockkeeping unit choice. *Journal of Marketing Research*, 49(4):502–513, 2012.
- Ningyuan Chen, Guillermo Gallego, and Zhuodong Tang. The use of binary choice forests to model and estimate discrete choices. *arXiv preprint*, 2019.
- Yi-Chun Chen and Velibor V Mišić. Decision forest: A nonparametric approach to modeling irrational choice. *Management Science*, 68(10):7090–7111, 2022.
- Cheryl Claypoole. What is the markup percentage for retail clothing? <https://smallbusiness.chron.com/markup-percentage-retail-clothing-80777>, 2019.
- Yiting Deng, Yuexing Li, and Jing-Sheng Jeannette Song. A unified parsimonious model for structural demand estimation accounting for stockout and substitution. *Available at SSRN*, 2022.
- Omar El Housni, Vineet Goyal, Salal Humair, Omar Mouchtaki, Ali Sadighian, and Jingchen Wu. Joint assortment and inventory planning for heavy tailed demand. 2021.
- V. F. Farias, S. Jagabathula, and D. Shah. A nonparametric approach to modeling choice with limited data. *Management science*, 59(2):305–322, 2013.
- Emily Farra. What is the right price for fashion? <https://www.vogue.com/article/what-is-the-right-price-for-fashion>, 2019.
- Marshall Fisher and Ramnath Vaidyanathan. A demand estimation procedure for retail assortment optimization with results from implementations. *Management Science*, 60(10):2401–2415, 2014.
- Guillermo Gallego, Richard Ratliff, and Sergey Shebalov. A general attraction model and sales-based linear program for network revenue management under customer choice. *Operations Research*, 63(1):212–232, 2015.
- Jérémie Gallien, Adam J Mersereau, Andres Garro, Alberte Dapena Mora, and Martín Nóvoa Vidal. Initial shipment decisions for new products at Zara. *Operations Research*, 63(2):269–286, 2015.
- Vineet Goyal, Retsef Levi, and Danny Segev. Near-optimal algorithms for the assortment planning problem under dynamic substitution and stochastic demand. *Operations Research*, 64(1):219–235, 2016.

- Dorothee Honhon and Sridhar Seshadri. Fixed vs. random proportions demand models for the assortment planning problem under stockout-based substitution. *Manufacturing & Service Operations Management*, 15(3):378–386, 2013.
- Dorothee Honhon, Vishal Gaur, and Sridhar Seshadri. Assortment planning and inventory decisions under stockout-based substitution. *Operations research*, 58(5):1364–1379, 2010.
- J. Huber, J. W. Payne, and C. Puto. Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of consumer research*, 9(1):90–98, 1982.
- Srikanth Jagabathula and Paat Rusmevichientong. The limit of rationality in choice modeling: Formulation, computation, and implications. *Management Science*, 65(5):2196–2215, 2019.
- Srikanth Jagabathula, Dmitry Mitrofanov, and Gustavo Vulcano. Demand estimation under uncertain consideration sets. *Operations Research*, 2024.
- A Gürhan Kök and Marshall L Fisher. Demand estimation and assortment optimization under substitution: Methodology and application. *Operations Research*, 55(6):1001–1021, 2007.
- Joonkyum Lee, Vishal Gaur, Suresh Muthulingam, and Gary F Swisher. Stockout-based substitution and inventory planning in textbook retailing. *Manufacturing & Service Operations Management*, 18(1):104–121, 2016.
- Songtao Li, Lauren Xiaoyuan Lu, Susan F Lu, and Simin Huang. Estimating the stockout-based demand spillover effect in a fashion retail setting. *Manufacturing & service operations management*, 2023.
- AJ Liang, Stefanus Jasin, and Joline Uichanco. Assortment and inventory planning under dynamic substitution with mnl model: An LP approach and an asymptotically optimal policy. *SSRN*, 2021.
- Siddharth Mahajan and Garrett Van Ryzin. Stocking retail assortments under dynamic consumer substitution. *Operations Research*, 49(3):334–351, 2001.
- Reza Yousefi Maragheh, Alexandra Chronopoulou, and James Mario Davis. A customer choice model with halo effect. *arXiv preprint*, 2018.
- Victor Martínez-de Albéniz and Sumit Kunnumkal. A model for integrated inventory and assortment planning. *Management Science*, 68(7):5049–5067, 2022.
- Geoffrey J McLachlan and Thiriyambakam Krishnan. *The EM algorithm and extensions*. John Wiley & Sons, 2007.
- Andrés Musalem, Marcelo Olivares, Eric T Bradlow, Christian Terwiesch, and Daniel Corsten. Structural estimation of the effect of out-of-stocks. *Management Science*, 56(7):1180–1197, 2010.
- Daniele Pennesi. A foundation for cue-triggered behavior. *Management Science*, 67(4):2403–2419, 2021.
- Jörg Rieskamp, Jerome R Busemeyer, and Barbara A Mellers. Extending the bounds of rationality: Evidence and theories of preferential choice. *Journal of Economic Literature*, 44(3):631–661, 2006.
- Paat Rusmevichientong and Huseyin Topaloglu. Robust assortment optimization in revenue management under the multinomial logit choice model. *Operations research*, 60(4):865–882, 2012.
- Paat Rusmevichientong, David Shmoys, Chaoxu Tong, and Huseyin Topaloglu. Assortment optimization under the multinomial logit model with random choice parameters. *Production and Operations Management*, 23(11):2023–2039, 2014.
- Garrett van Ryzin and Siddharth Mahajan. On the relationship between inventory costs and variety benefits in retail assortments. *Management Science*, 45(11):1496–1509, 1999.
- Stephen A Smith and Dale D Achabal. Clearance pricing and inventory policies for retail chains. *Management Science*, 44(3):285–300, 1998.
- Kalyan Talluri and Garrett Van Ryzin. Revenue management under a general discrete choice model of consumer behavior. *Management Science*, 50(1):15–33, 2004.
- Huseyin Topaloglu. Joint stocking and product offer decisions under the multinomial logit model. *Production and Operations Management*, 22(5):1182–1199, 2013.
- Kenneth E Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.
- Amos Tversky and Itamar Simonson. Context-dependent preferences. *Management science*, 39(10):1179–1189, 1993.
- G. van Ryzin and G. Vulcano. A market discovery algorithm to estimate a general class of nonparametric choice models. *Management Science*, 61(2):281–300, 2014.
- Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018.
- Jingwei Zhang, Will Ma, and Huseyin Topaloglu. Leveraging the degree of dynamic substitution in assortment and inventory planning. *Operations Research*, 2024.

Appendix A: Model Estimation: The EM Algorithm

We propose an estimation procedure for the average style-size choice model (6) based on the expectation-maximization (EM) algorithm. We will show that the model leads to a clean M step in the EM algorithm, which only requires solving a single concave maximization problem. We will highlight the simplicity and the flexibility of the EM algorithm of the style-size choice model in Section A.3 when comparing to that of other choice models.

We assume that we have access to stores' inventory and sales data. In particular, let N^{mt} be the number of visitors to store $m \in \mathcal{M}$ during week $t \in \mathcal{T}$. Let the non-negative integer $I_{(j,k)}^{mt}$ be the inventory of product $(j,k) \in \mathcal{N}$ at store $m \in \mathcal{M}$ at the beginning of week $t \in \mathcal{T}$. Let $A^{mt} = \{(j,k) \in \mathcal{N} \mid I_{(j,k)}^{mt} > 0\}$ denote the set of available products (assortment) at the beginning of week t at store m . Let $Q_{(j,k)}^{mt}$ be the sales during week t at store m for product $(j,k) \in A^{mt}$. This form of data $(N^{mt}, \{Q_{(j,k)}^{mt}\}_{(j,k) \in A^{mt}})_{m \in \mathcal{M}, t \in \mathcal{T}}$ has been widely used in the inventory management literature (Boada-Collado and Martínez-de Albéniz 2020). Notice that by definition $Q_{(0,0)}^{mt} = N^{mt} - \sum_{(j,k) \in A^{mt}} Q_{(j,k)}^{mt}$ is the number of customers who visited the store m at week t but didn't make a purchase (or made an outside choice). Finally, let $\mathbb{I}_E$ be the indicator function that equals one if event E is true.

Note that we *do not* observe customer types in the dataset and they can be considered a latent variable in this case. Therefore, we consider an expectation-maximization (EM) approach to estimate the model from data, which is a popular procedure to estimate predictive models with latent variables (McLachlan and Krishnan 2007). We also incorporate the fixed effects for seasonality in the estimation. As observed in Section 4, our dataset consists of sales across thirty-three weeks, covering the spring and fall sale seasons. It is thus important to capture how sales were affected by seasonality. To this end, we assume that the size-independent part of the utility of product (j,k) in week t is $v_j + v_t$, instead of v_j , in Equation (1).

A.1. The Complete Data Log-Likelihood Function

Recall that with the average style-size choice model (6), our goal is to estimate $\alpha = \alpha_0$, the average size sensitivity parameter, along with the style utility parameters $(v_j)_{j \in \mathcal{J}}$, the seasonality parameters $(v_t)_{t \in \mathcal{T}}$, and the distribution over customer types μ_τ , where a type is $\tau = (s, \sigma, \alpha_0)$. Note that in the average model, the collection of customer types is reduced to $\Gamma = \{(s, \sigma, \alpha_0) \mid s \in \mathcal{K}, \sigma \in \{+, -\}\}$.

For now, assume that we have the “complete” data $(N_\tau^{mt}, \{Q_{\tau,(j,k)}^{mt}\}_{(j,k) \in A^{mt}})_{\tau \in \Gamma, m \in \mathcal{M}, t \in \mathcal{T}}$, which include customers' types. Here N_τ^{mt} is the number of type- τ visitors at store m during week t and $Q_{\tau,(j,k)}^{mt}$ is the number of sales of product (j,k) made by type- τ visitors at store m during week

t . Obviously, we have $N^{mt} = \sum_{\tau \in \Gamma} N_{\tau}^{mt}$ and $Q_{(j,k)}^{mt} = \sum_{\tau \in \Gamma} Q_{\tau,(j,k)}^{mt}$. The likelihood of the complete data for store m during week t is

$$f_{\text{complete}}^{mt} = \frac{N^{mt}!}{\prod_{\tau} N_{\tau}^{mt}!} \cdot \prod_{\tau} (\mu_{\tau})^{N_{\tau}^{mt}} \cdot \prod_{\tau} f_{\tau,\text{complete}}^{mt},$$

where factor $(N^{mt}! / \prod_{\tau} N_{\tau}^{mt}!) \cdot \prod_{\tau} (\mu_{\tau})^{N_{\tau}^{mt}}$ is the multinomial distribution of customer types and

$$f_{\tau,\text{complete}}^{mt} (N_{\tau}^{mt}, \{Q_{\tau,(j,k)}^{mt}\}_{(j,k) \in A^{mt}}) = \frac{N_{\tau}^{mt}!}{\left(N_{\tau}^{mt} - \sum_{(j,k) \in A^{mt}} Q_{\tau,(j,k)}^{mt}\right)! \cdot \prod_{(j,k) \in A^{mt}} Q_{\tau,(j,k)}^{mt}!} \cdot \left(\prod_{(j,k) \in A^{mt}} \mathbb{P}_{\tau}^{mt}((j,k) | A^{mt})^{Q_{\tau,(j,k)}^{mt}} \right) \cdot \left(1 - \sum_{(j,k) \in A^{mt}} \mathbb{P}_{\tau}^{mt}((j,k) | A^{mt}) \right)^{N_{\tau}^{mt} - \sum_{(j,k) \in A^{mt}} Q_{\tau,(j,k)}^{mt}}.$$

