On (Re-Scaled) Multi-Attempt Approximation of Customer Choice Model and its Application to Assortment Optimization

Hakjin Chung
Stephen M. Ross School of Business
University of Michigan

Hyun-Soo Ahn
Stephen M. Ross School of Business
University of Michigan

Stefanus Jasin
Stephen M. Ross School of Business
University of Michigan

Ross School of Business Working Paper
Working Paper No. 1322
June 2016

On (Re-scaled) Multi-Attempt Approximation of Customer Choice Model and Its Application to Assortment Optimization

Hakjin Chung, Hyun-Soo Ahn, Stefanus Jasin
Stephen M. Ross School of Business, University of Michigan
Ann Arbor, MI 48109, hakjin, hsahn, sjasin@umich.edu

Motivated by the classic exogenous demand model and the recently developed Markov chain model, we propose a new approximation to the general customer choice model based on random utility called *multi-attempt* model, in which a customer may consider several substitutes before finally deciding to not purchase anything. We show that the approximation error of multi-attempt model decreases exponentially in the number of attempts. However, despite its strong theoretical performance, the empirical performance of multi-attempt model is not satisfactory. This motivates us to construct a modification of multi-attempt model called *re-scaled multi-attempt* model. We show that re-scaled 2-attempt model is exact when the underlying true choice model is Multinomial Logit (MNL); if, however, the underlying true choice model is not MNL, we show numerically that the approximation quality of re-scaled 2-attempt model is very close to that of Markov chain model. The key feature of our proposed approach is that the resulting approximate choice probability can be explicitly written. From a practical perspective, this allows the decision maker to use off-the-shelf solver, or borrow existing algorithms from literature, to solve a general assortment optimization problem with a variety of real-world constraints.

Key words:
History:

1. Introduction

Assortment optimization is one of the most important problems in operations and marketing; it is both mathematically challenging and practically prevalent. Despite a few decades of research on the topic, the pursuit of a new approach that can efficiently solve a general assortment optimization problem that takes into account a wide variety of real-world business constraints is still very vibrant. There are many reasons why assortment optimization is difficult. First, the estimation of customer

choice behavior itself is far from trivial — it continues to be one of the most important topics in the academic literature (Hess and Daly 2014). Second, even after a customer choice model has been successfully estimated, the resulting model is sometimes difficult to optimize, which limits how a decision maker can operationalize assortment, pricing, and inventory decisions based on the solution of a model. Our work is motivated by these very concerns. Working under the framework of mixed logit model (this assumption is without loss of generality since McFadden et al. (2000) show that any random utility model can be approximated to any degree of accuracy by a mixed logit model), we propose an approximation scheme that improves the approximation quality of the so-called *exogenous* demand model (see below) and show that this approximation can be potentially used to solve a general assortment optimization problem with a wide variety of real-world constraints.

Exogenous demand model is perhaps the most popular choice model used in the operations literature. It assumes that each customer behaves in the following way: when faced with an assortment of products (an offer set), she first looks for her favorite product in the assortment; if this product is not available, she considers a substitute product, and if this substitute product is also not available, she will not purchase anything. Since a customer is only making two attempts when purchasing a product, we also call this 2-attempt model. As explained in Kök and Fisher (2007), the assumption that a customer does not consider further substitutions in her search is not necessarily restrictive, at least in some settings. The strengths of exogenous demand model are obvious: Not only it is intuitively appealing, it also provides a *tractable* estimation and optimization framework. (We are not aware of works that study the theoretical complexity of assortment optimization under exogenous demand model. However, per our experience on running numerical experiments with exogenous demand model, its Mixed Integer Program (MIP) formulation can be solved very efficiently within a few seconds for a reasonable sized problem; see Table 2.) The main weakness of exogenous demand model is that, when many customers are willing to consider more than one substitute, it does not necessarily provide an accurate approximation of the true choice model. Thus, an important research question is how to improve the approximation quality of exogenous demand model without significantly compromising its strengths.

