Application of Nested Logit Model on Recommendation System

Wei-Yu Fan*, Yu-Chan Chen

May 2024

Abstract

We investigated the application of the nested logit (NL) model in recommendation systems. By employing random utility models and the nested logit distribution, we conceptualize consumer interactions with products as stochastic events, modulated by individual preferences across product categories. Subsequently, within the constraints of limited advertising slots, we optimize product assortments to maximize the likelihood of consumer engagement. Moreover, we derive a closed-form solution for this model and furnish accompanying algorithms.

Keywords: Random Utility Model, Nested Logit Model, Recommendation System, Assortment Optimization.

*Graduate student in Department of Economics, National Taiwan University. Email address: entrencemania@gmail.com

1 Introduction

The random utility model has long been a cornerstone in recommendation systems, enabling the modeling of the relationship between consumers and products. Through this framework, we derive consumer preferences towards products, with the logit model being the most widely recognized, further extended to handle choices involving three or more alternatives, known as the multinomial logit model (MNL). The advantage of the logit family lies in its computationally tractable probabilities and compatibility with regression models. McFadden (1977) introduced the generalized extreme value (GEV) distribution, a generalization of the logit model, with the NL model being a specific instance of GEV distribution. Notably, the NL model addresses dependencies among products, a crucial consideration in recommendation systems. For instance, in a scenario where advertisements for screens, graphics cards, keyboards, and headphones are placed in four separate slots, such a placement strategy may boost click-through rates for consumers interested in purchasing tech products. However, consumers with no immediate intention to purchase tech items may lack sufficient incentive to engage, highlighting the oversight of not considering the high correlation among these four products.

The NL model is specifically designed to address scenarios where correlations among products are prevalent. Its distinguishing feature lies in its capability to handle dependencies among products by clustering them,

thereby making the correlations more apparent and consequently enhancing the effectiveness of recommendation systems.

The other benefit of NL model is about fairness and diversity. In Zhao et al. (2024), the conventional recommendation system tends to over-recommend popular items, while unpopular items are under-represented. This is called popularity bias (Abdollahpouri et al. (2019)) and it may lead to several problems. For example, small buisness is harder to grow even if it has similar quality as the popular and bigger one. Moreover, these small buisness may drop out from the market, imply the monopolization of the market.

2 Related Literature

From conventional choice and behavior models, McFadden (1972) proposed the MNL model which depicits the consumers's behavior of choosing among multiple products. An extension of the MNL model is the NL model proposed by Williams (1977). The NL model is designed to handle the hierachical structure of the choice set, which is common in the real world. The NL model has been widely applied in various fields, such as transportation, marketing, and economics. For example, Hensher and Greene (2002) studies the estimation of NL model using data of travellers choosing different transportation modes. Another usecase is the study of assortments optimization probelm in pricing and expected revenue maximization. Rusmevichientong,

Shen, and Shmoys (2009) study that given the price of each product, how to choose the optimal assortment to maximize the expected revenue with the space constraint. The space constarin means that the number of products that can be displayed is limited and each product has a different size. Though this problem is NP-complete, they provide a polynomial time approximation acheme to solve this problem. Following Rusmevichientong, Shen, and Shmoys (2009), Gallego and Topaloglu (2014) study the assortment optimization problem with different space constraint which limited the space of product in **each** category. Moreover, Davis, Gallego, and Topaloglu (2014) finds out that the complexity and worst-case performance in the assortment optimization problem differs in model settings. The settings are 1. the range of dissimilarity parameter 2. whether consumer may leave after selecting a nest. [1]

From the aspect of machine learning and deep learning, recommendation systems are widely used in internet document serch, personalized advertisements, and online shopping platform. The first recommendation system was metioned by SALTON (1983) for algorithm searching amongst textual document. Kim et al. (2001) uses decision tree model to generate marketing rules that match customer demographics to product categories. The common flow of recommendation systems is 1. Using the availible user data to create user embedding 2. compute the item embedding 3. use the dot product of user embedding and item embedding to get the relevence score between user-item

[1]The dissimilarity parameter will be discussed in section 3.4

pair 4. ranking the items from hightest to lowest accroding to the relevence score and select top K items to form a recommendation list 5. optional use re-ranking model to select top K_2 items among the top K items.

3 Fundamental Concepts

3.1 Random Utility Model

Discrete choice model is a fundamental approach in econometrics for depicting consumer behavior. Within this framework, the random utility theory permits consumers to exhibit a degree of randomness in their choices among goods. This randomness elucidates why consumers opt for different products within the same choice set. In this research, we assume consumers are homogeneous and let the representative consumer be the example, the utility derived from goods $i \in \{1, 2\}$ respectively is:

$$\begin{aligned} U_1 &= V_1 + \epsilon_1, \\ U_2 &= V_2 + \epsilon_2, \qquad (\epsilon_1, \epsilon_2)^T \sim ((0,0)^T, \Sigma). \end{aligned}$$

U_i represents the utility of consumer for product i, V_i represents the deterministic utility of product i, and (ϵ_i, ϵ_j) represents the random utility of product i and j. Additionally, (ϵ_i, ϵ_j) follows a joint distribution with an expected value of 0. Expressing consumer choices with the choice variable y,

the probability of a consumer choosing product 1 is:

$$\begin{aligned} P(y=1) &= P(U_1 > U_2) = P(\epsilon_1 - \epsilon_2 > V_2 - V_1) \\ &= \int_{\epsilon_1 - \epsilon_2 > V_2 - V_1} f(\epsilon_1, \epsilon_2) d\epsilon_1 d\epsilon_2. \end{aligned}$$

