Enhancing Customer Baskets through Assortment Optimization

Qiuyu Tian[1,*], Hongwei Tang[1,2], Kun Wang[3]
[1]*Institute of Information Superbahn, Nanjing, Jiangsu, China*
[2]*University of Chinese Academy of Sciences, Nanjing, Jiangsu, China*
[3]*Institute of Computing Technology Chinese Academy of Sciences, Beijing, China*
**Corresponding Author.*

Abstract: This paper introduces models that support a novel system designed to optimize store selection when customers shop for baskets of items (e.g., bread, milk, snacks). Such behavior is typical in traditional grocery shopping and applicable to various e-commerce platforms. These platforms are often limited in the assortment they can provide, constrained to a fraction of the potential millions of products available (e.g., constrained to tens of thousands of products), making the problem of selection optimization both pertinent and computationally complex. To address this challenge, methods are proposed that leverage customer behavior data to identify groups of interchangeable items, enabling the formulation of customer choice models that reflect category complementarity and product substitutability. The model was initially implemented for selection optimization at one of Amazon's fulfillment centers in early 2020. Following its success, plans were made to expand its application to multiple additional sites by the fourth quarter of the same year. Retrospective analysis shows that compared to Amazon's existing selection strategy, sales have significantly increased, with sales in specific service areas growing by 4.8%, which translates to a significant increase in annual revenue. Additionally, the deployment of the model demonstrated a 13.7% reduction in basket abandonment rates and a 13.8% increase in units per order (UPO). Efforts are ongoing to extend the model to a wider selection process, with experimental implementation within Amazon anticipated by the end of 2020.

Keywords: Basket Optimization; Product Assortment; Item Substitutability; Category Complementarity

1. Introduction

This paper introduces models that underpin a novel system for optimizing store assortments when customers engage in basket shopping for various necessities (e.g., bread, milk, snacks). This behavior is prevalent not only in conventional grocery shopping but also across various e-commerce platforms, including those similar to Amazon's Fresh and same-day delivery services. These platforms face inherent constraints in the product selection they can offer, typically encompassing only a fraction of the potential product universe, which may extend to hundreds of thousands or even millions of items. Consequently, the challenge of optimizing product selection becomes highly relevant and intrinsically complex [1].

To address this issue, methodologies were devised that leverage customer behavior data to identify clusters of substitutable products. This foundation enabled the development of customer choice models that effectively represent category complementarity (e.g., how a limited assortment of bread affects the sales of milk) and product substitutability (e.g., the presence of numerous varieties of organic milk reducing each other's demand). An optimization problem was then formulated that incorporates these elements alongside basket abandonment—a scenario where customers may abandon their entire purchase if they cannot find all desired items [2]. Efficient algorithms were constructed to solve this problem.

Historically, optimizing selection in the presence of complementarity and substitutability has posed significant challenges, leading to the adoption of models that primarily rank products based on attractiveness or utilize discrete optimization without fully integrating the nuances of customer behavior. Existing solutions in e-commerce have often relied on econometric models or machine learning approaches focused on product feature ranking, overlooking the complexities of category interactions and the risk of basket abandonment. The academic literature on this topic is sparse, particularly in providing a unified framework that integrates these multifaceted aspects.

This approach departs from traditional models that impose strong behavioral assumptions, such as requiring customers to compute utilities for all items and make sequential choices [3]. Instead, a behavioral model is proposed that is both intuitively straightforward and plausible, offering a more realistic depiction of consumer decision-making processes [4]. This model has been successfully prototyped and is in the process of being expanded to other domains, showcasing its potential applicability across various contexts.

Research on assortment or selection optimization has primarily concentrated on customer choice models like multinomial logit (MNL), mixed logit, and Markov models, which largely address substitutability [5]. These problems are typically framed as binary maximization issues with quasi-concave objective functions under linear constraints. While such models have demonstrated efficacy in handling product substitutability, they encounter significant challenges when extended to incorporate additional customer behaviors, such as complementarity and the risk of basket abandonment.

In response to these challenges, the objective function was formulated from fundamental principles, deriving expressions for expected profit under the proposed behavioral assumptions. This formulation resulted in a stochastic nonlinear binary optimization problem, which was approached using a first-order optimizer such as ADAM within the PyTorch framework [6]. This method continues the recent trend of employing deep neural networks for solving complex mixed 0-1 linear and nonlinear programs [7], marking a significant advancement in assortment optimization research.