The most relevant advancement to the above question that we are aware of is made recently by Blanchet et al. (2016). (There exist other approximation schemes such as ranking-based model, see Bertsimas and Mišic (2016) for literature review; however, these models are not the focus of our work.) The authors propose an iterative Markov search model where a customer does not stop after the first substitution attempt but continues searching until either she finds a product that she likes within the assortment or she hits the *no-purchase* option. More precisely, they interpret the substitution probability as a transition probability in a Markov chain where both the no-purchase option and the set of products in the assortment act as the absorbing states. The authors show that their proposed model approximates the true choice model well, and they develop a polynomial-time algorithm to solve the corresponding unconstrained assortment optimization problem. Although they do not benchmark the performance of Markov chain model against the exogenous demand model, we show using numerical experiments in Section 2.3 that the former significantly improves the accuracy of the later. Moreover, since Markov chain model requires exactly the same number of parameters as exogenous demand model, it is as tractable as exogenous demand model from the estimation perspective.

The main drawback of Markov chain model is that its corresponding choice probability cannot be explicitly written. This makes it rather difficult for practitioners to use Markov chain model in conjunction with off-the-shelf optimization solvers, especially for the setting of assortment optimization with constraints. As noted in Bertsimas and Mišic (2016), firms typically have many business rules that limit the set of possible assortments. To name a few, a firm may have a limited shelf space which dictates that only a finite number of products can be displayed at any time; a firm may require that some products be offered together; and, a firm may also require that only a number of products within a certain category to be offered at any time, etc. There are two typical approaches taken by researchers to solve assortment optimization problem with constraints. The first approach is what Bertsimas and Mišic (2016) call the *fix-then-exploit* approach, where the researchers first *fix* a particular choice model and then *exploit* the structure of the resulting assortment problem to develop either an exact or approximate solution (e.g., Rusmevichientong

et al. 2010 and Désir et al. 2015). The second approach is the so-called *Mixed Integer Optimization* (MIO) approach where an assortment problem is formulated as an MIP (or its variant) and then solved using an off-the-shelf MIP solver; this approach typically requires that the corresponding choice probability can be explicitly written. Note that while the fix-then-exploit approach allows researchers to develop a highly efficient algorithm for a specific model, the MIO approach is highly flexible in the sense that no problem-specific effort to develop a specialized algorithm is required and practitioners can simply *declare* their constraints to the solver. Since the approximate choice probability under Markov chain model cannot be explicitly written, a specific algorithm needs to be developed to solve a constrained assortment optimization problem under Markov chain model. Indeed, this is the approach taken by Feldman and Topaloglu (2014) and Désir et al. (2015), where the authors focus on specific forms of constraints. In particular, Feldman and Topaloglu (2014) develop a linear programming algorithm in the context of the network revenue management problem, and Désir et al. (2015) develop constant factor approximations for assortment optimization problem with the cardinality and capacity constraints. In contrast to this, constrained assortment optimization problems under exogenous demand model, at least for the types of constraint discussed above, can be easily formulated as a Mixed Integer Linear Program (MILP) and solved using an off-the-shelf solver. (Per our numerical experiments in Section 3, the resulting MILP can be solved within a few seconds for a reasonable sized problem.)

Our contribution. In this work, we wish to bridge the gap between the classical exogenous demand model and the recently introduced Markov chain model. The central question we ask is whether it is possible to improve the approximation quality of exogenous demand model without sacrificing its tractability and versatility in dealing with real-world constraints. We are particularly interested in a type of approximation whose corresponding choice probability can be explicitly written as it allows practitioners to simply use off-the-shelf optimization solvers to solve a variety of constrained assortment problems without having to develop a specific algorithm for a specific set of constraints. Thus, our work shares the same spirit as the recent work of Bertsimas and Mišic (2016). (Our work differs from theirs in that they use a ranking-based approximation whereas we