3.2 Logit Model

If $(\epsilon_1, \epsilon_2) \overset{\text{iid}}{\sim}$ type-I Generalized Extreme Value Distribution, then this model degenerates into the logit model, and the probability of a consumer choosing product 1 is:

$$P(y=1) = \frac{e^{V_1}}{e^{V_1} + e^{V_2}}.$$

When we assume $i = 2$ is not buying product and let $V_2 = 0$, this model degenerates into binary logit model. The probability of a consumer choosing product 1 is:

$$P(y=1) = \frac{e^{V_1}}{e^{V_1} + 1}.$$

Furthermore, if the number of products is generalized to n types, $(\epsilon_1, \epsilon_2, ..., \epsilon_n) \overset{\text{iid}}{\sim}$ type-I Generalized Extreme Value Distribution, then this model degenerates into the MNL model, and the probability of a consumer

choosing product i is:

$$\begin{aligned} P(y=i) &= P(U_i \geq U_k), \quad \forall k \in \{1,2,...,n\} \\ &= \frac{e^{V_i}}{\sum_{k=1}^{n} e^{V_k}}. \end{aligned}$$

3.3 Generalized Extreme Value Model

McFadden (1977) introduced the generalized extreme value (GEV) distribution, which extends the logit model to a more general form. The advantage of this distribution lies in its ease of computation for choice probabilities when applied in random utility models. As long as the joint cumulative distribution function of random utilities $(\epsilon_1, \epsilon_2, ..., \epsilon_n)$ satisfies the following equation:

$$F(\epsilon_1, \epsilon_2, ..., \epsilon_n) = \exp[-G(e^{-\epsilon_1}, e^{-\epsilon_2}, ..., e^{-\epsilon_n})]. \tag{1}$$

The sufficient and necessary conditions for aggregate function $G(y_1, y_2, ..., y_n)$ are:

1. Nonnegativity
2. Homogeneity of degree 1
3. Mixed partial derivatives are continuous and nonpositive for even oreder and nonnegativity for odd order
4. $G(y_1, y_2, ..., y_n) \to \infty$ when $y_i \to \infty$

The probability of a consumer choosing product i is:

$$P(y=i) = e^{V_i}\frac{G_i(e^{-V_1}, e^{-V_2}, ...e^{-V_n})}{G(e^{-V_1}, e^{-V_2}, ...e^{-V_n})}, \tag{2}$$
$$G_i(y_1, y_2, ..., y_n) := \frac{\partial G(y_1, y_2, ..., y_n)}{\partial y_i}.$$

Let $G(y_1, y_2, ..., y_n) = \sum_{j=1}^{n} y_j$ be the aggregate functionm, the model degenerates into MNL model.

$$\begin{aligned} P(y=i) &= e^{V_i}\frac{G_i(e^{-V_1}, e^{-V_2}, ...e^{-V_n})}{G(e^{-V_1}, e^{-V_2}, ...e^{-V_n})} \\ &= \frac{e^{V_i}}{\sum_{k=1}^{n} e^{V_k}}. \end{aligned}$$

3.4 Nested Logit (NL) Model

The conventional logit model assumes independence among the random utilities of different products, which is an unreasonable assumption in the real world. Therefore, we allow for correlation among products within the same category (nest) while maintaining independence among products of different categories (nests). For example, products such as screens, graphics cards, keyboards, and headphones, all falling under the category of technology, may exhibit correlated random utilities, whereas the random utility of a screen (in technology category) would be independent of that of chocolate (in food category).

Let's assume all products are divided into M categories (nests), with

the number of products in the m'th group denoted as N_m. Without loss of generality, let's assume $N_1 = N_2 = ... = N_M = N$. In this scenario, a consumer's choice $y \in \mathbf{R}^2$ (i.e., y = (i,j) represents the consumer's selection of the j'th item within the i'th category of products).

Constructing a model that fits this scenario using the GEV model yields the NL model. Its definition and properties are as follows:

Definition 3.1 (Nested logit (NL) model)**.** Under random utility model, if the joint cumulative distribution function of $(\epsilon_{1,1}, \epsilon_{1,2}, ..., \epsilon_{M,N})$ satisfies:

$$\begin{aligned} & F(\epsilon_{1,1}, \epsilon_{1,2}, ..., \epsilon_{M,N}) \\ & = \exp[-G(e^{-\epsilon_{1,1}}, e^{-\epsilon_{1,2}}, ..., e^{-\epsilon_{M,N}})] \\ & = \exp\{-\sum_{m=1}^{M}[\sum_{n=1}^{N}\exp(-\frac{\epsilon_{m,n}}{\tau_m})]^{\tau_m}\}, \end{aligned}$$

this model is called NL model.

The parameters τ_m are called dissimilarity parameters which are required to satisfy $0 \leq \tau_m \leq 1$ for all $m = 1, 2, ..., M$.[2]

Proposition 3.1 (Probability of nested logit (NL) model)**.** Under NL model,

[2]NL model is consistent with the randomn utility model if this condition is satisfied. Davis, Gallego, and Topaloglu (2014) shows that the assortments optimization problem is tractable under this assumption. However, Börsch-Supan (1990) shows that local consistency can be preserved even if this condition is violated.

the probability of consumer choosing item (i,j) is:

$$P(y=(i,j)) = \frac{e^{\frac{V_{i,j}}{\tau_i}}}{\sum_{n=1}^{N} e^{\frac{V_{i,n}}{\tau_i}}} \frac{(\sum_{n=1}^{N} e^{\frac{V_{i,n}}{\tau_i}})^{\tau_i}}{\sum_{m=1}^{M}(\sum_{n=1}^{N} e^{\frac{V_{m,n}}{\tau_m}})^{\tau_m}},$$

this proposition can be proved by equation 2.