2. Problem Definition

A typical grocery shopping experience can be conceptualized as a sequence of decisions made by the customer [8]. When entering the store, the customer typically has a predefined intention to purchase a variety of items, such as groceries, cleaning supplies, meal preparation ingredients, body care products, or medication. During the shopping process, the customer will focus on specific categories of products that align with their needs for that particular trip. For instance, if the shopping goal includes body care items, the customer might consider products such as soap, toothbrushes, toothpaste, floss, and mouthwash. Within each product category, the customer is likely to browse through a subset of available products. This browsing stage involves evaluating different options before deciding whether to add a particular product to their basket. It is also possible that the customer might not find a satisfactory option within a category, in which case they may decide not to select any product from that category.

As the customer progresses through each category, they continuously evaluate whether to continue their shopping journey or to terminate it prematurely. The decision to abandon the basket might be influenced by various factors, including dissatisfaction with the available product selection or the perceived quality and price of items. This abandonment signifies the choice to fulfill their needs through an alternative store or channel. In the context of e-commerce, such as on platforms like Amazon, this behavior can be observed when customers navigate through product listings and make decisions based on the assortment available to them [9].

This abstracted model of shopping behavior forms the foundation for developing optimization strategies for product assortment, considering factors such as customer satisfaction, category completeness, and the risk of basket abandonment.

3. Methodology

This section describes the methodology employed to optimize product assortment, encapsulated within a framework that is referred to as the Discrete Choice Model (DCM) [10]. This model comprises four essential components: Substitution Group Learning, Within Group Demand Model, Cross Group Complementarity Model, and Constrained Assortment Optimization.

Together, these modules form an integrated approach to understanding customer behavior and making optimal product selection decisions. Each module is introduced below, with a more detailed explanation provided in the appendices.

3.1 Substitution Group Learning

Substitution Group Learning is the process of classifying products into distinct categories known as substitution groups [11]. This classification relies on analyzing customer search queries and purchase data to identify items that can be considered substitutes. The assumption is that products meeting similar customer needs can be categorized together, which can be inferred through the analysis of search keywords used by customers when browsing and purchasing products.

For instance, in a large-scale e-commerce setting, the platform might categorize millions of products into thousands of substitution groups. This classification is based on patterns in customer behavior, which indicate that certain products can fulfill similar needs. Formally, let A represent the set of all candidate products (ASINs), and $A_g \subseteq A$ be the set of candidate products in a specific substitution group g. The model identifies such groups and assigns products to them, ensuring that products serving similar functions are classified under the same category.

3.2 Within Group Demand Model

The Within Group Demand Model focuses on understanding customer preferences within each substitution group and predicting the probability of product-level purchases [12]. This model employs a bias-corrected Multinomial Logit (MNL) framework to estimate the likelihood that a customer will purchase a particular product j given that they are considering a specific substitution group S_g. Let $X_j \in \{0, 1\}$ denote whether the product j is purchased by the customer. The model uses the following equation to determine the purchase probability:

$$P(X_j = 1 \,|S_g) \approx \frac{w_j}{1+\sum_{j' \in S_g} w_{j'}} \gamma_j(S_g) \tag{1}$$

In this equation, w_j represents the attractiveness of the product j which is an encoded measure of its appeal to the customer based on various factors such as price, quality, and brand. The term $\gamma_j(S_g)$ is a bias correction factor that adjusts for the fact that customers only browse a subset of products within a substitution group before making a purchase decision. This correction is crucial for accurately capturing the decision-making process, as it accounts for the limited scope of customer consideration within the group.

3.3 Cross Group Complementary Model

The Cross Group Complementarity Model extends the analysis to consider customer purchasing behavior across multiple substitution groups [13]. It accounts for the likelihood that customers might abandon their basket if they cannot find products they need across different groups. This model, therefore, captures both intra-group substitutability and inter-group complementarity.