use a new multi-attempt approximation.) We first study the approximation quality of a natural generalization of exogenous demand model, called *multi-attempt* model. To be precise, assuming that all customers are willing to consider at most $k-1$ substitutes, how much improvement does this extra flexibility give, in general, as a function of k? We show that the approximation error of multi-attempt model relative to the true choice probability decreases exponentially in k. This confirms our intuition that capturing higher substitution dynamics leads to a better approximation. Unfortunately, while the theoretical bound of multi-attempt model is encouraging, its empirical performance is somewhat discouraging as it heavily depends on the number of products n. (Per our results in Table 1, for $n = 10$, 4-attempt model is better than Markov chain model; for $n = 100$, even 5-attempt model is still a lot worse than Markov chain model. This is not satisfactory because k-attempt model with $k \geq 3$ requires a lot more parameters than Markov chain model.) Upon a closer examination, however, it turns out that multi-attempt model consistently underestimates the true choice probability, which leads to its poor empirical performance. This motivates us to construct a modified multi-attempt model, which we call *re-scaled multi-attempt* model. The idea is to start with the original k-attempt model and then re-scale it with a non-constant factor to make the sum of probability equals one. The proposed re-scaling significantly improves the performance of the original multi-attempt model: If the true choice model is Mutinomial Logit (MNL), we show that re-scaled k-attempt model is *exact* for all $k \geq 1$ (this result is reminiscent of the result in Blanchet et al. (2016) that Markov chain model is exact for MNL); if, on the other hand, the true choice model is not MNL, our numerical experiments show that the approximation quality of re-scaled 2-attempt model is very close to Markov chain model and the approximation quality of re-scaled 3-attempt consistently dominates the Markov chain model (see Table 1).

Both re-scaled 2-attempt and Markov chain models share exactly the same number of parameters; and yet, the corresponding choice probability under re-scaled 2-attempt model can be explicitly written. This allows us to more easily formulate an assortment optimization problem with constraints. In Section 3, we show that the resulting constrained assortment optimization problems (with typical constraints discussed before) under re-scaled 2-attempt model can be written as a

Mixed Integer Linear Fractional Program (MILFP). Although MILFP in general is difficult to solve, many important problems in engineering and science can be formulated as MILFPs; these have motivated intensive researches in the scientific community to develop efficient methods (both exact and approximate) for solving large-scale MILFPs (e.g., Tawarmalani and Sahinidis 2002 and Yue et al. 2013). On another note, the MILFP formulation of assortment optimization under re-scaled 2-attempt model can be equivalently transformed into a 0-1 quadratic programming. Again, although 0-1 quadratic programming is in general difficult to solve (i.e., from theoretical complexity perspective), we do have a 50-year deep of literature on the topic of approximation algorithm for 0-1 quadratic programming (e.g., Kochenberger et al. 2014). Thus, we are not lacking of sophisticated algorithms that can be used to solve the resulting assortment problem under our proposed approach. Indeed, this is another advantage of having an explicit expression of approximate choice probability as it allows us to borrow tools from existing literature in addition to using off-the-shelf solvers. For the purpose of numerical illustrations, in this work, we will only focus on one approach, the so-called *Dinkelbach algorithm*. We discuss this in more detail in Section 3.

2. Choice Approximation Models

In this section, we describe both the *multi-attempt* and *re-scaled multi-attempt* models. In addition, we also provide results from numerical experiments to compare the approximation accuracy of these models with Markov chain model. We denote the universe of n products by the set $\mathcal{N} = \{1, ..., n\}$ and the no-purchase alternative as product 0. Since McFadden et al. (2000) show that any random utility choice model can be approximated by a mixture of Multinomial Logits (MNLs) at any degree of accuracy, we will assume that the underlying true model is a mixture of M MNL models. Let θ_m, $m = 1, \ldots, M$, denote the probability that a random customer belongs to segment m (by construction, we must have $\theta_1 + ... + \theta_M = 1$) and let the MNL parameters for segment m be denoted by $u_{im} \geq 0$ for $i \in \mathcal{N}_0 = \mathcal{N} \cup \{0\}$ and $m = 1, ..., M$. Then, for any offer set $S \subset \mathcal{N}$, the true choice probability of product $i \in S_0 := S \cup \{0\}$ is given by