Proposition 3.2 (Evaluation of probability)**.** Directly from proposition 3.1,

$$\begin{aligned} P(y=(i,j)) &= \frac{e^{\frac{V_{i,j}}{\tau_i}}}{\sum_{n=1}^{N} e^{\frac{V_{i,n}}{\tau_i}}} \frac{(\sum_{n=1}^{N} e^{\frac{V_{i,n}}{\tau_i}})^{\tau_i}}{\sum_{m=1}^{M}(\sum_{n=1}^{N} e^{\frac{V_{m,n}}{\tau_m}})^{\tau_m}} \\ &= P(y=(i,j) \mid y=(i,*))P(y=(i,*)), \end{aligned}$$

the first fraction is the conditional probability of choosing item (i,j) given that the consumer chooses category i, and the second fraction is the probability of choosing category i.

Proposition 3.3 (Covariance and correlation of nested logit (NL) model)**.** Under NL model, the covariance and corelation of random utility of **different** items (i,j) and (k,l) are:

$$Cov(\epsilon_{i,j}, \epsilon_{k,l}) = \begin{cases} \frac{\pi^2}{6}(1-\tau_i^2), & \text{if i=k} \\ 0, & \text{otherwise} \end{cases}$$

$$Cor(\epsilon_{i,j}, \epsilon_{k,l}) = \begin{cases} 1 - \tau_i^2, & \text{if i=k} \\ 0, & \text{otherwise} \end{cases}$$

4 Methodologies

Suppose there are M nests, each with N alternatives. The benefit of the NL model is that the decision is decomposed into two steps: first, the consumer chooses a nest, then the consumer chooses an alternative within the chosen nest. In the scenario of a recommendation system, the consumer clicks on an ad instead of buying a product. Futhermore, assume the consumers are homogeneous which means their deterministic utility are the same. Moreover, the consumer can choose not to click on any ad, which can also be considered as a kind of ad. The "not-click" ad is a special case which is the only choice in the "not-click" nest. Let the 0'th nest be the "not-click" nest; the utility of a consumer clicking on any of these ads is as follows:

$$U_{i,j} = \begin{cases} V_{i,j} + \epsilon_{i,j}, & \text{click j'th ad of i'th nest} \\ V_0 + \epsilon_0, & \text{not click any ad.} \end{cases} \tag{3}$$

Notice that we only observe the consumer's choice, not the consumer's utility. This means that equation 3 cannot be identified. A common way to solve this problem is to shift the entire model by subtracting V_0 (i.e., let

$\tilde{U}_{i,j} = U_{i,j} - V_0$). For readability, we will omit the tilde in the following discussion. Instead of model 3, we will discuss the following model:

$$U_{i,j} = \begin{cases} V_{i,j} + \epsilon_{i,j}, & \text{click j'th ad of i'th nest} \\ \epsilon_0, & \text{not click any ad.} \end{cases} \tag{4}$$

Another issue is the identification of the dissimilarity parameter τ_i. Heiss (2002) found that if a nest contains only one alternative, the dissimilarity parameter of this nest cannot be identified. In this case, only the "not-click" nest is affected. The solution is rather simple; we set $\tau_0 = 1$.

Combining the discussions above, the probability of a consumer clicking on ad (i,j) is:

$$P(y = (i,j)) = \frac{e^{\frac{V_{i,j}}{\tau_i}}}{\sum_{n=1}^{N} e^{\frac{V_{i,n}}{\tau_i}}} \frac{(\sum_{n=1}^{N} e^{\frac{V_{i,n}}{\tau_i}})^{\tau_i}}{1 + \sum_{m=1}^{M} (\sum_{n=1}^{N} e^{\frac{V_{m,n}}{\tau_m}})^{\tau_m}}, \tag{5}$$

and the probability of a consumer not click on any ad is:

$$P(y = (0,0)) = \frac{1}{1 + \sum_{m=1}^{M} (\sum_{n=1}^{N} e^{\frac{V_{m,n}}{\tau_m}})^{\tau_m}}. \tag{6}$$

The third issue is that the consumer only views a limited number of ads, which means the consumer can only choose from a subset of all ads. Therefore, we define the set of ads (i.e. assortment) that the consumer can choose from as an exposure matrix $\mathbf{S} \in \mathbb{R}^{M \times N}$.

Definition 4.1 (Exposure Matrix)**.** $\mathbf{S} \in \mathbb{R}^{M \times N}$ is the exposure matrix if and only if

$$\mathbf{S_{i,j}} = \begin{cases} 1, & \text{if ad (i,j) is exposed to consumer} \\ 0, & \text{otherwise.} \end{cases}$$

Furthermore, let $\mathbf{s_i} \in \mathbb{R}^N$ denote the exposure vector that the consumer can choose from nest i. The relation between $\mathbf{S}$ and $\mathbf{s_i}$ is that $\mathbf{s_i} = \mathbf{S}^T \mathbf{e_i}$, where $\mathbf{e_i} \in \mathbb{R}^M$ is a vector with 1 at the i-th position and 0 elsewhere.