Let $C = (C_1, C_2, \ldots, C_G)$ represent the customer's consideration set, where $C_g \in \{0, 1\}$ indicates whether the customer is considering substitution group g for a purchase. Similarly, $B_g \in \{0, 1\}$ denotes whether the customer buys at least one item from group g. If B_g is solely dependent on C_g, the random reward $R_g(S_g)$ for a purchase from the group is expressed as:

$$R(S_g \,|C_g) = C_g \sum_{j \in S_g} r_j P(X_j = 1|S_g) = C_g \frac{\sum_{j \in S_g} r_j \; w_j \, \gamma_j}{1 + \sum_{j' \in S_g} w_{j'}} \tag{2}$$

where r_j represents the per-unit benefit of product j, which could be defined in terms of maximized demand or profit. The total reward for the selection S, given $C$and B, is then:

$$\pi(S|C, B) = \left(\prod_{g=1}^{G} (1 - p_b)^{C_g(1-B_g)} \right) \sum_{g=1}^{G} R(S_g) \,\# \tag{3}$$

Here, p_b is the basket abandonment probability, reflecting the likelihood that a customer might abandon their shopping basket if they do not find satisfactory items in their desired categories. When $C_g = 1$ and $B_g = 0$, the expected total reward becomes:

$$\pi(S) \approx \mathbb{E}_C \left[\sum_g \left(\frac{C_g \sum_{j \in S_g} r_j w_j \gamma_j}{1 + \sum_{j' \in S_g} w_{j'}} \prod_{h \neq g, B_h = 1}^{G} \left(1 - \frac{p_b}{1 + \sum_{j \in S_h} w_j}\right) \right) \right] \tag{4}$$

which leads to a sample-based approximation.

3.4 Constrained Assortment Optimization

The final component, Constrained Assortment Optimization, formulates the overall problem as a stochastic binary optimization task. The goal is to maximize the expected reward, which can be defined in terms of revenue or profit, under various business constraints such as capacity limits and floor or ceiling bounds on the selection size.

Let A_g represent the set of all candidate products in substitution group g. For each product j within this group, $x_{jg} \in \{0,1\}$ is a binary indicator denoting whether product j is included in the assortment S_g. The revenue associated with the decision vector x can be expressed as:

$$\pi(x) = \mathbb{E}_C\left[\sum_g \left(C_g \frac{\sum_{j\in A_g} x_{jg} r_j w_j \gamma_j\left(S_g\right)}{1+\sum_{j'\in A_g} x_{j'g} w_{j'}} \prod_{h\neq g, C_h=1} \left(1 - \frac{p_b}{1+\sum_{j'\in A_g} x_{j'g} w_{j'}}\right)\right)\right] \quad (5)$$

The optimization problem is framed as:

$$\arg\max_x \pi(x) \quad (6)$$

subject to

$$\sum_{g=1}^{G}\sum_{j\in A_g} x_{jg} \leq N, \quad (7)$$

and the following constraints:

$$l_k(\mathrm{N}) \leq \sum_{\mathrm{j}=1}^{\mathrm{g}}\sum_{j\in A_g} x_{jg} L_{jk} \leq u_k(N)\ , \text{ for each } G, L, k, x_{jg} \in \{0,1\},\ \forall j \in A_g, g = 1, \ldots, G. \quad (8)$$

Here, $x_{jg} \in \{0,1\}$ is a binary variable indicating whether a product j in group g is included in the assortment, l_k is the lower (floor) bound on the fraction of group size k, and u_k is upper (ceiling) bound on the fraction of group size k. The constraints include a total capacity limit N across all selected products and mixture constraints that regulate the proportion of products selected within each group. The problem is inherently nonlinear and binary, requiring sophisticated optimization techniques such as Pytorch's constrained optimization library to solve [14].

4. Experiment

This subsection focuses on the collection and preparation of data necessary for the experiment. A comprehensive dataset was gathered, consisting of six months of customer search query data from an e-commerce platform. This data was utilized to identify common ASIN-keyword combinations, which were subsequently processed by the Substitution Group Module to classify 4.7 million candidate ASINs into 15,008 substitution groups. Additionally, session-level search data was collected over a one-month period (from December 27, 2019, to January 25, 2020) to analyze customer browsing and purchasing behavior within these substitution groups. This dataset was restricted to a specific regional area to ensure that the resulting ASIN propensity and attractiveness reflected regional preferences, encompassing aspects of substitutability and complementarity.