Take logarithm of $\prod_{m,t} f_{\text{complete}}^{mt}$, we obtain the complete data log likelihood. Specifically, the complete data log-likelihood function is equal to a constant plus $\mathcal{L}_{\text{complete}} = \mathcal{L}_1 + \mathcal{L}_2$, where $\mathcal{L}_1 \equiv \sum_{\tau \in \Gamma} \left(\sum_{m,t} N_{\tau}^{mt} \right) \cdot \log(\mu_{\tau})$ and

$$\begin{aligned} \mathcal{L}_2 \equiv & \sum_{m,t,\tau} \sum_{(j,k) \in A^{mt}} Q_{\tau,(j,k)}^{mt} \cdot \left[(v_j + v_t) \cdot \mathbb{I}_{(j,k) \in C_{\tau}^1(A^{mt})} + (v_j + v_t - \alpha_0) \cdot \mathbb{I}_{(j,k) \in C_{\tau}^2(A^{mt})} \right] \\ & - \sum_{m,t,\tau} N_{\tau}^{mt} \cdot \log \left(1 + \sum_{(j,k) \in C_{\tau}^1(A^{mt})} e^{v_j + v_t} + \sum_{(j,k) \in C_{\tau}^2(A^{mt})} e^{v_j + v_t - \alpha_0} \right). \end{aligned}$$

Note that the $\mathcal{L}_1$ only depends on $\boldsymbol{\mu} = (\mu_{\tau})_{\tau \in \Gamma}$ while $\mathcal{L}_2$ only depend on $(\mathbf{v}, \alpha_0)$, where $\mathbf{v} \equiv ((v_j)_{j \in \mathcal{J}}, (v_t)_{t \in \mathcal{T}})$. Therefore, to find the model parameter $(\boldsymbol{\mu}, \mathbf{v}, \alpha_0)$ that maximizes the complete data log likelihood $\mathcal{L}_{\text{complete}}$, we solve two separate optimization problems,

$$P_1^{\text{complete}} : \underset{\mathbf{1}^T \boldsymbol{\mu} = 1, \boldsymbol{\mu} \geq 0}{\text{maximize}} \left\{ \mathcal{L}_1 \mid \mu_{(s,+,\alpha_0)} = \mu_{(s,-,\alpha_0)}, \forall s \in \mathcal{K} \right\} \quad \text{and} \quad P_2^{\text{complete}} : \underset{\alpha \geq 0, \mathbf{v}}{\text{maximize}} \left\{ \mathcal{L}_2 \right\},$$

where the constraints in P_1^{complete} come from the symmetric-weight assumption in the average style-size model. Note that P_1^{complete} has a closed-form unique solution $\mu_{(k,+, \alpha_0)} = \mu_{(k,-, \alpha_0)} = \sum_{m,t} (N_{(k,+, \alpha_0)}^{mt} + N_{(k,-, \alpha_0)}^{mt}) / (2 \cdot \sum_{m,t,\tau} N_{\tau}^{mt})$. Meantime, the second problem P_2^{complete} is a concave maximization problem in $(\mathbf{v}, \alpha_0)$ that can be solved using standard optimization software.

A.2. The E and M steps of the EM algorithm

Recall that we do not observe customer types in the data. Therefore, the parameters N_{τ}^{mt} and $Q_{\tau,(j,k)}^{mt}$ in optimization problems P_1^{complete} and P_2^{complete} are *not* available. We will instead replace them by their conditionally expected values given the choice model parameter $\boldsymbol{\nu} = (\boldsymbol{\mu}, \mathbf{v}, \alpha_0)$.

We start with any initialization of values of $\boldsymbol{\nu}^{(0)}$. In the EM algorithm, we generate a sequence of parameters $\{\boldsymbol{\nu}^{(q)}, q = 1, 2, \dots\}$ until convergence. Assume that we are currently in the q th iteration. We describe how we generate model $\boldsymbol{\nu}^{(q+1)}$ based on $\boldsymbol{\nu}^{(q)}$ through an “E” step then an “M” step.

The E Step: By Bayes' rule, given an assortment A^{mt} at store m during week t , product $(j, k) \in A^{mt} \cup \{(0, 0)\}$, and model parameter $\boldsymbol{\nu}^{(q)}$, we can infer the likelihood that a sale of (j, k) was contributed by a type- τ customer via

$$\mathbb{P}^{mt}(\tau \mid A^{mt}, (j, k), \boldsymbol{\nu}^{(q)}) = \frac{\mathbb{P}_\tau^{mt}((j, k) \mid A^{mt}, \mathbf{v}^{(q)}) \times \mu_\tau^{(q)}}{\sum_{\tau' \mid (j, k) \in C_{\tau'}(A^{mt})} \mathbb{P}_{\tau'}^{mt}((j, k) \mid A^{mt}, \mathbf{v}^{(q)}) \times \mu_{\tau'}^{(q)}},$$

where $\mathbb{P}_\tau^{mt}((j, k) \mid A^{mt}, \mathbf{v}^{(q)})$ is defined as in Equation (4) with α replaced by $\alpha_0^{(q)}$ and v_j replaced by $v_j^{(q)} + v_i^{(q)}$ since we consider the fixed effect for seasonality. For a customer type τ such that $(j, k) \notin C_\tau(A^{mt})$, the conditional value is simply zero. With the conditional probability, we have that, for $(j, k) \in A^{mt} \cup \{(0, 0)\}$, the expected sales from customer type τ on product (j, k) at store m during week t is $\hat{Q}_{\tau, (j, k)}^{mt} = Q_{(j, k)}^{mt} \cdot \mathbb{P}^{mt}(\tau \mid A^{mt}, (j, k), \boldsymbol{\nu}^{(q)})$ and $\hat{N}_\tau^{mt} = \sum_{(j, k) \in A^{mt} \cup \{(0, 0)\}} \hat{Q}_{\tau, (j, k)}^{mt}$.

The M Step: The M step is to replace the parameters N_τ^{mt} and $Q_{\tau, (j, k)}^{mt}$ in the complete data log-likelihood $\mathcal{L}_{\text{complete}}$ from Section A.1 by the conditional expected value $\hat{N}_\tau^{mt}$ and $\hat{Q}_{\tau, (j, k)}^{mt}$ obtained in the E step, respectively, and then optimize the log-likelihood. Therefore, $\boldsymbol{\nu}^{(q+1)} = (\boldsymbol{\mu}^{(q+1)}, \mathbf{v}^{(q+1)}, \alpha_0^{(q+1)})$ is updated by $\mu_\tau^{(q+1)} = \sum_{m, t} \left(\hat{N}_{(s, +, \alpha_0)}^{mt} + \hat{N}_{(s, -, \alpha_0)}^{mt} \right) / \left(2 \cdot \sum_{m, t, \tau'} \hat{N}_{\tau'}^{mt} \right)$ if $\tau = (s, +, \alpha_0)$ or $(s, -, \alpha_0)$ and $(\mathbf{v}^{(q+1)}, \alpha_0^{(q+1)})$ is the unique optimizer of P_2^{complete} with a problem instance of $\hat{N}_\tau^{mt}$ and $\{\hat{Q}_{\tau, (j, k)}^{mt}\}$.

The procedure alternates between the E and M steps until the model parameter $\boldsymbol{\nu}^{(q)}$ converges.

A.3. Remarks

The EM algorithm has been used for estimating choice models from data. Examples include the estimation of the LC-MNL model (Train 2009), the general attraction model (GAM) (Gallego et al. 2015), the ranking-based model (van Ryzin and Vulcano 2014), and the decision forest model (Chen and Mišić 2022). In general, the efficiency of the EM algorithm depends on whether one can solve the M step easily under a choice model. For example, in the LC-MNL model, the M step requires to solve K concave maximization problems, where K is the number of the customer types. In the GAM model, the M step cannot be solved as a concave maximization problem. Gallego et al. (2015) thus instead consider minimizing the square error by ignoring the no-purchase option. In the ranking-based model, the M step is amount to solve a linear ordering problem, which is NP-hard, and van Ryzin and Vulcano (2014) solve it by a mixed-integer linear program.

In contrast, the M step for the style-size choice model is surprisingly simple, as it only requires solving a single concave maximization problem P_2^{complete} . This simplicity comes from our model formulation, especially from the design of the consideration sets $C_\tau^1(A)$ and $C_\tau^2(A)$ and the fact that the choice from the two sets can be separated in the log-likelihood function. Furthermore, such structure in the M step still exists even if we generalize the style-size model and incorporate store-specific parameters, such as the store intrinsic utility $(v_m)_{m \in \mathcal{M}}$ and store-dependent best-fit

distribution $(\mu_{m,\tau})_{m \in \mathcal{M}, \tau \in \Gamma}$. These parameters can be useful for designing localized assortments and local inventory levels (Fisher and Vaidyanathan 2014). This highlights the flexibility of the style-size choice model and its EM estimation procedure.

Appendix B: Proofs

B.1. Proof for Lemma 1

We focus on a customer type $(s, +, \alpha)$. The proof for customer type $(s, -, \alpha)$ follows a similar argument. We first prove the necessary condition. That is, we will show that $\mathbb{P}_{(s,+, \alpha)}$ satisfies the substitutability property if $\alpha \geq 0$. We first define functions

$$F_j(A) = \mathbb{I}[(j, s) \in A] + \beta \cdot \mathbb{I}[(j, s) \notin A \text{ and } (j, \text{ADJ}_+(s)) \in A],$$

for all assortment $A \subseteq \mathcal{N}$ and style $j \in \mathcal{J}$, where $\beta = \exp(-\alpha)$. One can easily verify that $F_j(A)$ will not decrease if we add a new product to A as long as $\beta \leq 1$.

Now we will show that the choice probability $\mathbb{P}_{(s,+, \alpha)}((j, k) | A)$ for a given product (j, k) will not increase after adding any new product to assortment A . For simplicity, we write $w_j = \exp(v_j)$ for all $j \in \mathcal{J}$. We consider three cases.

- Case 1: $k = s$. Then we can write

$$\mathbb{P}_{(s,+, \alpha)}((j, s) | A) = \frac{w_j}{1 + w_j + \sum_{i \neq j} w_i F_i(A)}.$$

Then no matter which product we add to A , $\mathbb{P}_{(s,+, \alpha)}((j, k) | A)$ will not increase due to the monotonicity of $F_i(A)$ for all $i \neq j$.

- Case 2: $k = \text{ADJ}_+(s)$. If $(j, s) \in A$, then $\mathbb{P}_{(s,+, \alpha)}((j, k) | A) = 0$ stays as zero no matter what we add to A ; else if $(j, s) \notin A$ and we add (j, s) to A , then $\mathbb{P}_{(s,+, \alpha)}((j, k) | A)$ decreases to zero; else if $(j, s) \notin A$ and we add a product other than (j, s) to A , then $\mathbb{P}_{(s,+, \alpha)}((j, k) | A)$ will not increase, since the denominator in the expression

$$\mathbb{P}_{(s,+, \alpha)}((j, k) | A) = \frac{\beta w_j}{1 + \beta w_j + \sum_{i \neq j} w_i F_i(A)}$$

will not decrease no matter what we add to the assortment.

- Case 3: $k \notin \{s, \text{ADJ}_+(s)\}$. The choice probability $\mathbb{P}_{(s,+, \alpha)}((j, k) | A)$ is always zero and thus will not increase.