$$\pi(i, S) = \sum_{m=1}^{M} \theta_m \frac{u_{im}}{\sum_{j \in S_0} u_{jm}}.$$

2.1. Multi-Attempt Model

Per our discussions in Section 1, under the k-attempt model, each customer considers up to $k-1$ substitutes, beyond her favorite product, before she decides to not purchase anything. To illustrate, suppose that $\mathcal{N}=\{1,2,3,4\}$ and $S=\{1,2\}$. Under 2-attempt model, a customer will purchase product 1 if either (1) it is her favorite product among all four products and it is preferred to the no-purchase alternative, or (2) she likes either product 3 or 4 best but unfortunately neither of these is included in S and her next favorite product is 1. Let U_i denote the utility of product i and let $\bar{S}$ denote the complement of S. Mathematically, we can write the probability that a customer will purchase product 1 as follows: $\hat{\pi}_2(1,S) = P\left(U_1 > \max\{U_0,U_2,U_3,U_4\}\right) + P\left(U_3 > U_1 > \max\{U_0,U_2,U_4\}\right) + P\left(U_4 > U_1 > \max\{U_0,U_2,U_3\}\right) := \lambda_1 + \lambda_{31} + \lambda_{41}$. Note that this choice probability is the same as the choice probability under the classic exogenous demand model. Similarly, under 3-attempt model, a customer will purchase product 1 if either (1) it is her favorite product among all four products and the no-purchase alternative, or (2) it is her second favorite product after either product 3 or 4, or (3) it is her third favorite product after both products 3 and 4. We can write the probability that a customer will purchase product 1 as follows:

$$\begin{aligned}\hat{\pi}_3(1,S) \;=\; & P\left(U_1 > \max\{U_0,U_2,U_3,U_4\}\right) \\ & + P\left(U_3 > U_1 > \max\{U_0,U_2,U_4\}\right) + P\left(U_4 > U_1 > \max\{U_0,U_2,U_3\}\right) \\ & + P\left(\min\{U_3,U_4\} > U_1 > \max\{U_0,U_2\}\right) \\ := \; & \lambda_1 + \lambda_{31} + \lambda_{41} + \lambda_{\{3,4\}1}.\end{aligned}$$

More generally, given a set of products $\mathcal{N}$ and an offer set S, the probability that a customer will purchase product $i \in S_0$ under k-attempt model is given by

$$\hat{\pi}_k(i,S) = \lambda_i + \sum_{j_1 \in \bar{S}} \lambda_{j_1 i} + \sum_{\{j_1,j_2\} \subseteq \bar{S}} \lambda_{\{j_1,j_2\}i} + \cdots + \sum_{\{j_1,j_2,\ldots,j_{k-1}\} \subseteq \bar{S}} \lambda_{\{j_1,j_2,\ldots,j_{k-1}\}i},$$

where $\lambda_{\{j_1,j_2,\ldots,j_{k-1}\}i}$ is the probability that a customer values product $j \in \{j_1,j_2,\ldots,j_{k-1}\}$ better than i and product $j' \in \mathcal{N} - \{j_1,j_2,\ldots,j_{k-1}\} \cup \{0\}$ worse than i. That is,

$$\lambda_{\{j_1,j_2,\ldots,j_{k-1}\}i} = P\left(\min\{U_{j_1},\ldots,U_{j_{k-1}}\} > U_i \geq \max\{U_l : l \in \mathcal{N} \setminus \{j_1,j_2,\ldots,j_{k-1}\} \cup \{0\}\}\right).$$

Since a customer makes a purchase as soon as her next favorite product is in S, she only needs to consider at most $|\bar{S}|$ substitutes (beyond her most favorite product) before making a purchase. This means that, under multi-attempt model, we must have: $\hat{\pi}_{|\bar{S}|+1}(i,S) = \pi(i,S)$ for all $i \in S_0$. Moreover, by construction, we also have $\hat{\pi}_k(i,S) < \pi(i,S)$ for all $k < |\bar{S}|+1$ and $i \in S_0$.