4.1 Data Collection and Preparation

This subsection focuses on the collection and preparation of data necessary for the experiment. A comprehensive dataset was gathered, including six months of customer search query data from an e-commerce platform. This data was used to identify common ASIN-keyword combinations, which were then processed by the Substitution Group Module to classify 4.7 million candidate ASINs into 15,008 substitution groups. Additionally, session-level search data was collected over a one-month period (from December 27, 2019, to January 25, 2020) to analyze customer browsing and purchasing behavior within these substitution groups. This dataset was restricted to a specific regional area to ensure that the resulting ASIN propensity and attractiveness reflected regional preferences, including aspects of substitutability and complementarity.

4.2 Model Implementation and Training

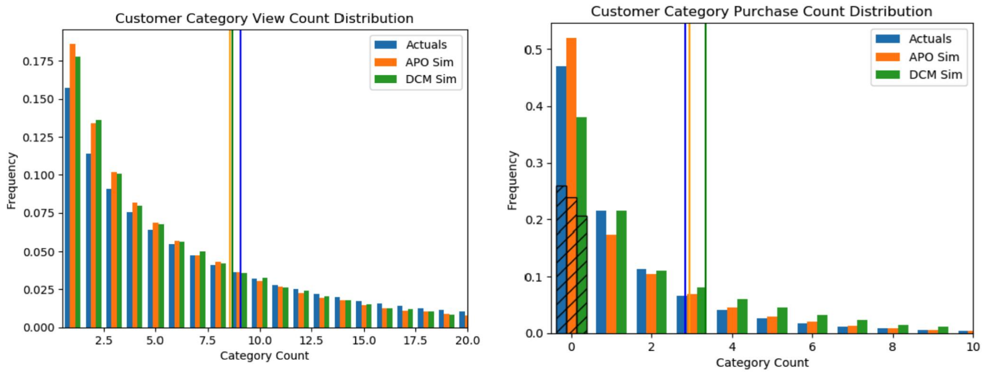
This section describes the implementation and training of the Discrete Choice Model (DCM) using the collected data. The Within Group Module of DCM was employed to estimate the attractiveness of each ASIN within a substitution group. Two primary components of attractiveness were computed: the search-based attractiveness w_j^S, derived from the view propensity and conditional purchase probability, and the forecast-based attractiveness w_j^F, calculated by normalizing the ASIN's forecasted demand with the number of customers who purchased from the same group. The final utility w_j was obtained by combining these two components using a weighted sum $w_j = \theta\, w_j^S + \left(1 - \theta\right) w_j^F$ with $\theta = 0.1$. A variational inference approach implemented in PyTorch Probability [4] to train the Multinomial Logit (MNL) model, applying weight regularization through prior distributions to impose a structure on the model parameters.

4.3 Policy Generation and Optimization

This section details the generation of weekly selection policies using both the DCM and the existing production model (APO). The production model addresses a constrained integer programming problem to optimize forecasted demand. In contrast, the DCM leverages historical search data and forecast inputs to generate the selection policy. The Constrained Optimization library in PyTorch was employed to optimize the DCM's selection policy stochastically, utilizing a basket of completed orders to account for complementarity across substitution groups. Shipping order data was mapped to the substitution groups to form 1,931,675 completed baskets, which were used to construct the optimization objective as outlined in Equation (5). During the optimization process, a tunable parameter representing the basket abandonment rate ρ_b was considered, with optimization conducted using a stochastic approach over 30 epochs, a batch size of 26,000, and a learning rate of 0.1 with a decay factor of 0.9 per epoch.

4.4 Evaluation and Simulation

This subsection focuses on the evaluation metrics and simulation techniques used to assess the performance of the Discrete Choice Model (DCM) relative to the Alternative Product Offering (APO). The selection outputs were benchmarked using key metrics, including regional glance views (GV), regional unit sales, the number of ASINs with regional GV (AWAGV), and the number of ASINs with regional sales (AWAS) within the fulfillment center's service area. Since observed sales and GV metrics might not fully capture the impact of selection differences on customer behavior—particularly regarding lost sales due to basket abandonment—a simulation approach was employed. This simulation utilized the DCM framework to estimate key behavioral metrics, such as no-purchase rate, effective basket abandonment rate, average basket size, and the average number of substitution groups viewed per trip. By setting the basket abandonment rate $\rho_b = 4\%$, the simulation results were aligned with historical search data, thereby approximating the impact of selection differences on customer behavior in a manner similar to a randomized experiment.