For the sufficient condition, it amounts to show that if $\alpha < 0$, then there exists an assortment A such that the choice probability of an option increases as A enlarges. Consider $A = \{(j, \text{ADJ}_+(s))\}$. Then we have

$$\mathbb{P}_{(s,+, \alpha)}((0, 0) | A) = \frac{1}{1 + \beta w_j} < \frac{1}{w_j} = \mathbb{P}_{(s,+, \alpha)}((0, 0) | A \cup \{(j, s)\}),$$

where the inequality holds since $\beta = \exp(-\alpha) > 1$ when $\alpha < 0$ and the assortment A is enlarged by adding product (j, s) . $\square$

B.2. Proof of Theorem 1

The main idea is to show that the optimal revenue $R_\tau(A)$ from each customer type $\tau = (s, \sigma, \alpha)$, where $s \in \mathcal{K}$, $\alpha \geq 0$, and $\sigma \in \{+, -\}$, is upper bounded by the optimal value z_{MNL}^* of the style-MNL assortment optimization problem (10). Therefore, the overall expected revenue would be upper bounded by the same value, i.e.,

$$R(A) = \sum_{s, \sigma} \int_0^\infty \mu_{(s, \sigma, \alpha)} \cdot R_{(s, \sigma, \alpha)}(A) d\alpha \leq \sum_{s, \sigma} \int_0^\infty \mu_{(s, \sigma, \alpha)} \cdot z_{\text{MNL}}^* d\alpha = z_{\text{MNL}}^*.$$

We then show that this upper bound is attached by the revenue-ordered assortment (11) in Theorem 1.

We first focus on the revenue collected from a fixed customer type $\tau = (s, +, \alpha)$ and provide several lemmas around it. We denote $\beta = \exp(-\alpha)$ and $w_j = \exp(v_j)$ for all $j \in \mathcal{J}$ to simplify the notation. Define $\mathcal{N}_s^+ = \{(j, k) \mid j \in \mathcal{J}, k \in \{s, \text{ADJ}_+(s)\}\}$, which is a subset of $\mathcal{N}$ that includes all products of sizes s and $\text{ADJ}_+(s)$. Note that function $R_\tau(A)$ can be written as

$$R_\tau(A) = \frac{\sum_{(j, k) \in C_\tau^1(A)} r_j w_j + \sum_{(j, k) \in C_\tau^2(A)} \beta r_j w_j}{1 + \sum_{(j, k) \in C_\tau^1(A)} w_j + \sum_{(j, k) \in C_\tau^2(A)} \beta w_j}.$$

Notice that $R_\tau(A) = R_\tau(A \cap \mathcal{N}_s^+)$ for any assortment $A \subseteq \mathcal{N}$, since any product of sizes other than s and $\text{ADJ}_+(s)$ will not be considered by customer $\tau = (s, +, \alpha)$ and thus will not contribute to the revenue R_τ . Therefore, to discuss the revenue $R_\tau(A)$, it suffices to only discuss $R_\tau(A)$ for $A \subseteq \mathcal{N}_s^+$.

The following lemma tells us that it is always good to introduce a style of correct size if the style is more profitable than the current assortment.

LEMMA 2. *Consider a fixed customer type $\tau = (s, +, \alpha)$. Suppose $A \subseteq \mathcal{N}_s^+$ and $(i, s) \notin A$ for a style $j \in \mathcal{J}$. If $r_j > R_\tau(A)$, then $R_\tau(A \cup \{(j, s)\}) > R_\tau(A)$.*

Proof: Denote the larger-adjacent size of the customer by $\ell = \text{ADJ}_+(s)$. Let $\mathbb{I}_{(j, \ell) \in A}$ be the indicator that whether the larger-adjacent size ℓ of style j is in the assortment A . We can write the revenue of $R_\tau(A \cup \{(j, s)\})$ as

$$\begin{aligned} R_\tau(A \cup \{(j, s)\}) &= \frac{r_j w_j + \sum_{(i, k) \in C_\tau^1(A): i \neq j} r_i w_i + \sum_{(i, k) \in C_\tau^2(A): i \neq j} \beta r_i w_i}{1 + w_j + \sum_{(i, k) \in C_\tau^1(A): i \neq j} w_i + \sum_{(i, k) \in C_\tau^2(A): i \neq j} \beta w_i} \\ &= \left(\frac{w_j \cdot (1 - \mathbb{I}_{(j, \ell) \in A} \beta)}{1 + w_j + \sum_{(i, k) \in C_\tau^1(A): i \neq j} w_i + \sum_{(i, k) \in C_\tau^2(A): i \neq j} \beta w_i} \right) \cdot r_j + \\ &\quad \left(\frac{\beta r_j w_j \mathbb{I}_{(j, \ell) \in A} + \sum_{(i, k) \in C_\tau^1(A): i \neq j} r_i w_i + \sum_{(i, k) \in C_\tau^2(A): i \neq j} \beta r_i w_i}{1 + w_j + \sum_{(i, k) \in C_\tau^1(A): i \neq j} w_i + \sum_{(i, k) \in C_\tau^2(A): i \neq j} \beta w_i} \right) \\ &= \left(\frac{w_j \cdot (1 - \mathbb{I}_{(j, \ell) \in A} \beta)}{1 + w_j + \sum_{(i, k) \in C_\tau^1(A): i \neq j} w_i + \sum_{(i, k) \in C_\tau^2(A): i \neq j} \beta w_i} \right) \cdot r_j + \\ &\quad \left(\frac{1 + \beta w_j \mathbb{I}_{(j, \ell) \in A} + \sum_{(i, k) \in C_\tau^1(A): i \neq j} w_i + \sum_{(i, k) \in C_\tau^2(A): i \neq j} \beta w_i}{1 + w_j + \sum_{(i, k) \in C_\tau^1(A): i \neq j} w_i + \sum_{(i, k) \in C_\tau^2(A): i \neq j} \beta w_i} \right) \cdot R_\tau(A) \end{aligned}$$

where we note that the revenue function $R_\tau(A)$ can be re-written as

$$R_\tau(A) = \frac{\beta r_j w_j \mathbb{I}_{(j,\ell) \in A} + \sum_{(i,k) \in C_\tau^1(A): i \neq j} r_i w_i + \sum_{(i,k) \in C_\tau^2(A): i \neq j} \beta r_i w_i}{1 + \beta w_j \mathbb{I}_{(j,\ell) \in A} + \sum_{(j,k) \in C_\tau^1(A): i \neq j} w_i + \sum_{(j,k) \in C_\tau^2(A): i \neq j} \beta w_i}$$

Therefore, $R_\tau(A \cup \{(j,s)\})$ is a convex combination of r_j and $R_\tau(A)$. If $r_j > R_\tau(A)$, then $R_\tau(A \cup \{(j,s)\}) > R_\tau(A)$. $\square$

The next lemma tells us that if a product is less profitable than the current assortment, no matter it is of the correct size or of the adjacent size, then we should exclude it from the current assortment.

LEMMA 3. Consider a fixed customer type $\tau = (s, +, \alpha)$. Suppose $(j, k) \in A \subseteq \mathcal{N}_s^+$ for a style $j \in \mathcal{J}$. If $r_j \leq R_\tau(A)$, then $R_\tau(A \setminus \{(j, k)\}) \geq R_\tau(A)$.

Proof: Again, we denote the larger-adjacent size of the customer by $\ell = \text{ADJ}_+(s)$. Let $\mathbb{I}_{(j,\ell) \in A}$ be the indicator that whether the larger-adjacent size ℓ of style j is in the assortment A . We consider two cases.

1. For $k = s$. Similar to the construction in the proof of Lemma 2, we have $R_\tau(A) = \gamma \cdot r_j + (1 - \gamma) \cdot R_\tau(A \setminus \{(j, s)\})$, where

$$\gamma = \frac{w_j \cdot (1 - \mathbb{I}_{(j,\ell) \in A} \beta)}{1 + w_j + \sum_{(i,k) \in C_\tau^1(A): i \neq j} w_i + \sum_{(i,k) \in C_\tau^2(A): i \neq j} \beta w_i}$$

and

$$R_\tau(A \setminus \{(j, s)\}) = \frac{r_j w_j \beta \cdot \mathbb{I}_{(j,\ell) \in A} + \sum_{(i,k) \in C_\tau^1(A): i \neq j} r_i w_i + \sum_{(i,k) \in C_\tau^2(A): i \neq j} \beta r_i w_i}{1 + w_j \beta \cdot \mathbb{I}_{(j,\ell) \in A} + \sum_{(i,k) \in C_\tau^1(A): i \neq j} w_i + \sum_{(i,k) \in C_\tau^2(A): i \neq j} \beta w_i}$$

Therefore,

$$R_\tau(A \setminus \{(j, s)\}) = \frac{R_\tau(A) - \gamma r_j}{1 - \gamma} \geq \frac{R_\tau(A) - \gamma R_\tau(A)}{1 - \gamma} = R_\tau(A).$$

2. For $k = \ell$. Recall that $\ell = \text{ADJ}_+(s)$. If $(j, s) \in A$, then $R_\tau(A \setminus \{(j, \ell)\}) = R_\tau(A)$, as (j, ℓ) are already not considered when (j, s) is available. Now we assume $(j, s) \notin A$, and we have $R_\tau(A) = \gamma' \cdot r_j + (1 - \gamma') \cdot R_\tau(A \setminus \{(j, s)\})$, where

$$\gamma' = \frac{w_j \beta}{1 + w_j \beta + \sum_{(i,k) \in C_\tau^1(A): i \neq j} w_i + \sum_{(i,k) \in C_\tau^2(A): i \neq j} \beta w_i}$$

and

$$R_\tau(A \setminus \{(j, \ell)\}) = \frac{\sum_{(i,k) \in C_\tau^1(A): i \neq j} r_i w_i + \sum_{(i,k) \in C_\tau^2(A): i \neq j} \beta r_i w_i}{1 + \sum_{(i,k) \in C_\tau^1(A): i \neq j} w_i + \sum_{(i,k) \in C_\tau^2(A): i \neq j} \beta w_i}$$

Therefore,

$$R_\tau(A \setminus \{(j, \ell)\}) = \frac{R_\tau(A) - \gamma' r_j}{1 - \gamma'} \geq \frac{R_\tau(A) - \gamma' R_\tau(A)}{1 - \gamma'} = R_\tau(A).$$

□

The next lemma shows that given a customer type $\tau = (s, +, \alpha)$, its expected revenue $R_\tau(A)$ is upper bounded by z_{MNL}^* and the upper bound is attached by a revenue-ordered assortment of products of the customer's best-fit size s .

LEMMA 4. Consider a customer type $\tau = (s, +, \alpha)$. Denote $z^* \equiv \max_{A \subseteq \mathcal{N}_s^+} R_\tau(A)$. Then $A_o = \{(j, k) \mid r_j > z^*, j \in \mathcal{J}\}$ is an optimal solution to the problem $\max_{A \subseteq \mathcal{N}_s^+} R_\tau(A)$. In addition,

$$z^* = z_{MNL}^* \equiv \max_{j \in \mathcal{J}} \left\{ \frac{\sum_{i=1}^j r_i w_i}{1 + \sum_{i=1}^j w_i} \right\}.$$

Proof: Obviously, z^* exists and it is finite since $\mathcal{N}_s^+$ is a finite set. We prove the first part of the statement by contradiction. Suppose A_o is not an optimal solution and let A is an optimal solution with smallest cardinality. The fact that A is optimal and $z^* = R_\tau(A) > R_\tau(A_o)$ imply that one of the following statements must be true: (i) there exists a style j such that $r_j > z^*$ and $(j, s) \notin A$; and (ii) there exists a product $(j, k) \in A$ such that $r_j \leq z^*$ and $k \in \{s, \text{ADJ}_+(s)\}$. Otherwise, if none of them is true, then $(j, k) \in A$ for all $j \in \mathcal{J}$ satisfying $r_j > z^*$ and $(j, k) \notin A$ for all j satisfying $r_j \leq z^*$ and all $k \in \{s, \text{ADJ}_+(s)\}$. One can then easily verify that $R_\tau(A) = R_\tau(A_o)$, which is a contradiction (that is to say, if none of (i) and (ii) is true, then A and A_o would be only different from each other for size $\text{ADJ}_+(s)$ of styles $j \in \{j \mid r_j > z^*\}$. Given that the correct size s of these styles are already in both A and A_o , these products of the larger-adjacent size do not change the expected revenue of A from A_o . That means $R_\tau(A) = R_\tau(A_o)$, a contradiction.)