Error bound for multi-attempt model. We now derive an error bound for k-attempt model. Let $u_{max}(\bar{S})$ be the maximum probability that the most favorite product of a random customer from any segment $m = 1, \ldots, M$ is included in a compliment of offer set S, $\bar{S} := \mathcal{N} \setminus S \cup \{0\}$. That is, $u_{max}(\bar{S}) = \max_m \sum_{i \in \bar{S}} u_{im}$. The following theorem tells us that the relative error of multi-attempt model decreases exponentially with the number of attempts k.

THEOREM 1. *For any $S \subset \mathcal{N}$ and $i \in S_0$, we have:*

$$\left(1 - u_{max}(\bar{S})^k\right) \cdot \pi(i,S) \;\leq\; \hat{\pi}_k(i,S) \;\leq\; \pi(i,S). \tag{1}$$

Proof. Per our discussions above, $\hat{\pi}_k(i,S) = \pi(i,S)$ for $k \geq |\bar{S}|+1$. So, we only need to consider the case $k \leq |\bar{S}|$. We first consider the case where the true choice model is MNL with parameters $\{u_0, u_1, ..., u_n\}$, $\sum_{i=0}^n u_i = 1$. Note that, for any preference sequence $j_1, j_2, \ldots, j_l, i \in \mathcal{N}$, we have:

$$\begin{aligned} &P(U_{j_1} > U_{j_2} > \cdots > U_{j_l} > U_i \geq \max\{U_m : m \in \mathcal{N} \setminus \{j_1, j_2, \ldots, j_l\} \cup \{0\}\}) \\ &\qquad = \left(\frac{u_{j_1}}{1-u_{j_1}}\right)\left(\frac{u_{j_2}}{1-u_{j_1}-u_{j_2}}\right)\cdots\left(\frac{u_{j_l}}{1-u_{j_1}-\cdots-u_{j_l}}\right)\cdot u_i. \end{aligned}$$

The above probability is an immediate consequence of the assumption of i.i.d noises with Gumbel distribution in the construction of MNL model and not difficult to prove (we omit the details). Given the above formula, we can bound $\hat{\pi}_k(i,S)$ as follows:

$$\begin{aligned} \hat{\pi}_k(i,S) &= \lambda_i + \sum_{j_1 \in \bar{S}} \lambda_{j_1 i} + \sum_{\{j_1,j_2\} \subseteq \bar{S}} \lambda_{\{j_1,j_2\}i} + \cdots + \sum_{\{j_1,j_2,\ldots,j_{k-1}\} \subseteq \bar{S}} \lambda_{\{j_1,j_2,\ldots,j_{k-1}\}i} \\ &= \sum_{l=0}^{k-1} \sum_{j_1,\cdots,j_l \in \bar{S}} \left(\frac{u_{j_1}}{1-u_{j_1}}\right)\left(\frac{u_{j_2}}{1-u_{j_1}-u_{j_2}}\right)\cdots\left(\frac{u_{j_l}}{1-u_{j_1}-\cdots-u_{j_l}}\right)\cdot u_i \\ &\geq \sum_{l=0}^{k-1} \sum_{j_1,\cdots,j_l \in \bar{S}} \left(\frac{u_{j_1}}{1-u_{j_1}}\right)\left(\frac{u_{j_2}}{1-u_{j_2}}\right)\cdots\left(\frac{u_{j_l}}{1-u_{j_l}}\right)\cdot u_i \end{aligned}$$