(a) View distribution (b) Purchase distribution

Figure 1. Distribution of Categories Viewed (left) and Purchased (right) by Customers during the Week of Feb 23 under Observed APO, Simulated APO, and Simulated DCM Selections

5. Results

The simulation conducted under the APO selection successfully replicates key characteristics observed in the actual search and order data (Table 1), where the APO policy was in place. Notably, 47% of customer sessions ended without a purchase, which is closely matched by the 52% seen in the simulation. Additionally, 26% of customers added items to their cart but did not complete the purchase, while the simulation shows a similar abandonment rate of 24%. The distributions for basket size in terms of views (Figure 1(a)) and purchases (Figure 1(b)) exhibit a long-tail pattern, indicating that customers explore a wide range of categories but purchase from a smaller subset. Specifically, customers viewed an average of 9.07 categories but only completed orders in 2.85 categories on average (Table 1). The APO simulation mirrors this behavior with an average of 8.55 categories viewed and 2.95 units per order (UPO).

When applying the DCM selection in the simulation, improvements were observed across several crucial metrics related to store completeness and basket building.

Table 1. Simulation Results for Customer Behavior under APO and DCM Selection Offerings

Model	No Purchase Rate	Effective Basket Abandon Rate	Avg Basket Size (UPO)	Avg Number Categories Viewed
Actuals	47.0%	25.9%	2.8	9.1
Sim APO	52.0%	23.9%	2.9	8.5
Sim DCM	38.1%	20.6%	3.4	8.7
% Change	-26.8%	-13.7%	13.8%	1.5%

The percentage of sessions without purchases dropped significantly to 38%, a 26.7% reduction compared to APO. Furthermore, the rate of abandoned carts decreased by 13.7%, resulting in a rate of 20.6% (Table 1). The mean UPO experienced an increase of 13.8%, reaching 3.35, while the average number of categories viewed per customer remained comparable between the two models, with DCM showing an average of 8.68 categories. Overall, the simulation suggests that implementing DCM could lead to a 4.31% lift in sales, primarily due to the reduction in basket abandonment from 23.9% to 20.6%.

Further analysis adjusting for shipping promises and differences in basket abandonment revealed that DCM increased regional glance views (GV) by 0.8% and sales by 4.8% in comparison to the current production policy generated by APO (Table 2). Additionally, DCM's selection included a higher number of ASINs with non-zero regional GV (+0.4%, or 509 ASINs) and sales (+0.2%, or 250 ASINs) than APO. The selection generated by DCM differed from APO's by 13,498 ASINs, representing 10.6% of the total selection. It's worth noting that with a more advanced view propensity estimation, the DCM model could further enhance performance, potentially leading to an additional increase of 2.3% in GV and 2.2% in sales, as seen in the DCM* scenario in Table 2.

Table 2. Benchmarks on SAZ1 Policy of Week Deb 23 for APO and DCM

Model	Regional GV	Regional Sales	% AWAGV	% AWAS
APO	3,908,818	614,112	93.20%	68.00%
DCM	3,939,528	643,923	93.60%	68.20%
DCM*	4,032,779	657,932	93.90%	69.20%

Table 3. Selection Difference between APO and DCM for Substitution Group: "Carpet Cleaner — Carpet Cleaner Solution — Carpet Shampoo"

ASIN	Item Name	View Propensity	Attractive-ness *w*	Forecast Demand	GV	Sales	APO Selected	DCM Selected
B00XPSS33A	arm & hammer pet fresh carpet odor eliminator plus oxi clean dirt fighters (pack of 3), 48.9 ounce	4.1E-02	2.3E-03	3.1	135	33	N	Y
B00NB8N7OO	rug doctor 04127 portable machine and upholstery cleaner, 2-pack	6.8E-04	2.1E-05	4.4	2	1	Y	N
B01DL1ZKFE	hoover expert pet 64 ounce carpet washer liquid detergent, ah15072, 64 oz, red	1.5E-03	7.4E-05	3.9	0	0	Y	N
B07RCNGVP2	hoover carpet paws & claws premixed spot machine cleaning shampoo, pet stain solution and odor remover, 32oz cleaner formula, ah30940, white	1.1E-03	6E-05	4.1	1	0	Y	N
B00GJABIRY	hoover pro plus 2x carpet washer and upholstery detergent solution, 120 oz, ah30051nf, red	2.3E-03	5.7E-05	4.0	13	0	Y	N

An illustrative example can be found in the substitution group "carpet cleaner — carpet cleaner solution — carpet shampoo" (Table 3). Here, it can be observed that the demand forecast significantly underestimated the sales potential of ASIN B00XPSS33A, resulting in APO excluding it from the selection. However, DCM identified this item as popular based on its high view propensity during training, which contributed to a high enough attractiveness score for the optimization process to select it. This choice was validated by the higher regional GV and sales figures during the target week, surpassing other ASINs chosen by APO but not by DCM.