Now we know one of the statements (i) and (ii) about A must be true. However, if (i) is true, we can conclude that

$$R_\tau(A \cup \{(j, s)\}) > R_\tau(A)$$

by Lemma 2, which contradicts the fact that A is an optimal solution. If (ii) is true, we can conclude from Lemma 3 that the assortment $A \setminus \{(j, k)\}$ has a no-worse revenue

$$R_\tau(A \setminus \{(j, k)\}) \geq R_\tau(A)$$

but has a smaller cardinality than A , which would contradicts the fact that A is an optimal assortment with smallest cardinality. Therefore, neither (i) nor (ii) is true, which leads to a contradiction. Thus, A_o is an optimal solution.

For the second part of the theorem, we first notice that $A_o \in \mathcal{A}_{\text{order}}$, where $\mathcal{A}_{\text{order}}$ is the collection of all revenue-ordered assortments that only consist of products of size s

$$\mathcal{A}_{\text{order}} = \{A_o^j \mid j \in \mathcal{J}\}, \text{ where we define } A_o^j \equiv \{(1, k), (2, k), \dots, (j, k)\}.$$

For each revenue-ordered assortment A_o^j , we have $R_\tau(A_o^j) = \sum_{i=1}^j r_i w_i / (1 + \sum_{i=1}^j w_i)$. Therefore,

$$z^* = \max_{A \subseteq \mathcal{N}_k^+} R_\tau(A) = \max_{A_o \in \mathcal{A}_{\text{order}}} R_\tau(A_o) = \max_{j \in \mathcal{J}} \left\{ \frac{\sum_{i=1}^j r_i w_i}{1 + \sum_{i=1}^j w_i} \right\},$$

where the second equality follows the first part of the theorem that we just proved. $\square$

We note that Lemma 4 holds for any other customer types, as all the arguments in Lemmas 2 and 3 can easily follow for customer types in form of $\tau = (s, \sigma, \alpha)$. In other words, $\max_A R_\tau(A) = z_{\text{MNL}}^*$ for any $\tau = (s, \sigma, \alpha)$ where $s \in \mathcal{K}$, $\alpha \geq 0$, and $\sigma \in \{+, -\}$. It also holds when the best-fit size s is a boundary size of $\mathcal{K}$. For example, if $k_{\max}$ is the maximal size among $\mathcal{K}$, then the corresponding customer type (s, σ, α) for $s = k_{\max}$ behaves like a classic MNL model over products $\{(j, k_{\max}) \mid j \in \mathcal{J}\}$, as there is no larger-adjacent size to substitute to, again implying that $\max_A R_\tau(A) = z_{\text{MNL}}^*$. Applying Lemma 4 to all customer types $\tau = (s, \sigma, \alpha)$, we simply prove Theorem 1 as follows.

Proof of Theorem 1: By Lemma 4 and the discussion above, we know that $R_\tau(A) \leq z_{\text{MNL}}^*$ for all customer types $\tau = (s, \sigma, \alpha)$. Therefore, $R(A) = \sum_{s, \sigma} \int_0^\infty \mu_{(s, \sigma, \alpha)} \cdot R_{(s, \sigma, \alpha)}(A) d\alpha \leq \sum_{s, \sigma} \int_0^\infty \mu_{(s, \sigma, \alpha)} \cdot z_{\text{MNL}}^* d\alpha = z_{\text{MNL}}^*$. On the other hand, one can easily verify that $R(A^*) = z_{\text{MNL}}^*$ for the assortment defined in Equation (11). Therefore, A^* is an optimal solution. $\square$

Lastly, we remark that our proof of Theorem 1 follow a first-principle argument to determine whether we can further improve the expected revenue by adding or removing products from the assortment. The same proof technique was used by Rusmevichientong and Topaloglu (2012) to show that the robust assortment optimization under the MNL model has a revenue-ordered structure.

B.3. Proof of Proposition 1

Our proof closely follows the argument in Zhang et al. (2024); see the proof of Theorem 3.1 in the paper. To simplify the expression, we label products in $\mathcal{N}$ as $\{1, 2, \dots, n\}$. There, $n = JK$ when there are J styles and K sizes. We consider an inventory problem under the following assumption: *Customers choose product $i \in \mathcal{N}$ according to the initial set $S_0 = \{i \mid I_i \geq 1\}$ of available products. If the product they choose is out of stock, then they leave without a purchase.* We call this optimization problem P_{static} . Given a inventory vector $\mathbf{I}$, the profit of the inventory model is

$$\Pi_{\text{static}}(\mathbf{I}) = \sum_{i \in \mathcal{N}} p_i \cdot \mathbb{E} \left\{ \min \left\{ I_i, \sum_{\ell=1}^L C_{i\ell}(\mathbf{I}^{\text{FA}}) \right\} \right\} - \sum_{i \in \mathcal{N}} c_i I_i \quad (21)$$

Here, L is a random variable that represents the number of customer visiting in period $[0, T]$ and $C_{i\ell}(\mathbf{I})$ is the indicator that whether customer ℓ would choose product i from S_0 . If the underlying choice model $\mathbb{P}(\cdot \mid \cdot)$ is a substitutive model, then the profit Π_{static} of this inventory problem is a lower bound to the original dynamic inventory problem. This because if a product is out of stock,

then the demand of other available products should increase (or stay the same) in the dynamic inventory model. However, in Problem P_{static} , we assume that the demand of other products remains the same. That implies we underestimate the revenue collected after the stock-out happens in Problem P_{static} , resulting a lower bound to the dynamic problem. From here, we can also see why the same argument does not apply to non-substitutive choice models. In non-substitutive models, other products' demand could shrink to a lower value after each stock-out, and the objective Π_{static} is thus no longer a lower bound.

Define $\pi_i = \mathbb{P}(i \mid A^*)$, the choice probability of product i under the optimal assortment A^* . We consider bounding the gap between $\Pi_{\text{static}}(\mathbf{I})$ with $\mathbf{I} = \mathbf{I}^{\text{FA}} \equiv (\lceil T\lambda\pi_i \rceil)_{i \in \mathcal{N}}$ and $V_{\text{fluid}} = T\lambda \sum_{i=1}^n (p_i - c_i)\pi_i$. For simplicity, we call $\mathbf{I}^{\text{FL}} = (T\lambda\pi_i)_{i \in \mathcal{N}}$, which is a vector that consists of fractional numbers.

$$\begin{aligned} V_{\text{fluid}} - \Pi_{\text{static}}(\mathbf{I}^{\text{FA}}) &= \sum_{i \in \mathcal{N}} (p_i - c_i) I_i^{\text{FL}} - \sum_{i \in \mathcal{N}} p_i \cdot \mathbb{E} \left\{ \min \left\{ I_i^{\text{FA}}, \sum_{\ell=1}^L C_{i\ell}(\mathbf{I}^{\text{FA}}) \right\} \right\} + \sum_{i \in \mathcal{N}} c_i I_i^{\text{FA}} \\ &= \sum_{i \in A^*} p_i \cdot \mathbb{E} \left\{ I_i^{\text{FL}} - \min \left\{ I_i^{\text{FA}}, \sum_{\ell=1}^L C_{i\ell}(\mathbf{I}^{\text{FA}}) \right\} \right\} + \sum_{i \in \mathcal{N}} c_i \cdot (I_i^{\text{FA}} - I_i^{\text{FL}}) \\ &\leq \sum_{i \in A^*} p_i \cdot \mathbb{E} \left\{ I_i^{\text{FL}} - \min \left\{ I_i^{\text{FL}}, \sum_{\ell=1}^L C_{i\ell}(\mathbf{I}^{\text{FA}}) \right\} \right\} + \sum_{i \in A^*} c_i, \end{aligned}$$

where in the last inequality we use the fact that $I_i^{\text{FL}} \leq I_i^{\text{FA}} < I_i^{\text{FL}} + 1$. The second term is upper bounded by $|A^*| \cdot c_{\max}$. The first term can be bounded as follows.

$$\begin{aligned} \sum_{i \in A^*} p_i \cdot \mathbb{E} \left\{ I_i^{\text{FL}} - \min \left\{ I_i^{\text{FL}}, \sum_{\ell=1}^L C_{i\ell}(\mathbf{I}^{\text{FA}}) \right\} \right\} &\leq \sum_{i \in A^*} p_i \cdot \mathbb{E} \left\{ \left| I_i^{\text{FL}} - \sum_{\ell=1}^L C_{i\ell}(\mathbf{I}^{\text{FA}}) \right| \right\} \\ &\leq \sum_{i \in A^*} p_i \cdot \sqrt{\mathbb{E} \left\{ \left(I_i^{\text{FL}} - \sum_{\ell=1}^L C_{i\ell}(\mathbf{I}^{\text{FA}}) \right)^2 \right\}} \end{aligned}$$

Notice that the random variable $\sum_{\ell=1}^L C_{i\ell}(\mathbf{I}^{\text{FA}})$ has expectation

$$\mathbb{E} \left[\sum_{\ell=1}^L C_{i\ell}(\mathbf{I}^{\text{FA}}) \right] = \mathbb{E}_L \left[\sum_{\ell=1}^L \mathbb{E}[C_{i\ell}(\mathbf{I}^{\text{FA}})] \right] = \mathbb{E}_L \left[\sum_{\ell=1}^L \pi_i \right] = \mathbb{E}_L [L\pi_i] = T\lambda\pi_i = I_i^{\text{FL}}$$

Therefore,

$$\begin{aligned} \sum_{i \in A^*} p_i \cdot \mathbb{E} \left\{ I_i^{\text{FL}} - \min \left\{ I_i^{\text{FL}}, \sum_{\ell=1}^L C_{i\ell}(\mathbf{I}^{\text{FA}}) \right\} \right\} &\leq \sum_{i \in A^*} p_i \cdot \sqrt{\text{Var} \left\{ \sum_{\ell=1}^L C_{i\ell}(\mathbf{I}^{\text{FA}}) \right\}} \\ &= \sum_{i \in A^*} p_i \cdot \sqrt{\mathbb{E}_L \left[\text{Var} \left\{ \sum_{\ell=1}^L C_{i\ell}(\mathbf{I}^{\text{FA}}) \mid L \right\} \right]} \\ &= \sum_{i \in A^*} p_i \cdot \sqrt{\mathbb{E}_L [L\pi_i(1 - \pi_i)]} \\ &= \sum_{i \in A^*} p_i \cdot \sqrt{T\lambda\pi_i(1 - \pi_i)} \leq p_{\max} \sqrt{T\lambda} \cdot \sum_{i \in A^*} \sqrt{\pi_i} \leq p_{\max} \sqrt{T\lambda} \sqrt{|A^*|}, \end{aligned}$$

where the last step is obtained by Cauchy-Schwartz inequality $(\sum_{i \in A^*} \sqrt{\pi})^2 \leq (\sum_{i \in A^*} 1) \cdot (\sum_{i \in A^*} \pi) = |A^*|$. Therefore, $V_{\text{fluid}} - \Pi_{\text{static}}(\mathbf{I}) \leq p_{\max} \sqrt{T\lambda n} + c_{\max} n$. Finally, we note that $\Pi^* \leq V_{\text{fluid}}$ according to Lemma (6), which is introduced below. Also, $\Pi(\mathbf{I}^{\text{FA}}) \geq \Pi_{\text{static}}(\mathbf{I}^{\text{FA}})$. Thus, $\Pi^* - \Pi(\mathbf{I}^{\text{FA}}) \leq V_{\text{fluid}} - \Pi_{\text{static}}(\mathbf{I}^{\text{FA}}) = O(\sqrt{nT\lambda})$. $\square$