In the following sections, we will discuss several cases from simple to complex. The first case is that the consumer is an all-knowing being, which means the consumer can observe any number of ads with ease. The second case is that the consumer is a normal being, and the exposure constraint is that the consumer can observe a limited number of ads per nest. The third case is about the substitute effect within ads. We will discuss the effect of another ad being exposed to the consumer and the impact on the probability of the consumer clicking on the original ad. The fourth case is the more familiar case, where the consumer can only observe a limited number of ads in total.

4.1 Case 1: All-knowing Consumer

The goal is to maximize the probability of the consumer click on any ad. In other words, we aim to minimize the probability of consumer not click on

any ad. The optimization problem is:

$$\begin{aligned} \text{Objective funtion:} \qquad & \min_{\mathbf{S}} P(y=(0,0)) \\ & = \frac{1}{1+\sum_{m=1}^{M}(\sum_{n=1}^{N}\mathbf{S_{m,n}}e^{\frac{V_{m,n}}{\tau_m}})^{\tau_m}} \end{aligned} \tag{7}$$

To achieve notation simplicity, we define $\mathbf{H} \in \mathbb{R}^{M\times N}$ as the exponent utility matrix.

Definition 4.2 (Exponent Utility Matrix)**.** $\mathbf{H} \in \mathbb{R}^{M\times N}$ is the exponent utility matrix if and only if

$$\mathbf{H_{i,j}} = e^{\frac{V_{i,j}}{\tau_i}}.$$

Similarly, we define $\mathbf{h_i} \in \mathbb{R}^N$ as the exponent utility vector of nest i. The relationship between $\mathbf{H}$ and $\mathbf{h_i}$ is such that $\mathbf{h_i} = \mathbf{H}^T\mathbf{e_i}$. Therefore, we can rewrite the optimization problem as follows:

$$\begin{aligned} \text{Objective funtion:} \qquad & \min_{\mathbf{S}} P(y=(0,0)) \\ & = \frac{1}{1+\sum_{m=1}^{M}(\mathbf{h_m}^T\mathbf{s_m})^{\tau_m}} \end{aligned} \tag{8}$$

The optimal solution $\mathbf{S}^*$ is:

$$\begin{aligned} \mathbf{S}^* &= \arg\min_{\mathbf{S}} P(y=(0,0)) = \arg\min_{\mathbf{S}} \frac{1}{1+\sum_{m=1}^{M}(\mathbf{h_m}^T\mathbf{s_m})^{\tau_m}} \\ &= \arg\max_{\mathbf{S}} \sum_{m=1}^{M}(\mathbf{h_m}^T\mathbf{s_m})^{\tau_m}. \end{aligned} \tag{9}$$

Since the components of the exponent utility vector $\mathbf{h_m}$ are all positive, combined with the fact that the components of the exposure vector $\mathbf{s_m}$ are all binary, the optimal solution should be for the all-knowing consumer to be exposed to all ads to achieve the maximum probability of the consumer clicking on any ad.

$$\mathbf{S^*_{i,j}} = 1, \quad \forall i \in \{1, 2, ..., M\}, j \in \{1, 2, ..., N\}. \tag{10}$$

4.2 Case 2: Limited Exposure in each Nest

Case 2 is case 1 with the exposure consraint. The problem can be written as:

$$\begin{aligned} \text{Objective funtion:} \quad & \min_{\mathbf{S}} P(y = (0,0)) = \frac{1}{1 + \sum_{m=1}^{M} (\mathbf{h_m}^T \mathbf{s_m})^{\tau_m}}, \\ \text{Subject to:} \quad & \sum_{n=1}^{N} \mathbf{s_{m,n}} \leq K, \quad \forall m \in \{1, 2, ..., M\}, \\ & \text{where K is some integer less than N.} \end{aligned} \tag{11}$$

Case 1 implies that the inner solution (i.e., $\sum_{n=1}^{N} \mathbf{s_{m,n}} < K$) can never be optimal. Hence, we focus specifically on the bounded solution. The optimal solution is rather simple as well. Without loss of generality, we can reorder the components of $\mathbf{H}$ such that $\mathbf{H_{m,1}} \geq \mathbf{H_{m,2}} \geq ... \geq \mathbf{H_{m,N}}, \quad \forall m \in$

$\{1, 2, ..., M\}$. The optimal solution is:

$$\mathbf{S}^*_{\mathbf{m,n}} = \begin{cases} 1, & \text{if n} \leq \text{K} \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

The intepratation is that the consumer should be exposed to the K highest exponent utility ads in each nest.