6. Conclusion

This study has developed innovative models for optimizing store selection, tailored for e-commerce programs that serve customers shopping for a variety of needs. These programs include options similar to Amazon's Sub-Same Day (SSD), Fresh, and Prime Now. The system was first deployed for an SSD site in March 2020 and is designed to account for key customer behaviors such as category complementarity, product substitutability, and basket abandonment. These factors have traditionally been considered separately in both customer behavior modeling and assortment optimization literature, making this integrated approach unique. A complex non-linear, non-convex constrained optimization problem was effectively addressed using PyTorch.

Simulation results indicate that the Discrete Choice Model (DCM) enhances key metrics related to store completeness and basket building. The model reduces the rate of abandoned customer trips by 13.7% and increases units per order by 13.8% compared to the current production policy. Additionally, backtests demonstrate a 4.8% increase in regional sales when employing DCM.

Looking ahead, it is anticipated that DCM will evolve into a unified system for selection optimization across multiple programs, potentially incorporating aspects of Prime Now and F3 (Fresh, Food, Fast) selections. Future work will involve extending the model to accommodate heterogeneous customer segments. Moreover, the model aims to enhance search and discovery processes by highlighting items that are most appealing to users. This approach could also address the cold-start problem by using estimated utilities to better position new products, thereby accelerating their sales growth.

Acknowledgements

This research was supported by The Research of Key Technologies and Industrialization for Intelligent Computing in Large-Scale Video Scenarios under Digital Social Governance of Henan, China (grant number 241100210100).

References

[1] Kok, A.G., Fisher, M.L., Vaidyanathan, R.: Assortment planning: Review of literature and industry practice. In: Retail Supply Chain Management, 99–153. Springer (2008).

[2] Li, G., Wu, J., Humair, S., Razoumov, L.: Deciding what jeans to sell – Demand forecasting and optimization problems in softline assortment planning. Tech. rep., Amazon (2018).

[3] Kingma, D.P., Ba, L.J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (2015).

[4] Cotter, A., Jiang, H., Sridharan, K.: Two-player games for efficient non-convex constrained optimization. Proceedings of the 30th International Conference on Algorithmic Learning Theory, 300–332 (2019).

[5] Davis, J.M., Gallego, G., Topaloglu, H.: Assortment optimization under variants of the nested logit model. Operations Research, 250–273 (2014).

[6] Paszke, A., Gross, S., Massa, F., et al.: PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Advances in Neural Information Processing Systems (NeurIPS) (2019).

[7] Fischetti, M., Jo, J.: Deep neural networks as 0-1 mixed integer linear programs: A feasibility study. arXiv preprint arXiv:1712.06174 (2017).

[8] Ruiz, F.J.R., Athey, S., Blei, D.M.: SHOPPER: A probabilistic model of consumer choice with substitutes and complements. Annals of Applied Statistics, 1–27 (2020).

[9] Kok, A.G., Fisher, M.L., Vaidyanathan, R.: Assortment planning: Review of literature and industry practice. In: Retail Supply Chain Management, 99–153. Springer (2008).

[10] Liu, Z., Li, G., Lu, Y., et al.: Basket optimization in pantry assortment planning using discrete choice model. Tech. rep., Amazon (2019).

[11] Li, G.: Learning substitutable product group from customer search behavior and purchase intent. Tech. rep., Amazon (2019).

[12] Li, G., Wu, J., Humair, S., et al.: Deciding what jeans to sell – Demand forecasting and optimization problems in softline assortment planning. Tech. rep., Amazon (2018).

[13] Keane, M., Demirel, S., Ma, X., et al.: Assortment planning with complementary products. Tech. rep., Amazon (2018).

[14] Cooper: A toolkit for Lagrangian-based constrained optimization in PyTorch. GitHub repository. Available at: https://github.com/cooper-org/cooper.