$$
\begin{aligned}
&= \sum_{l=0}^{k-1} l! \cdot \left[\sum_{\substack{\{j_1,\cdots,j_l\}\subseteq \bar{S} \\ j_1<j_2<\cdots<j_l}} \left(\frac{u_{j_1}}{1-u_{j_1}}\right)\left(\frac{u_{j_2}}{1-u_{j_2}}\right)\cdots\left(\frac{u_{j_l}}{1-u_{j_l}}\right)\cdot u_i \right] \\
&= \sum_{l=0}^{k-1} l! \cdot \left[\sum_{\substack{\{j_1,\cdots,j_l\}\subseteq \bar{S} \\ j_1<j_2<\cdots<j_l}} \left(u_{j_1}+u_{j_1}^2+\cdots\right)\left(u_{j_2}+u_{j_2}^2+\cdots\right)\cdots\left(u_{j_l}+u_{j_l}^2+\cdots\right)\cdot u_i \right] \\
&= u_i \cdot \Bigg[0! + 1! \sum_{j_1\in\bar{S}} (u_{j_1}+u_{j_1}^2+\cdots) + 2! \sum_{\substack{\{j_1,j_2\}\subseteq \bar{S} \\ j_1<j_2}} (u_{j_1}+u_{j_1}^2+\cdots)(u_{j_2}+u_{j_2}^2+\cdots) \\
&\qquad + 3! \sum_{\substack{\{j_1,j_2,j_3\}\subseteq \bar{S} \\ j_1<j_2<j_3}} (u_{j_1}+u_{j_1}^2+\cdots)(u_{j_2}+u_{j_2}^2+\cdots)(u_{j_3}+u_{j_3}^2+\cdots) + \cdots \\
&\qquad + (k-1)! \sum_{\substack{\{j_1,\cdots,j_{k-1}\}\subseteq \bar{S} \\ j_1<j_2<\cdots<j_{k-1}}} (u_{j_1}+u_{j_1}^2+\cdots)\cdots(u_{j_{k-1}}+u_{j_{k-1}}^2+\cdots) \Bigg] \\
&\geq u_i \cdot \Bigg[1 + \sum_{j_1\in\bar{S}} u_{j_1} + \left(\sum_{j_1\in\bar{S}} u_{j_1}^2 + 2! \sum_{\substack{\{j_1,j_2\}\subseteq \bar{S} \\ j_1<j_2}} u_{j_1}u_{j_2} \right) \\
&\qquad + \left(\sum_{j_1\in\bar{S}} u_{j_1}^3 + 2! \sum_{\substack{\{j_1,j_2\}\subseteq \bar{S} \\ j_1<j_2}} (u_{j_1}^2u_{j_2}+u_{j_1}u_{j_2}^2) + 3! \sum_{\substack{\{j_1,j_2,j_3\}\subseteq \bar{S} \\ j_1<j_2<j_3}} u_{j_1}u_{j_2}u_{j_3} \right) + \cdots \\
&\qquad + \left(\sum_{j\in\bar{S}} u_{j_1}^{k-1} + 2! \sum_{\{j_1,j_2\}\subseteq\bar{S}} \prod_{\substack{t_1+t_2=k-1 \\ t_a\in\mathbb{N}}} u_{j_1}^{t_1}u_{j_2}^{t_2} + \cdots + (k-1)! \sum_{\substack{\{j_1,\cdots,j_{k-1}\}\subseteq \bar{S} \\ j_1<\cdots<j_{k-1}}} u_{j_1}\cdots u_{j_{k-1}} \right) \Bigg] \\
&= u_i \cdot \left[1 + \left(\sum_{j\in\bar{S}} u_j\right) + \left(\sum_{j\in\bar{S}} u_j\right)^2 + \left(\sum_{j\in\bar{S}} u_j\right)^3 + \cdots + \left(\sum_{j\in\bar{S}} u_j\right)^{k-1} \right],
\end{aligned}
$$

where the fourth equality follows from identity $\frac{x}{1-x} = \sum_{n=1}^{\infty} x^n$ for all $x \in [0,1]$ and the last inequality follows by collecting polynomial terms with the same degree.

Now, if the true choice probability is a mixture of M MNL models with parameters $\{u_{im}\}$ and $\{\theta_m\}$ for all $i \in S_0$ and $m \in \{1,\cdots,M\}$, applying the result above, we can bound $\hat{\pi}_k(i,S)$ as follows:

$$
\begin{aligned}
\hat{\pi}_k(i,S) &\geq \sum_{m=1}^{M} \theta_m u_{im} \cdot \left[1 + \sum_{j\in\bar{S}} u_{jm} + \cdots + \left(\sum_{j\in\bar{S}} u_{jm}\right)^{k-1} \right] \\
&= \sum_{m=1}^{M} \theta_m u_{im} \cdot \frac{1-\left(\sum_{j\in\bar{S}} u_{jm}\right)^k}{1-\left(\sum_{j\in\bar{S}} u_{jm}\right)} \\
&\geq \left(1-\max_m u_m(\bar{S})^k\right) \cdot \sum_{m=1}^{M} \theta_m \cdot \frac{u_{im}}{1-\left(\sum_{j\in\bar{S}} u_{jm}\right)} \\
&= \left(1-u_{max}(\bar{S})^k\right) \cdot \pi(i,S).
\end{aligned}
$$