B.4. Proof of Proposition 2

Define $\text{supp}(\mathbf{I})$ as the support of an inventory vector $\mathbf{I}$, i.e., $\text{supp}(\mathbf{I}) = \{(j, k) \mid I_{jk} > 0\}$. We further define $\mathcal{C}(A)$ as the class of inventory vectors with support S , i.e., $\mathcal{C}(A) = \{\mathbf{I} \in \mathbb{N}_+^{JK} \mid \text{supp}(\mathbf{I}) = A\}$. We will first show that when $\bar{L}$ is sufficiently large, any inventory vector from class $\mathcal{C}(A')$ for $A' \neq A^*$, where A^* is the optimal assortment, cannot be an optimal solution to Problem P_{LB} . In particular, for any $\mathbf{I} \in \cup_{A \neq A^*} \mathcal{C}(A)$, we have

$$\begin{aligned} z'_{\text{LB}} &= \max_{A \neq A^*, \mathbf{I} \in \mathcal{C}(A), \mathbf{I} \in \mathbb{N}_+^{JK}} \left[\sum_{(j,k)} p_j \cdot \min \{ \bar{L} \cdot \pi_{jk}(\mathbf{I}), I_{jk} \} - \sum_{(j,k)} c_j \cdot I_{jk} \right] \\ &\leq \max_{A \neq A^*, \mathbf{I} \in \mathcal{C}(A), \mathbf{I} \in \mathbb{N}_+^{JK}} \left[\sum_{(j,k)} (p_j - c_j) \cdot \min \{ \bar{L} \cdot \pi_{jk}(\mathbf{I}), I_{jk} \} \right] \\ &\leq \max_{A \neq A^*, \mathbf{I} \in \mathcal{C}(A), \mathbf{I} \in \mathbb{N}_+^{JK}} \left[\sum_{(j,k)} (p_j - c_j) \cdot \bar{L} \cdot \pi_{jk}(\mathbf{I}) \right] \leq \bar{L} \cdot \max_{A \neq A^*} \sum_{(j,k)} (p_j - c_j) \cdot \mathbb{P}((j, k) \mid A). \end{aligned}$$

Meanwhile, we consider the objective value of fluid policy $\mathbf{I}^{\text{FA}}$ in Problem P_{LB} as follows

$$\begin{aligned} z_{\text{LB}}(\mathbf{I}^{\text{FA}}) &= \sum_{(j,k) \in \mathcal{N}} p_j \cdot \bar{L} \cdot \mathbb{P}((j, k) \mid A^*) - \sum_{(j,k) \in \mathcal{N}} c_j \cdot \left[\bar{L} \cdot \mathbb{P}((j, k) \mid A^*) + (\lceil \bar{L} \cdot \mathbb{P}((j, k) \mid A^*) \rceil - \bar{L} \cdot \mathbb{P}((j, k) \mid A^*)) \right] \\ &\geq \sum_{(j,k) \in \mathcal{N}} (p_j - c_j) \cdot \bar{L} \cdot \mathbb{P}((j, k) \mid A^*) - \sum_{(j,k) \in \mathcal{N}} c_j = \bar{L} R_{\text{asst}}(A^*) - \sum_{(j,k) \in \mathcal{N}} c_j. \end{aligned}$$

Therefore, for sufficiently large $\bar{L}$, we have $z'_{\text{LB}} < z_{\text{LB}}(\mathbf{I}^{\text{FA}})$, which implies that the support of the optimal solution of Problem P_{LB} must be A^* when $\bar{L}$ is sufficiently large.

Now we shall show that $I_{jk}^{\text{IP}}/I_{jk}^{\text{FA}} \rightarrow 1$ for all $(j, k) \in A^*$ when $\bar{L} \rightarrow \infty$. This is equivalent to show that $I_{jk}^{\text{IP}}/(\bar{L} \cdot \mathbb{P}((j, k) \mid A^*)) \rightarrow 1$ for all $(j, k) \in A^*$. Consider a sufficiently large $\bar{L}$. As $\mathbf{I}^{\text{IP}}$ returns the optimal solution to P_{LB} , we know that $\mathbf{I}^{\text{IP}}$ has support A^* . Assume there exists a pair (j, k) such that $\liminf_{\bar{L}} |I_{jk}^{\text{IP}}/(\bar{L} \cdot \mathbb{P}((j, k) \mid A^*)) - 1| > \epsilon$ for a constant ϵ . Then, as $\bar{L} \rightarrow \infty$, either $\liminf \{ (z_{\text{LB}}(\mathbf{I}^{\text{FA}}) - z_{\text{LB}}(\mathbf{I}^{\text{IP}})/\bar{L}) \} > \epsilon(p_j - c_j)\mathbb{P}((j, k) \mid A^*) > 0$ or $\liminf \{ (z_{\text{LB}}(\mathbf{I}^{\text{FA}}) - z_{\text{LB}}(\mathbf{I}^{\text{IP}})/\bar{L}) \} > \epsilon c_j \mathbb{P}((j, k) \mid A^*) > 0$ holds. In both cases, it contradicts the fact that $\mathbf{I}^{\text{IP}}$ maximizes $z_{\text{LB}}(\cdot)$. Therefore, $I_{jk}^{\text{IP}}/I_{jk}^{\text{FA}} \rightarrow 1$ for all $(j, k) \in A^*$ when $\bar{L} \rightarrow \infty$. $\square$

Appendix C: Comparing the Style-Size Choice Model and Nested Logit Model

To compare the style-size model with the nested logit model, we examine two variants of the nested logit framework that incorporate the structure of apparel styles and sizes. Figure 4 illustrates these two configurations. In the left panel, nests (or baskets) are defined by apparel styles, while in the

right panel, nests are organized by apparel sizes. The size-basket variant (right panel) serves as a benchmark in Section 4.3, as it includes $|\mathcal{J}| + |\mathcal{K}| + 1$ parameters, making it comparable to the style-size choice model and the size aggregation approach. In contrast, the style-based variant (left panel) has $2|\mathcal{J}| + 1$ parameters. We will delve into the details of parameter counts for each variant shortly. Next, we demonstrate that both variants of the nested logit model result in unrealistic demand substitution within the context of the apparel industry. This highlights a key distinction between the nested logit models and the style-size choice model proposed in this paper.

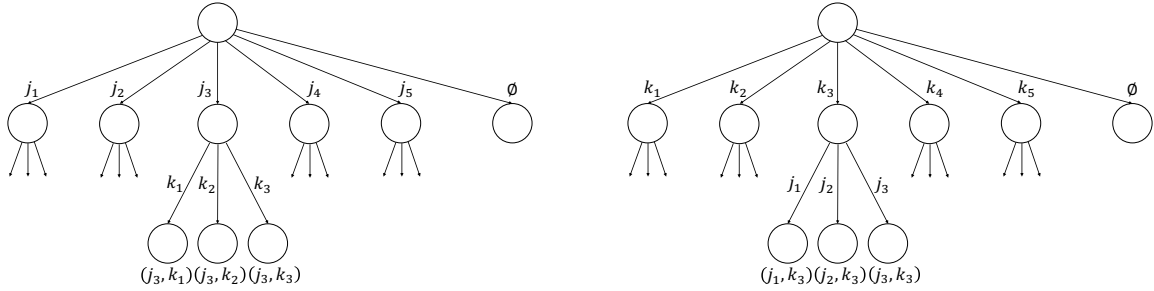


Figure 4 Two variants of the nested logit model that encode the apparel style and size structure.

Let us first consider the variant of the nested logit model where each nest is defined with respect to style, i.e., the model in Figure 4 (left). The model has $2|\mathcal{J}| + 1$ parameters. The first $|\mathcal{J}|$ parameters correspond to the utility parameters v_j for each style $j \in \mathcal{J}$. The second $|\mathcal{J}|$ parameters represent the similarity parameters $\eta_j \in (0, 1]$ associated with each nest defined for $j \in \mathcal{J}$. The final parameter, v_0 , captures the utility of no-purchase option. Unlike the MNL, mixed-MNL, and the style-size choice models, the presence of the similarity parameters $(\eta_j)_{j \in \mathcal{J}}$ prevents us from rescaling the utility of each style relative to v_0 via $v_j - v_0$ to eliminate v_0 as a parameter.

Consider the following toy example. Suppose a store sells T-shirts in five sizes, $\mathcal{K} = \{XS, S, M, L, XL\}$, where we let $k_1 = XS$ and $k_5 = XL$. Now, compare two assortments $A_1 = \{(j_1, k_1)\}$ and $A_2 = \{(j_1, k_1), (j_1, k_5)\}$. In assortment A_2 , an additional T-shirt of the same style but in size XL is available compared to assortment A_1 . The choice probability (i.e., the demand) of product (j_1, k_1) given assortment A_1 under the nested logit model is

$$\mathbb{P}^{\text{NL}}((j_1, k_1) | A_1) = \frac{(e^{v_{j_1}})^{\eta_1}}{e^{v_0} + (e^{v_{j_1}})^{\eta_1}}.$$

Now we introduce product (j_1, k_5) to the assortment A_1 , resulting assortment A_2 . The choice probability of product (j_1, k_1) follows as

$$\mathbb{P}^{\text{NL}}((j_1, k_1) | A_2) = \frac{(e^{v_{j_1}} + e^{v_{j_1}})^{\eta_1}}{e^{v_0} + (e^{v_{j_1}} + e^{v_{j_1}})^{\eta_1}} \cdot \frac{e^{v_{j_1}}}{e^{v_{j_1}} + e^{v_{j_1}}} = \frac{2^{\eta_1} \cdot (e^{v_{j_1}})^{\eta_1}}{e^{v_0} + 2^{\eta_1} \cdot (e^{v_{j_1}})^{\eta_1}} \cdot \frac{1}{2}.$$

Since $2^{\eta_1} \in (1, 2]$, we have

$$\mathbb{P}^{\text{NL}}((j_1, k_1) | A_2) = \frac{2^{\eta_1} \cdot (e^{v_{j_1}})^{\eta_1}}{e^{v_0} + 2^{\eta_1} \cdot (e^{v_{j_1}})^{\eta_1}} \cdot \frac{1}{2} < \frac{2 \cdot (e^{v_{j_1}})^{\eta_1}}{e^{v_0} + (e^{v_{j_1}})^{\eta_1}} \cdot \frac{1}{2} = \mathbb{P}^{\text{NL}}((j_1, k_1) | A_1).$$

Therefore, it implies that, under the nested logit model, introducing a T-shirt in size XL would reduce the demand for the size XS of the same style. However, this is unrealistic, as customers who wear size XL T-shirts are unlikely to consider purchasing size XS.