4.3 Case 3: Substitute Effect

Here, we discuss the effect of exposing another ad to the consumer and its impact on the probability of the consumer clicking on the original ad. We define the substitute effect of ad (o,p) on ad (i,j) as:

Definition 4.3 (Substitute Effect)**.**

$$\text{SE}[(o,p) \to (i,j)] = P(y = (i,j) \mid \mathbf{S_{o,p}} = 1) - P(y = (i,j) \mid \mathbf{S_{o,p}} = 0).$$

The substitute effect can be calculated by the following formula

$$P(y = (i,j) \mid \mathbf{S_{o,p}} = 0) = \frac{\mathbf{H_{i,j}}}{\mathbf{h_i}^T \mathbf{s_i}} \frac{(\mathbf{h_i}^T \mathbf{s_i})^{\tau_i}}{1 + \sum_{m=1}^{M} (\mathbf{h_m}^T \mathbf{s_m})^{\tau_m}}$$

$$P(y = (i,j) \mid \mathbf{S_{o,p}} = 1) = \begin{cases} \frac{\mathbf{H_{i,j}}}{\mathbf{h_i}^T \mathbf{s_i}} \frac{(\mathbf{h_i}^T \mathbf{s_i})^{\tau_i}}{1 + \sum\limits_{m=1, m \neq o}^{M} (\mathbf{h_m}^T \mathbf{s_m})^{\tau_m} + (\mathbf{h_o}^T \mathbf{s_o} + \mathbf{H_{o,p}})^{\tau_o}}, & o \neq i, \\ \frac{\mathbf{H_{i,j}}}{\mathbf{h_i}^T \mathbf{s_i} + \mathbf{H_{o,p}}} \frac{(\mathbf{h_i}^T \mathbf{s_i} + \mathbf{H_{o,p}})^{\tau_i}}{1 + \sum\limits_{m=1, m \neq o}^{M} (\mathbf{h_m}^T \mathbf{s_m})^{\tau_m} + (\mathbf{h_o}^T \mathbf{s_o} + \mathbf{H_{o,p}})^{\tau_o}}, & o = i. \end{cases}$$

Assume $\mathbf{H_{o,p}}$ is rather small, the first order Taylor expansion of $P(y=(i,j) \mid \mathbf{S_{o,p}}=1)$ at $\mathbf{H_{o,p}}=0$ is:

$$
\begin{aligned}
&P(y=(i,j) \mid \mathbf{S_{o,p}}=1) \\
&\approx \begin{cases} P(y=(i,j) \mid \mathbf{S_{o,p}}=0)[1-\frac{\mathbf{H_{o,p}}}{\mathbf{h_o}^T\mathbf{s_o}}\frac{\tau_o(\mathbf{h_o}^T\mathbf{s_o})^{\tau_o}}{1+\sum_{m=1}^{M}(\mathbf{h_m}^T\mathbf{s_m})^{\tau_m}}], & o \neq i, \\ P(y=(i,j) \mid \mathbf{S_{o,p}}=0)[1-\frac{\mathbf{H_{o,p}}}{\mathbf{h_o}^T\mathbf{s_o}}(\frac{\tau_o(\mathbf{h_o}^T\mathbf{s_o})^{\tau_o}}{1+\sum_{m=1}^{M}(\mathbf{h_m}^T\mathbf{s_m})^{\tau_m}}+(1-\tau_o))], & o = i. \end{cases}
\end{aligned} \tag{13}
$$

From equation 13, we can derive the substitute effect of ad (o,p) on ad (i,j) as:

$$
\begin{aligned}
&\mathrm{SE}[(o,p) \to (i,j)] = P(y=(i,j) \mid \mathbf{S_{o,p}}=1) - P(y=(i,j) \mid \mathbf{S_{o,p}}=0) \\
&\approx \begin{cases} -P(y=(i,j) \mid \mathbf{S_{o,p}}=0)\frac{\mathbf{H_{o,p}}}{\mathbf{h_o}^T\mathbf{s_o}}\frac{\tau_o(\mathbf{h_o}^T\mathbf{s_o})^{\tau_o}}{1+\sum_{m=1}^{M}(\mathbf{h_m}^T\mathbf{s_m})^{\tau_m}}, & o \neq i, \\ -P(y=(i,j) \mid \mathbf{S_{o,p}}=0)\frac{\mathbf{H_{o,p}}}{\mathbf{h_o}^T\mathbf{s_o}}[\frac{\tau_o(\mathbf{h_o}^T\mathbf{s_o})^{\tau_o}}{1+\sum_{m=1}^{M}(\mathbf{h_m}^T\mathbf{s_m})^{\tau_m}}+(1-\tau_o)], & o = i. \end{cases}
\end{aligned} \tag{14}
$$

The result is actually intuitive. First, the substitute effect of ad (o,p) on ad (i,j) is negative in both cases, which means that the probability of a consumer clicking on ad (i,j) decreases when another ad is exposed to the consumer. This makes sense because in the NL model, all the ads are substitutes for each other with respect to probability. Second, the magnitude of the substitute effect is larger (the $1-\tau_o$ term in equation 14) when the old ad shares the same nest with the new ad. This is reasonable because the ads in the same nest are more similar to each other, which means the consumer

is more likely to substitute the old ad with the new one. For example, the substitute effect of a new iPhone ad on a Samsung mobile ad is larger than the substitute effect of a new iPhone ad on a cereal ad due to the fact that the iPhone and Samsung mobile are both in the smartphone category, but cereal isn't.

4.4 Case 4: Limited Exposure in Total

Case 4 is case 2 with a different exposure consraint. The problem can be written as:

$$\begin{aligned} \text{Objective funtion:} \qquad & \min_{\mathbf{S}} P(y=(0,0)) = \frac{1}{1+\sum_{m=1}^{M}(\mathbf{h_m}^T\mathbf{s_m})^{\tau_m}}, \\ \text{Subject to:} \qquad & \sum_{m=1}^{M}\sum_{n=1}^{N}\mathbf{S_{m,n}} \leq K,\ K\in\mathbb{N},\ K\leq M\times N. \end{aligned} \tag{15}$$

The constraint can be written as $\sum_{m=1}^{M}\mathbf{s_m}^T\mathbf{s_m} \leq K$, which seems like the familiar L2 regularization in machine learning. The difference is that the L2 regularization is a continuous constraint while the constraint in the random utility problem is a discrete constraint. This kind of problem is called the knapsack problem in combinatorial optimization and

Lemma 4.1. Knapsack problem is NP-complete.