Now, let us consider the second variant of the nested logit model, illustrated in Figure 4(right). In this model, each size corresponds to a nest, with parameters $(\eta_k)_{k \in \mathcal{K}}$. Therefore, there are $|\mathcal{J}| + |\mathcal{K}| + 1$ parameters. Following the same setup for apparel products and assortments in the toy example, we have

$$\mathbb{P}^{\text{NL}}((j_1, k_1) | A_2) = \frac{(e^{v_{j_1}})^{\eta_1}}{e^{v_0} + (e^{v_{j_1}})^{\eta_1} + (e^{v_{j_1}})^{\eta_5}} < \frac{(e^{v_{j_1}})^{\eta_1}}{e^{v_0} + (e^{v_{j_1}})^{\eta_1}} = \mathbb{P}^{\text{NL}}((j_1, k_1) | A_1).$$

Consequently, the T-shirt in size XL once again reduces the demand for the size XS T-shirt of the same style, which is unrealistic.

Finally, it is easy to verify that in the proposed style-size choice model, we have $\mathbb{P}((j_1, k_1) | A_2) = \mathbb{P}((j_1, k_1) | A_1)$, implying the demand of T-shirts of size XS and XL will not cannibalize each other. This highlights the difference between the proposed model and nested logit model.

Appendix D: Additional Numerical Results on Performance of Inventory Policies

D.1. Runtime of the IP-based Policy

Table 4 reports the runtime of optimally solving the MILP (19) in each instance listed in Table 3. Across all instances, the runtime remains under five minutes on a MacBook Pro with an Apple M2 chip. The table also shows that as α_0 decreases (i.e., $\beta_0 = e^{-\alpha_0}$ increases), the runtime for optimally solving the MILP increases. This trend arises because stronger size substitution (lower α_0) allows the firm to leverage demand spillovers across sizes to better meet customer needs. Consequently, the inventory optimization becomes more complex, leading to longer runtimes.

D.2. Performance of the IP-based Policy under the Major-minor Size Constraint

In the following table, we report the expected profit per customer $\Pi_{\text{BT-per}}(\cdot)$ to the casual booties sector for each inventory policy. All notations follow Table 3, except that in each sub-table, we include the performance of the IP-based policy that enforces the major-minor size constraint described in Section 5.3. We call the resulting inventory vector as $\mathbf{I}^{\text{IP}^2}$.

We have the following observations from the table. First, the major-minor size constraint affects the performance of the IP-policy more severely when β is larger, i.e., when the size substitution is more prevalent. This is expected, since the constraint restricts how the IP can utilize the size

$\beta_0 = 24.9\%$		$\beta_0 = 100.0\%$	
$\bar{L}$	T	$\bar{L}$	T
4W	0.1	4W	0.1
8W	0.2	8W	0.4
12W	0.4	12W	172.3
16W	0.6	16W	297.1
20W	0.6	20W	71.1
24W	42.0	24W	28.0
32W	11.3	32W	156.2

Table 4 The runtime T (sec) for solving the mixed-integer program.

$\bar{L}$	$\mathbf{I}^{\text{NV}}$	$\mathbf{I}^{\text{FA}}$	$\mathbf{I}^{\text{IP}}$	$\mathbf{I}^{\text{IP2}}$	$\bar{L}$	$\mathbf{I}^{\text{NV}}$	$\mathbf{I}^{\text{FA}}$	$\mathbf{I}^{\text{IP}}$	$\mathbf{I}^{\text{IP2}}$
4W	-18.88	-18.65	0.11	0.03	4W	-18.88	-18.65	0.29	0.22
8W	-4.45	-3.89	0.79	0.41	8W	-3.07	-2.39	2.07	0.67
12W	-0.16	0.53	1.72	1.49	12W	1.57	2.34	3.74	2.57
16W	1.30	2.09	2.49	2.44	16W	2.50	3.23	4.13	3.52
20W	2.71	3.28	3.42	3.42	20W	3.51	4.16	4.69	4.30
24W	3.02	3.53	3.59	3.59	24W	3.93	4.68	5.05	4.87
32W	3.71	4.31	4.34	4.34	32W	4.98	5.54	5.67	5.67

Table 5 Expected profit per customer of each inventory policy for a given $T\lambda$ ($\beta_0 = 24.9\%$ in the left panel and $\beta_0 = 100.0\%$ in the right panel)

substitution. Meanwhile, when the number of customer visits is sufficiently large, this constraint does not impact the performance of the IP-based policy, as we can see in the case of $\bar{L} = 32W$. This is because the major-minor size constraint no longer alters the optimal solution of the original integer program.

D.3. Performance on a Synthetic Setup: Don't ignore both style and size substitutions

In this section, we consider a toy model to show that the newsvendor policy could have poor performance as it fails to consider style substitution. Notice that in Section 5.3, the attraction of products are low since less than 1% of customers make a purchase from the casual booties. The resulting optimal assortment is to offer all styles there, and the difference between the newsvendor policy and the fluid approximation is only at whether to include a safety stock. In the following toy model, each product' attraction is higher than the ones we considered in Section 5.3 and thus the optimal strategy is not always to offer all styles.

Specifically, we assume that each style $j \in \mathcal{J}$ has the attraction $w_j \sim U([0, 2])$ and price $p_j \sim U([0, 100])$, where U is the uniform distribution, with 100% markup pricing scheme. Note that with such markup scheme, we can isolate the performance of the newsvendor model from its safety stock strategy. For simplicity, we set the size distribution μ as a uniform distribution. We set both $|\mathcal{J}|$ and $|\mathcal{K}|$ to be five and set $\beta_0 = 24.9\%$, which is the size substitution parameter we estimated in

Section 4. We consider the following quantity, $G(\mathbf{I}) \equiv (LR_{\text{asst}}^* - \Pi(\mathbf{I})) / (LR_{\text{asst}}^*)$, which is an upper bound on the optimality gap, where R_{asst}^* is defined in Lemma 6. The same lemma shows that LR_{asst}^* is an upper bound to the inventory problem (12).

We present the performance of the three inventory policies by showing their G value in Table 6. The results are quite consistent with what we have observed in Section 5.3 in terms of the relative performance of the policies. Particularly, the IP-based policy has the best performance and the fluid approximation catches up as there are more customers. In the table, we further observe that these two policies can reach a small optimality gap, less than one percent, as the total number of customer visits increases. Different from Section 5.3, the newsvendor policy has a much worse performance compared to the other two policies since it fails to account for substitution between styles. In particular, it cannot narrow the optimality gap below twenty percent even when the other two policies can reach small gaps. This demonstrates that while it is not fatal to ignore the size substitution as the fluid approximation, ignoring both the size and style substitutions, like the newsvendor policy, could be catastrophic and result in poor performance if customers follow the style-size choice model to make decisions.

L	$\mathbf{I}^{\text{NV}}$	$\mathbf{I}^{\text{FA}}$	$\mathbf{I}^{\text{IP}}$
50	37.72%	16.86%	11.04%
100	27.71%	11.20%	9.91%
200	24.64%	4.97%	3.17%
400	27.19%	1.98%	1.79%
800	19.57%	1.27%	1.21%
1600	22.14%	0.54%	0.52%

Table 6 Bound on optimality gap for each inventory policy

Appendix E: Asymptotic Performance of Fluid Policies under General Choice Models

We end this paper by discussing an extension result for the asymptotic performance of fluid-like inventory policies under general choice models. The literature has mainly focused on choice models that satisfy the substitutability property (Definition 1). For example, Honhon and Seshadri (2013) show that if the underlying choice model is a ranking-based model, a fluid-like approximation solved by a dynamic program proposed by Honhon et al. (2010) can have an $O(n\sqrt{Q})$ optimality gap, where n is the number of products and Q is the total order quantity over these n products. El Housni et al. (2021) achieve an $O\left(n + \sqrt{nL_D \log(nL_D)}\right)$ gap using fluid approximation and sample-average approximation, where L_D is the deterministic number of customer visits. Zhang et al. (2024) improve the optimality gap to $O(\sqrt{nL_D})$ by exploring the gap between the full relaxation upper bound and a lower bound like Problem (13). Given the emerging literature on general choice

models that do not satisfy the substitutability property, such as tree-based models (Akchen and Mišić 2021, Chen and Mišić 2022, Chen et al. 2019) and models inspired by behavioral economics (Maragheh et al. 2018), providing an encompassing performance guarantee can be valuable. We state our result as follows for a general choice model and then we discuss its application to our style-size choice model.

PROPOSITION 3. *Let $P(\cdot | \cdot)$ be any choice model over n products. Assume that $p_{\max} = \max_{i=1, \dots, n} p_i$ and $c_{\max} = \max_{i=1, \dots, n} c_i$ are independent of $\bar{L} = T\lambda$ and n . Let A^* be the optimal assortment and define $\pi_i = \mathbb{P}(i | A^*)$. Consider the inventory policy $\mathbf{I} = (\lceil \bar{L}(\pi_i + \epsilon) \rceil \cdot \mathbb{I}_{i \in A^*})_{i=1, \dots, n}$, where*

$$\epsilon = \frac{1}{2} \cdot \sqrt{\frac{\log(\tilde{L})}{\tilde{L}} \cdot \left(1 - 2\sqrt{\frac{e \log(\tilde{L})}{\tilde{L}} - \frac{1}{\tilde{L}}}\right)^{-1}}$$

with $\tilde{L} = \max\{\bar{L}, e^4\}$. Then the policy $\mathbf{I}$ in the stockout-based inventory problem (12) has an $O(n\sqrt{\bar{L} \log \bar{L}})$ optimality gap and thus it is asymptotically optimal.

We prove the proposition by recognizing that the first stockout is a stopping time and quantifying the revenue collected until the point of the first stockout through a series of concentration inequalities (Vershynin 2018). Compared to the fluid approximation, the inventory policy in Proposition 3 introduces a safety stock $\bar{L}\epsilon = O\left(\sqrt{\bar{L} \cdot \log(\bar{L})}\right)$, which prevents the stockouts from happening too early. Asymptotically, this safety stock is negligible compared to $\bar{L}\pi_i$, making the inventory policy in Proposition 3 converge to the fluid approximation as $\bar{L}$ tends to infinity. When applying this policy to the style-size choice model (5), we again obtain a size-substitution-invariant inventory policy that is asymptotically optimal, although the theoretical optimality gap is larger than the gap of the fluid approximation shown in Proposition 1. On the other hand, in contrast to Proposition 1, the result in Proposition 3 is applicable to the case when the substitutability property does not hold. One of such examples include the scenario that some customers' utility discount α is negative, i.e., there exists a customer type $\tau = (s, \sigma, \alpha)$ such that $\alpha < 0$ with nonzero weight $\mu_\tau > 0$.

E.1. Proof of Proposition 3

We define $[n] \equiv \{1, 2, \dots, n\}$ throughout the proof. We first state two lemmas. The first lemma, Lemma 5, concerns the first stockout time for a specifically constructed inventory vector.