Proof. Refer to Garey, Johnson, et al. (1990). □

Nonetheless, in our setting, the optimal solution $\mathbf{S}^*$ is P problem.

In the first step, sort the components of each row of $\mathbf{H}$ in descending order i.e., $\mathbf{H_{m,1}} \geq \mathbf{H_{m,2}} \geq ... \geq \mathbf{H_{m,N}}, \quad \forall m \in \{1, 2, ..., M\}$. To proceed further, we need the following lemma.

Lemma 4.2 (Monotonicity of Choice)**.** $\mathbf{S}^*$ is the optimal solution. If $\mathbf{S^*_{i,j}} = 1$, then

$$\mathbf{S^*_{i,j-1}} = 1$$

Proof. The proof is by contradiction. Suppose $\mathbf{S^*_{i,j-1}} = 0$, then we can construct a new solution $\mathbf{S'}$ by setting $\mathbf{S'_{i,j}} = 0$ and $\mathbf{S'_{i,j-1}} = 1$. Since the objective function is the same as case 1, we can rewrite the objective function as

$$\text{Alternative Objective funtion:} \qquad \max_{\mathbf{S}} \sum_{m=1}^{M} (\mathbf{h_m}^T \mathbf{s_m})^{\tau_m}. \tag{16}$$

The difference between alternative value functions achieved by $\mathbf{S'}$ and $\mathbf{S}^*$ is:

$$\begin{aligned}
&\sum_{m=1}^{M} (\mathbf{h_m}^T \mathbf{s'_m})^{\tau_m} - \sum_{m=1}^{M} (\mathbf{h_m}^T \mathbf{s^*_m})^{\tau_m} \\
&= (\mathbf{h_i}^T \mathbf{s'_i})^{\tau_i} - (\mathbf{h_i}^T \mathbf{s^*_i})^{\tau_i} \\
&= (\sum_{n \neq j, n \neq j-1}^{N} \mathbf{H_{i,n}} \mathbf{S'_{i,n}} + \mathbf{H_{i,j-1}})^{\tau_i} - (\sum_{n \neq j, n \neq j-1}^{N} \mathbf{H_{i,n}} \mathbf{S^*_{i,n}} + \mathbf{H_{i,j}})^{\tau_i} \\
&= (\sum_{n \neq j, n \neq j-1}^{N} \mathbf{H_{i,n}} \mathbf{S^*_{i,n}} + \mathbf{H_{i,j-1}})^{\tau_i} - (\sum_{n \neq j, n \neq j-1}^{N} \mathbf{H_{i,n}} \mathbf{S^*_{i,n}} + \mathbf{H_{i,j}})^{\tau_i} > 0.
\end{aligned}$$

Therefore, the alternative solution $\mathbf{S'}$ is better than the optimal solution $\mathbf{S}^*$, which contradicts the assumption that $\mathbf{S}^*$ is the optimal solution. □

Lemma 4.3 (Bounded Constraint)**.** Optimal solution $\mathbf{S}^*$ always satisfies the bounded constraint i.e.

$$\sum_{m=1}^{M}\sum_{n=1}^{N}\mathbf{S^*_{m,n}} = K.$$

Proof. This is proof by contradiction. Suppose $\sum_{m=1}^{M}\sum_{n=1}^{N}\mathbf{S^*_{m,n}} < K$, then we can construct a new solution $\mathbf{S'}$ by setting $\mathbf{S'_{i,j}} = 1$ where $\mathbf{S^*_{i,j}} = 0$. Using the alternative objective function in lemma 4.2, we can show that the alternative solution $\mathbf{S'}$ is better than the optimal solution $\mathbf{S}^*$, which contradicts the assumption that $\mathbf{S}^*$ is the optimal solution. □

By Lemma 4.2, we can conclude that the optimal solution $\mathbf{S}^*$ is a matrix in which the components are all 1 on the left side and 0 on the right side. The number of 1's is exactly K, as proven by Lemma 4.3. The last piece of the puzzle is to find the distribution of 1's in each row. To achieve this, we need to construct a new variable $\Delta_{i,j}$, which denotes the increment of the alternative objective function when $\mathbf{S_{i,j}}$ is changed from 0 to 1. By Lemma 4.2, we only need to consider $\mathbf{S_{i,j}}$ when $\mathbf{S_{i,1}}, \mathbf{S_{i,2}}, ..., \mathbf{S_{i,j-1}}$ are all 1. Therefore, the formal definition of $\Delta_{i,j}$ is:

Definition 4.4 (Increment Variable)**.**

$$\Delta_{i,j} = \begin{cases} (\sum_{n=1}^{j}\mathbf{H_{i,n}})^{\tau_i} - (\sum_{n=1}^{j-1}\mathbf{H_{i,n}})^{\tau_i}, & \text{if j} \neq 1, \\ \mathbf{H_{i,1}}^{\tau_i}, & \text{if j} = 1. \end{cases}$$

Given that $\tau_m \leq 1, \forall m \in \{1, 2, ..., M\}$, the increment variable $\Delta_{i,j}$ is non-

increasing in j. Therefore, the optimal solution $\mathbf{S}^*$ is determined by the first K largest increment variables. The optimal solution $\mathbf{S}^*$ is:

$$\mathbf{S}^*_{\mathbf{i,j}} = \begin{cases} 1, & \text{if } \Delta_{i,j} \text{ is one of the K largest increment variables,} \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

The complete algorithm is as follows:

0. Identify the nest structure of the model, the deterministic utility of ads, and the dissimilarity parameters τ_i.[3]

1. Calculate the exponent utility matrix $\mathbf{H}$ and sort the components of each row in descending order.

2. Calculate the increment variables $\Delta_{i,j}$.

3. Find the optimal solution $\mathbf{S}^*$ by selecting the first K largest increment variables across all $\Delta_{i,j}$.

The time complexity of the algorithm is $O(MN)$, which is linear to the number of total ads. The algorithm is easy to implement and interpret. The results can be used in the recommendation system to maximize the probability of a consumer clicking on any ad.

[3]The nest structure can be identified by the domain knowledge. Besides, Hausman (1978) proposed Hausman's specification test which is used to identify the worngly classified products. The deterministic utility and τ_i can be identified by the sequential estimation proposed by Train (2002).

5 Diversity and Fairness

According to Zhao et al. (2024), the diversity and fairness are two different concepts with similiar goal. Fairness focuses on

1. Popularity Bias: the exposure of similar items should be fair,
2. Quality Discrepancy: the recommendation quality for different users/ groups should be similar, and
3. Discriminator Performance: the recommendation result should not take account of sensitive attributes (e.g. gender or sexual orientation).

On the other hand, diversity are categorized into two types:

1. Individual Diversity: the recommendation list should contain a variety of items, and
2. Aggregate Diversity: the recommendation list vary across different users/groups which means recommendation system is able to recommend less popular items to users.

There are two type of metric to measure the diversity and fairness. The first one is measuring the average distance of item embedding pairs within recommendation list i.e. Intra-List Diversity (ILD). Ziegler et al. (2005) start from a reverse aspect and propose Intra-List Similarity (ILS) and choose among items with least similarity. Vargas et al. (2014) critisize the distance metric because of inconsistency. Vargas et al. (2014) leverage the genre

information and propose a novel metric called BinomDiv($\mathbf{S}$) = Coverage($\mathbf{S}$)$\times$ NonRed($\mathbf{S}$). Coverage is the number of genres in the recommendation list and NonRed (i.e. non-redundancy) is the number of genres that are not repeated in the recommendation list. As for aggregate diversity, the most common metric are total number of unique items, Shanon entropy, and Gini index.

Definition 5.1 (Intra-List Diversity)**.**

$$ILD(\mathbf{S}) = \frac{1}{|\mathbf{S}|(|\mathbf{S}|-1)} \sum_{\mathbf{S}_{i',j'} \neq 0} \sum_{\mathbf{S}_{i,j} \neq 0, (i,j) \neq (i',j')} d(i,j,i',j'),$$

where $d(i,j,i',j')$ is the distance between item (i, j) and item (i', j').

Definition 5.2 (Shannon Entropy)**.**

$$H = -\sum_{i,j}^{N,M} freq(i,j) \log freq(i,j),$$

where freq(i,j) is the frequency of item (i,j) in the lists across users.

To achieve diversity, re-ranking is a common approach for the recommendation system. Combine the relevence score with diversity score, the system re-rank the recommendation list. Adding an item one by one with greedy algorithm, it can easily find the suboptimal solution. Many researches adapt this method (Santos, Macdonald, and Ounis (2010), Smyth and McClave (2001)).Likewise, learning-to-rank is another approach to achieve di-

versity by adding a regularization term to the loss function. There are one addtional approach utilizing fusion-based model to achieve diversity. Fusion-based model aggregate multiple recommendation scores to generate a list of recommendation. Ribeiro et al. (2012) combines the scores by weighted sum, then use strength Pareto evolutionary algorithm to optimize the weights.

6 Conclusion

In this paper, we aim to maximize the opportunity for a representative consumer to click on any ads. We first discuss the all-knowing consumer case and show that the optimal solution is to expose the consumer to all ads. Then we discuss the limited exposure in each nest case and show that the optimal solution is to expose the consumer to the K highest exponent utility ads in each nest. We also discuss the substitute effect of ads and show that the probability of a consumer clicking on an ad decreases when another ad is exposed to the consumer. The magnitude of the substitute effect is larger when the old ad shares the same nest with the new ad. Finally, we discuss the limited exposure in total case and show that the optimal solution is to expose the consumer to the K highest increment variables across all ads. The optimal solution is determined by the first K largest increment variables. An algorithm is provided to find the optimal solution. The results can be used in the recommendation system to maximize the probability of a consumer clicking on any ad.

However, there are some limitations in our paper. First, we assume consumers are homogeneous. The advantage of this assumption is data sufficiency because researchers can use all of the consumers' data to estimate the parameters. The drawback is the loss of flexibility because all of the consumers are exposed to the same ad set (i.e. assortment). The potential solution may lie in Bayesian statistics. The result from the nested model can be treated as a prior, and the user's history data can be treated as a likelihood function used to update the recommendation ad set over time. Second, model misspecification is another issue. Since we use the NL model, every ad is in exactly one nest, unlike the real world. To relax this assumption, one may adopt a more general model like the cross-nested logit (CNL) model Vovsha (1997). Third, the model is static. The optimal solution is determined by the current utility of ads. The ideal model should be able to capture the time-varying utility of ads. The time-varying and customized recommendation system is a potential research direction, and Bayesian statistics may be a good tool to solve this problem.