LEMMA 5. *Assume ϵ and ϵ' are two positive constants. Let l be an integer that satisfies $l \in (T(\lambda - \epsilon'), T(\lambda + \epsilon'))$. Let $\{X_\ell\}$ be a sequence of IID multinomial random variables such that $\mathbb{P}(X_\ell = m) = \nu_m$ for $m = 0, 1, \dots, M$, where $\sum_{m=0}^M \nu_m = 1$. For $m = 1, \dots, M$, we denote $Z_m^\ell = \sum_{i=1}^\ell \mathbb{I}[X_i = m]$ as the recurrence of outcome m up to random variable X_ℓ and let $U_m := \lceil T\lambda(\nu_m + \epsilon) \rceil$. We define*

$$\tau = \inf\{\ell \mid \exists m \in \{1, \dots, M\} \text{ such that } Z_m^\ell \geq U_m\}$$

as the first time that one of the Z_m^ℓ hits the corresponding bound U_m . Then, we have

$$\mathbb{P}[\tau \leq \lfloor l - T\epsilon' \rfloor] \leq M \cdot \exp(-2 \cdot (T\lambda - 2T\epsilon' - 1) \cdot \epsilon^2).$$

Proof: Define $l' = \lfloor l - T\epsilon' \rfloor$. Event $\{\tau \leq l'\}$ is equivalent to event $\{\exists m \in \{1, \dots, M\} \text{ such that } Z_m^{l'} \geq U_m\}$. Therefore, by union bound,

$$\mathbb{P}[\tau \leq l'] \leq \sum_{m=1}^M \mathbb{P}\left[Z_m^{l'} \geq U_m := \lceil T\lambda(\nu_m + \epsilon) \rceil\right].$$

On the other hand, we know that $Z_m^{l'} \sim B(l', \nu_m)$, a binomial distribution of l' trials with ν_m success rate. Therefore, by Hoeffding's inequality,

$$\begin{aligned} \mathbb{P}\left[Z_m^{l'} \geq \lceil T\lambda(\nu_m + \epsilon) \rceil\right] &\leq \exp\left(-2 \cdot l' \cdot \left(\nu_m - \frac{\lceil T\lambda(\nu_m + \epsilon) \rceil}{l'}\right)^2\right) \\ &\leq \exp\left(-2 \cdot l' \cdot \left(\nu_m - \frac{\lceil T\lambda(\nu_m + \epsilon) \rceil}{T\lambda}\right)^2\right) \\ &\leq \exp\left(-2 \cdot l' \cdot \left(\nu_m - \frac{T\lambda(\nu_m + \epsilon)}{T\lambda}\right)^2\right) \\ &\leq \exp(-2 \cdot (T\lambda - 2T\epsilon' - 1) \cdot \epsilon^2). \end{aligned}$$

As a result, $\mathbb{P}[\tau \leq l - T\epsilon'] \leq \sum_{m=1}^M \mathbb{P}\left[Z_m^{l'} \geq U_m\right] \leq M \exp(-2 \cdot (T\lambda - 2T\epsilon' - 1) \cdot \epsilon^2)$. $\square$

The second lemma, Lemma 6, provides an upper bound to the inventory optimization problem (12).

LEMMA 6. Define $R_{asst}^* = \max_{A \subseteq \mathcal{N}} \left\{ \sum_{j \in S} r_j \cdot \mathbb{P}(j \mid A) \right\}$ as the optimal objective value of the assortment optimization problem with margin $r_j = p_j - c_j$. Then for any inventory vector $\mathbf{I}$, its expected profit follows $\mathcal{P}(\mathbf{I}) \leq T\lambda R_{asst}^*$.

Proof: Let L be the number of customers arrived in the period $[0, T]$. The expected profit is given by

$$\mathcal{P}(\mathbf{I}) = \mathbb{E}\left[\sum_{\ell=1}^L r_{D^\ell} - \sum_{j \in [n]} c_j \cdot I_j\right] = \sum_{l=0}^{\infty} \mathbb{E}\left[\sum_{\ell=1}^L r_{D^\ell} - \sum_{j \in [n]} c_j \cdot I_j \mid L = l\right] \cdot \mathbb{P}[L = l].$$

In what follows, we will show that $\mathbb{E}\left[\sum_{\ell=1}^l r_{D^\ell} - \sum_{j \in [n]} c_j \cdot I_j\right] \leq l \cdot R_{asst}^*$ for any positive integer l . Let $Z_j^l = \sum_{\ell=1}^l \mathbb{I}[D^\ell = j]$ be the number of sales of product j up to customer l . Then

$$\begin{aligned} &\mathbb{E}\left[\left(\sum_{\ell=1}^l r_{D^\ell}\right) - \sum_{j \in [n]} c_j I_j\right] \\ &= \mathbb{E}\left[\sum_{\ell=1}^l \sum_{j \in [n]} (r_j - c_j) \cdot \mathbb{I}[D^\ell = j] + \sum_{\ell=1}^l \sum_{j \in [n]} c_j \cdot \mathbb{I}[D^\ell = j]\right] - \sum_{j \in [n]} c_j I_j \end{aligned}$$

$$\begin{aligned}
&= \sum_{\ell=1}^l \mathbb{E} \left[\sum_{j \in [n]} (r_j - c_j) \cdot \mathbb{E} \left[\mathbb{I}[D^\ell = j] \mid A^\ell \right] \right] + \mathbb{E} \left[\sum_{j \in [n]} c_j \cdot \sum_{\ell=1}^l \mathbb{I}[D^\ell = j] \right] - \sum_{j \in [n]} c_j I_j \\
&= \sum_{\ell=1}^l \mathbb{E} \left[\sum_{j \in [n]} (r_j - c_j) \cdot \mathbb{E} \left[\mathbb{I}[D^\ell = j] \mid A^\ell \right] \right] - \sum_{j \in [n]} \mathbb{E} \left[c_j \cdot (I_j - Z_j^l) \right].
\end{aligned}$$

Notice that $I_j \geq Z_j^l$ almost surely, as the sales of a product cannot be higher than its initial stock. Also, the term $\sum_{j \in [n]} (r_j - c_j) \cdot \mathbb{E}[\mathbb{I}[D^\ell = j] \mid A^\ell] = \sum_{j \in [n]} (r_j - c_j) \cdot \mathbb{P}(j \mid A^\ell)$ is the expected profit given assortment A^ℓ . Therefore, it is less or equal to R_{asst}^* according to Theorem (1). Therefore,

$$\mathbb{E} \left[\left(\sum_{\ell=1}^l r_{D^\ell} \right) - \sum_{j \in [n]} c_j I_j \right] \leq \sum_{\ell=1}^l R_{\text{asst}}^* = l \cdot R_{\text{asst}}^*$$

and the expected profit follows

$$\mathcal{P}(\mathbf{I}) \leq \sum_{\ell=0}^{\infty} l \cdot R_{\text{asst}}^* \cdot \mathbb{P}[L = l] = R_{\text{asst}}^* \cdot T\lambda,$$

where the last equality follows since L is a Poisson random variable with parameter $T\lambda$ and has expected value $T\lambda$. $\square$

Now, we are ready to prove Proposition 3.

Proof of Proposition 3: Define A^* as the optimal assortment defined in Lemma 6 and R_{asst}^* as its expected profit. Now we show that the inventory vector $\mathbf{I}$ defined in Proposition 3 is asymptotically optimal with rate $O\left(n\sqrt{\bar{L} \log(\bar{L})}\right)$. We separate the discussion into the following three parts: (a) bounding the expected revenue $\mathbb{E}\left[\sum_{\ell=1}^L r_{D^\ell}\right]$ from below; (b) bounding the cost $\sum_j c_j I_j$ from above; and (c) bounding the optimality gap.

(a) Bound the expected revenue. Recall that L is the number of arrived customers in time $[0, T]$.

The expected revenue follows as

$$\mathbb{E} \left[\sum_{\ell=1}^{\infty} r_{D^\ell} \cdot \mathbb{I}[t_\ell \leq T] \right] = \sum_{l=0}^{\infty} \mathbb{E} \left[\sum_{\ell=1}^L r_{D^\ell} \mid L = l \right] \cdot \mathbb{P}[L = l] \geq \sum_{l=\lceil T(\lambda - \epsilon_1) \rceil}^{\lfloor T(\lambda + \epsilon_1) \rfloor} \mathbb{E} \left[\sum_{\ell=1}^l r_{D^\ell} \right] \cdot \mathbb{P}[L = l],$$

where we will choose ϵ_1 carefully later. If there exists a lower bound $\tilde{R}_{\text{low}}$ that is independent of l and satisfy $\tilde{R}_{\text{low}} \leq \mathbb{E} \left[\sum_{\ell=1}^l r_{D^\ell} \right]$ for any positive integer $l \in (T(\lambda - \epsilon_1), T(\lambda + \epsilon_1))$, then

$$\begin{aligned}
\mathbb{E} \left[\sum_{\ell=1}^{\infty} r_{D^\ell} \cdot \mathbb{I}[t_\ell \leq T] \right] &\geq \sum_{l \in \mathbb{N}_+ : l \in (T(\lambda - \epsilon_1), T(\lambda + \epsilon_1))} \mathbb{E} \left[\sum_{\ell=1}^l r_{D^\ell} \right] \cdot \mathbb{P}[L = l] \\
&\geq \sum_{l \in \mathbb{N}_+ : l \in (T(\lambda - \epsilon_1), T(\lambda + \epsilon_1))} \tilde{R}_{\text{low}} \cdot \mathbb{P}[L = l] \\
&= \tilde{R}_{\text{low}} \cdot \sum_{l \in \mathbb{N}_+ : l \in (T(\lambda - \epsilon_1), T(\lambda + \epsilon_1))} \mathbb{P}[L = l]
\end{aligned}$$

$$\begin{aligned}
&\geq \tilde{R}_{\text{low}} \cdot \mathbb{P}\left[L \in (T(\lambda - \epsilon_1), T(\lambda + \epsilon_1))\right] \\
&= \tilde{R}_{\text{low}} \cdot \left(1 - \mathbb{P}\left[|L - T\lambda| \geq T\epsilon_1\right]\right) \\
&\geq \tilde{R}_{\text{low}} \cdot \left(1 - 2\exp\left(-\frac{(T\epsilon_1)^2}{2(T\lambda + T\epsilon_1)}\right)\right), \tag{22}
\end{aligned}$$

where in the last inequality, we use the concentration inequality for the Poisson random variable (Vershynin 2018). By choosing $\epsilon_1 = \lambda \cdot \sqrt{e \log(\bar{L})/\bar{L}}$, we have,

$$\exp\left(-\frac{(T\epsilon_1)^2}{2(T\lambda + T\epsilon_1)}\right) = \exp\left(-\frac{1}{2} \cdot \frac{e \cdot \log(\bar{L})}{1 + \sqrt{\frac{e \cdot \log(\bar{L})}{\bar{L}}}}\right) \leq \exp\left(-\frac{e \log(\bar{L})}{4}\right) \leq \exp\left(-\frac{\log(\bar{L})}{2}\right) = \frac{1}{\sqrt{\bar{L}}},$$

where the first inequality follows since $e \log x \leq x$ whenever $x \geq e$. Therefore, as long as we have the lower bound $\tilde{R}_{\text{low}}$, then the expected revenue follows as

$$\mathbb{E}\left[\sum_{\ell=1}^{\infty} r_{D^\ell} \cdot \mathbb{I}[t_\ell \leq T]\right] \geq \tilde{R}_{\text{low}} \cdot \left(1 - \frac{2}{\sqrt{\bar{L}}}\right).$$

Now we will obtain the lower bound $\tilde{R}_{\text{low}}$ for $\mathbb{E}\left[\sum_{\ell=1}^l r_{D^\ell}\right]$ for any positive integer $l \in (T(\lambda - \epsilon_1), T(\lambda + \epsilon_1))$. We define $Z_j^\ell = \sum_{i=1}^\ell \mathbb{I}[D^i = j]$ as the number of times that product $j \in A^*$ is chosen by the first ℓ customers. We further define a random variable

$$\tau = \inf\{\ell \mid \exists \text{ product } j \in A^* \text{ such that } Z_j^\ell = \lceil \bar{L}(\pi_j + \epsilon) \rceil\},$$

which is the first customer such that after she makes the decision, one of the products in the optimal assortment A^* is out of stock. More importantly, τ is a *stopping time*. Also, it only depends on customers' decisions and it is independent of customers' arrival times. Notice that for a fixed $l \in (T(\lambda - \epsilon_1), T(\lambda + \epsilon_1))$,

$$\mathbb{E}\left[\sum_{\ell=1}^l r_{D^\ell}\right] \geq \mathbb{E}\left[\sum_{\ell=1}^{\min\{\lfloor l - T\epsilon_1 \rfloor, \tau\}} r_{D^\ell}\right].$$

We will use a *Wald equation*-like argument to calculate the right-hand side. Notice that we cannot directly apply Wald's equation here, as $\{D^\ell\}_{\ell \in \mathbb{N}}$ is *not* a sequence of IID random variables. Indeed, as we discussed above, the set of available products A^ℓ that customer ℓ sees is not the same for all ℓ and thus the distribution of D^ℓ is not fixed.