In conclusion, we proposed a model to improve the click-through rate of ads in the recommendation system. The advantages of our model are 1. backed up by the random utility model, 2. consider the dependencies among similar ads, 3. have closed form solution, 4. the time complexity is linear to to number of total ad, and 5. easy to interpretate. The disadvantage of our model are 1. homogeneous consumer assumption which leads to no customization, 2. model and structure specification, and 3. static model.

References

Abdollahpouri, Himan et al. (2019). "The Unfairness of Popularity Bias in Recommendation". In: *CoRR* abs/1907.13286. arXiv: 1907.13286. URL: http://arxiv.org/abs/1907.13286.

Börsch-Supan, Axel (1990). "On the compatibility of nested logit models with utility maximization". In: *Journal of Econometrics* 43.3, pp. 373–388.

Davis, James M, Guillermo Gallego, and Huseyin Topaloglu (2014). "Assortment optimization under variants of the nested logit model". In: *Operations Research* 62.2, pp. 250–273.

Gallego, Guillermo and Huseyin Topaloglu (2014). "Constrained assortment optimization for the nested logit model". In: *Management Science* 60.10, pp. 2583–2601.

Garey, Michael R, David S Johnson, et al. (1990). "A Guide to the Theory of NP-Completeness". In: *Computers and intractability*, pp. 37–79.

Hausman, Jerry A (1978). "Specification tests in econometrics". In: *Econometrica: Journal of the econometric society*, pp. 1251–1271.

Heiss, Florian (2002). "Structural choice analysis with nested logit models". In: *The Stata Journal* 2.3, pp. 227–252.

Hensher, David A and William H Greene (2002). "Specification and estimation of the nested logit model: alternative normalisations". In: *Transportation Research Part B: Methodological* 36.1, pp. 1–17.

Kim, Jong Woo et al. (2001). "Application of Decision-Tree Induction Techniques to Personalized Advertisements on Internet Storefronts". In: *International Journal of Electronic Commerce* 5.3, pp. 45–62. ISSN: 10864415, 15579301. URL: http://www.jstor.org/stable/27750981 (visited on 08/24/2024).

McFadden, Daniel (1972). "Conditional logit analysis of qualitative choice behavior". In.

— (1977). "Modelling the choice of residential location". In.

Ribeiro, Marco Tulio et al. (2012). "Pareto-efficient hybridization for multi-objective recommender systems". In: *Proceedings of the Sixth ACM Conference on Recommender Systems*. RecSys '12. Dublin, Ireland: Association for Computing Machinery, pp. 19–26. ISBN: 9781450312707. DOI: 10.1145/2365952.2365962. URL: https://doi.org/10.1145/2365952.2365962.

Rusmevichientong, Paat, Zuo-Jun Max Shen, and David B Shmoys (2009). "A PTAS for capacitated sum-of-ratios optimization". In: *Operations Research Letters* 37.4, pp. 230–238.

SALTON, G. (1983). "Introduction to modern information retrieval". In: *McGraw-Hill*. URL: https://cir.nii.ac.jp/crid/1574231873785747328.

Santos, Rodrygo L.T., Craig Macdonald, and Iadh Ounis (2010). "Exploiting query reformulations for web search result diversification". In: *Proceedings of the 19th International Conference on World Wide Web*. WWW '10.

Raleigh, North Carolina, USA: Association for Computing Machinery, pp. 881–890. ISBN: 9781605587998. DOI: 10.1145/1772690.1772780. URL: https://doi.org/10.1145/1772690.1772780.

Smyth, Barry and Paul McClave (2001). "Similarity vs. Diversity". In: *Case-Based Reasoning Research and Development*. Ed. by David W. Aha and Ian Watson. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 347–361. ISBN: 978-3-540-44593-7.

Train, Kenneth E (2002). *Discrete choice methods with simulation*. Cambridge university press.

Vargas, Saúl et al. (2014). "Coverage, redundancy and size-awareness in genre diversity for recommender systems". In: *Proceedings of the 8th ACM Conference on Recommender Systems*. RecSys '14. Foster City, Silicon Valley, California, USA: Association for Computing Machinery, pp. 209–216. ISBN: 9781450326681. DOI: 10.1145/2645710.2645743. URL: https://doi.org/10.1145/2645710.2645743.

Vovsha, Peter (1997). "Application of cross-nested logit model to mode choice in Tel Aviv, Israel, metropolitan area". In: *Transportation Research Record* 1607.1, pp. 6–15.

Williams, Huw CWL (1977). "On the formation of travel demand models and economic evaluation measures of user benefit". In: *Environment and planning A* 9.3, pp. 285–344.

Zhao, Yuying et al. (2024). *Fairness and Diversity in Recommender Systems: A Survey*. arXiv: 2307.04644 [cs.IR]. URL: https://arxiv.org/abs/2307.04644.

Ziegler, Cai-Nicolas et al. (2005). "Improving recommendation lists through topic diversification". In: *Proceedings of the 14th International Conference on World Wide Web*. WWW '05. Chiba, Japan: Association for Computing Machinery, pp. 22–32. ISBN: 1595930469. DOI: 10.1145/1060745.1060754. URL: https://doi.org/10.1145/1060745.1060754.