Let $l' = \lfloor l - T\epsilon_1 \rfloor$. Define $\tilde{R}_{\text{asst}}^* := \sum_{j \in A^*} r_j \cdot \mathbb{P}(j \mid A^*)$ as the “revenue” part of the optimal assortment A^* (instead of profit, which doesn't have a tilde in the notation). For a given integer $l \in (T(\lambda - \epsilon_1), T(\lambda + \epsilon_1))$, we have

$$\mathbb{E}\left[\sum_{\ell=1}^{\min\{l', \tau\}} r_{D^\ell}\right] = \mathbb{E}\left[\sum_{\ell=1}^{\infty} r_{D^\ell} \cdot \mathbb{I}[\ell \leq l'] \cdot \mathbb{I}[\ell \leq \tau]\right]$$

$$\begin{aligned}
&= \sum_{\ell=1}^{\infty} \mathbb{E} \left[r_{D^\ell} \cdot \mathbb{I}[\ell \leq l'] \cdot \mathbb{I}[\ell \leq \tau] \right] \\
&= \sum_{\ell=1}^{\infty} \mathbb{E} \left[\mathbb{E} \left[\mathbb{I}[\ell \leq l'] \cdot \mathbb{I}[\ell \leq \tau] \cdot r_{D^\ell} \mid D^1, D^2, \dots, D^{\ell-1} \right] \right] \\
&= \sum_{\ell=1}^{\infty} \mathbb{E} \left[\mathbb{I}[\ell \leq l'] \cdot \mathbb{I}[\ell \leq \tau] \cdot \mathbb{E} \left[r_{D^\ell} \mid D^1, D^2, \dots, D^{\ell-1} \right] \right] \\
&= \sum_{\ell=1}^{\infty} \mathbb{E} \left[\mathbb{I}[\ell \leq l'] \cdot \mathbb{I}[\ell \leq \tau] \cdot \tilde{R}_{\text{asst}}^* \right] \\
&= \tilde{R}_{\text{asst}}^* \cdot \mathbb{E} \left[\min\{l', \tau\} \right].
\end{aligned}$$

Here the first equality follows the definition. The second equality follows Fubini's theorem as the random variables are nonnegative. The third equality follows the tower property of expectation. The fourth equality follows the fact that τ is a stopping time. In particular, event $\{\ell \leq \tau\} = \{\ell - 1 \geq \tau\}^c$ is in the filtration defined by $D^1, D^2, \dots, D^{\ell-1}$, as we can tell whether a product will out of stock at any time after the ℓ th customer makes her decision by checking the decisions of the first $\ell - 1$ customers. The fifth equality follows an observation: given that no products are out of stock after first $\ell - 1$ customers' visits, the set A^ℓ of available products that ℓ th customer will see remains the same as A^1 , which is the optimal assortment A^* , according to the construction of the inventory decision **I**. Therefore, if $\ell \leq \tau$, then $\mathbb{E}[r_{D^\ell} \mid D^1, \dots, D^{\ell-1}] = \mathbb{E}[r_{D^\ell} \mid A^\ell] = \mathbb{E}[r_{D^\ell} \mid A^*] = \tilde{R}_{\text{asst}}^*$.

Now we will further lower bound $\mathbb{E}[\min\{l', \tau\}]$ for any fixed $l' = \lfloor l - T\epsilon_1 \rfloor$ for $l \in (T(\lambda - \epsilon_1), T(\lambda + \epsilon_1))$ by Lemma 5. Notice that

$$\begin{aligned}
\mathbb{E}[\min\{l', \tau\}] &= \mathbb{E}[\min\{l', \tau\} \mid l' \geq \tau] \cdot \mathbb{P}[l' \geq \tau] + \mathbb{E}[\min\{l', \tau\} \mid l' < \tau] \cdot \mathbb{P}[l' < \tau] \\
&= \mathbb{E}[\tau \mid l' \geq \tau] \cdot \mathbb{P}[l' \geq \tau] + \mathbb{E}[l' \mid l' < \tau] \cdot \mathbb{P}[l' < \tau] \\
&\geq l' \cdot \mathbb{P}[l' < \tau] \\
&\geq l' \cdot (1 - |A^*| \cdot \exp(-2(T\lambda - 2T\epsilon_1 - 1) \cdot \epsilon^2)) \\
&\geq l' \cdot \left(1 - |A^*| \cdot \sqrt{\frac{1}{T\lambda}}\right) \\
&\geq (T\lambda - 2T\epsilon_1 - 1) \cdot \left(1 - |A^*| \cdot \sqrt{\frac{1}{T\lambda}}\right) \\
&= \bar{L} \cdot \left(1 - 2\sqrt{\frac{e \log(\bar{L})}{\bar{L}}} - \frac{1}{\bar{L}}\right) \cdot \left(1 - |A^*| \cdot \sqrt{\frac{1}{\bar{L}}}\right)
\end{aligned}$$

where the first and second equalities follow the definition, the first inequality follows as l' is a constant, and the second inequality follows by Lemma 5 and the construction of inventory vector $\mathbb{I}$ and τ . In particular, before the hitting time happens, the consumer decision D^ℓ follows $D^\ell = j$

with probability $\mathbb{P}(j \mid A^*) = \pi_j$. The last two inequalities follow as $l > T(\lambda - \epsilon_1)$, $l' \geq T\lambda - 2T\epsilon_1 - 1$, and $\epsilon = 0.5 \cdot \sqrt{\log(T\lambda)/(T\lambda - 2T\epsilon_1 - 1)}$. Combining all elements, we have

$$\begin{aligned} \mathbb{E} \left[\sum_{\ell=1}^l r_{D^\ell} \right] &\geq \mathbb{E} \left[\sum_{\ell=1}^{\min\{l', \tau\}} r_{D^\ell} \right] = \tilde{R}_{\text{asst}}^* \cdot \mathbb{E} \left[\min\{l', \tau\} \right] \\ &\geq \tilde{R}_{\text{asst}}^* \cdot T\lambda \left(1 - 2\sqrt{\frac{e \log(\bar{L})}{\bar{L}}} - \frac{1}{\bar{L}} \right) \cdot \left(1 - n\sqrt{\frac{1}{\bar{L}}} \right) := \tilde{R}_{\text{low}} \end{aligned}$$

whenever $l \in (T(\lambda - \epsilon_1), T(\lambda + \epsilon_1))$. Therefore, going back to Equation (22) and plugging in the defined $\tilde{R}_{\text{low}}$, we have the expected revenue bounded below as

$$\mathbb{E} \left[\sum_{\ell=1}^{\infty} r_{D^\ell} \cdot \mathbb{I}[t_\ell \leq T] \right] \geq \tilde{R}_{\text{low}} \cdot \left(1 - \frac{2}{\sqrt{\bar{L}}} \right) \geq \tilde{R}_{\text{asst}}^* \cdot \bar{L} \cdot \left(1 - 2\sqrt{\frac{e \log(\bar{L})}{\bar{L}}} - \frac{1}{\bar{L}} - n \cdot \sqrt{\frac{1}{\bar{L}}} - \frac{2}{\sqrt{\bar{L}}} \right)$$

(b) Bound the cost. We have $\sum_{j \in [n]} c_j \cdot I_j \leq \sum_{j \in A^*} c_j \cdot (1 + \bar{L}(\pi_j + \epsilon))$. Notice that whenever $\bar{L} \geq e^4$, $\epsilon_1 = \lambda \cdot \sqrt{e \log(\bar{L})/\bar{L}} \leq \lambda \sqrt{e \cdot 4/e^4} \leq 0.45 \cdot \lambda$, which results in

$$\epsilon = 0.5 \cdot \sqrt{\frac{\log(\bar{L})}{\bar{L} \left[1 - 2\sqrt{e \log(\bar{L})/\bar{L}} - 1/\bar{L} \right]}} \leq \frac{0.5}{\sqrt{1 - 2 \times 0.45 - \exp(-4)}} \cdot \sqrt{\frac{\log(\bar{L})}{\bar{L}}} \leq 1.8 \cdot \sqrt{\frac{\log(\bar{L})}{\bar{L}}}.$$

Plugging the upper bound for ϵ , we have the upper bound for the total procurement cost whenever $T\lambda \geq e^4$ as $\sum_{j \in [n]} c_j \cdot I_j \leq nc_{\max} + \bar{L} \sum_{j \in A^*} c_j \pi_j + 1.8nc_{\max} \sqrt{\bar{L} \cdot \log(\bar{L})}$.

(c) Approximate. Following the discussion in (a) and (b), the expected profit $\Pi(\mathbf{I})$, which is the expected revenue subtract the cost, has a lower bound

$$\begin{aligned} \Pi(\mathbf{I}) &\geq \tilde{R}_{\text{asst}}^* \cdot \bar{L} \cdot \left(1 - 2\sqrt{\frac{e \log(\bar{L})}{\bar{L}}} - \frac{1}{\bar{L}} - n \cdot \sqrt{\frac{1}{\bar{L}}} - \frac{2}{\sqrt{\bar{L}}} \right) - \left(nc_{\max} + \bar{L} \sum_{j \in A^*} c_j \pi_j + 1.8n \cdot c_{\max} \cdot \sqrt{\bar{L} \cdot \log(\bar{L})} \right) \\ &= R_{\text{asst}}^* \cdot \bar{L} - O \left((p_{\max} + c_{\max}) \cdot n \cdot \sqrt{\bar{L} \log \bar{L}} \right), \end{aligned}$$

where we use the fact $\tilde{R}_{\text{asst}}^* \leq p_{\max}$ and $\tilde{R}_{\text{asst}}^* - \sum_{j \in A^*} c_j \pi_j = \sum_{j \in A^*} (r_j - c_j) \pi_j = R_{\text{asst}}^*$. Therefore, the inventory vector $\mathbf{I}$ has an optimality gap

$$\Pi^* - \Pi(\mathbf{I}) \leq \bar{L} R_{\text{asst}}^* - \Pi(\mathbf{I}) \leq O \left(n \cdot \sqrt{\bar{L} \log \bar{L}} \right),$$

where the first inequality follows Lemma 6 and the second inequality follows that both $p_{\max}$ and $c_{\max}$ are independent of $\bar{L}$. Finally, we argue that $\mathbf{I}$ is asymptotically optimal as follows:

$$\frac{\Pi(\mathbf{I})}{\Pi^*} \geq \frac{\Pi(\mathbf{I})}{\bar{L} R_{\text{asst}}^*} \geq 1 - \frac{n}{R_{\text{asst}}^*} \cdot O \left(\sqrt{\frac{\log(\bar{L})}{\bar{L}}} \right) \rightarrow 1 \quad \text{as} \quad \bar{L} \rightarrow \infty.$$

□