This completes the proof. ■

Note that multi-attempt model best approximates the true choice model when $u_{max}(\bar{S})$ is small. Intuitively, this is likely to happen when S is large. As mentioned in Section 1, although the theoretical bound of multi-attempt model is encouraging, we will show that its empirical performance is not satisfactory: see numerical results in Table 1. This motivates us to construct a modified multi-attempt model, called re-scaled multi-attempt model which we discuss next.

2.2. Re-scaled Multi-Attempt Model

Under the re-scaled k-attempt model, we approximate $\pi(i,S)$ with $\hat{\pi}_k^R(i,S)$ defined below:

$$\hat{\pi}_k^R(i,S) \;=\; \frac{\hat{\pi}_k(i,S)}{\sum_{j\in S_0}\hat{\pi}_k(j,S)}.$$

Two comments are in order. First, since the re-scaled k-attempt model uses the k-attempt model as its primitive, they share the same set of parameters. In particular, all three models – 2-attempt model, re-scaled 2-attempt model, and Markov chain model – share exactly the same set of parameters. Second, the re-scaled 1-attempt model is identical to MNL approximation. Thus, if the underlying true choice model is MNL (i.e., there is only 1 segment of customer), the re-scaled 1-attempt model is exact.

Analogous to $u_{max}(\bar{S})$, we define $u_{min}(\bar{S})$, the minimum probability that the most favorite product of a random customer from any segment $m = 1,\ldots,M$ is included in $\bar{S}$. That is, $u_{min}(\bar{S}) = \min_m \sum_{i\in\bar{S}} u_{im}$. The following result is an immediate corollary of Theorem 1.

COROLLARY 1. *For any $S \subset \mathcal{N}$ and $i \in S_0$, we have:*

$$\frac{1-u_{max}(\bar{S})^k}{1-u_{min}(\bar{S})}\cdot\pi(i,S) \;\le\; \hat{\pi}_k^R(i,S) \;\le\; \frac{1}{1-u_{max}(\bar{S})}\cdot\pi(i,S). \tag{2}$$

Proof. Let $\hat{\pi}_k^m(i,S)$ denote the choice probability under k-attempt model by a customer that belongs to segment m. We can write:

$$\hat{\pi}_k^R(i,S) \;=\; \frac{\sum_m \theta_m \hat{\pi}_k^m(i,S)}{\sum_m\sum_{j\in S_0}\theta_m\hat{\pi}_k^m(j,S)}.$$

Given the lower bound of $\hat{\pi}_k(i,S)$ in Theorem 1, we can bound:

$$\hat{\pi}_k^R(i,S) \;\geq\; \frac{(1-u_{max}(\bar{S})^k)\cdot\pi(i,S)}{\sum_m\sum_{j\in S_0}\theta_m\hat{\pi}_k^m(j,S)} \;\geq\; \frac{(1-u_{max}(\bar{S})^k)\cdot\pi(i,S)}{\max_m\sum_{j\in S_0}\hat{\pi}_k^m(j,S)} \;=\; \frac{1-u_{max}(\bar{S})^k}{1-u_{min}(\bar{S})}\cdot\pi(i,S)\;.$$

Similarly, we also have:

$$\hat{\pi}_k^R(i,S) \;\leq\; \frac{\pi(i,S)}{\sum_m\sum_{j\in S_0}\theta_m\hat{\pi}_k^m(j,S)} \;\leq\; \frac{\pi(i,S)}{\min_m\sum_{j\in S_0}\hat{\pi}_k^m(j,S)} \;=\; \frac{1}{1-u_{max}(\bar{S})}\cdot\pi(i,S)\;.$$

This completes the proof. ■

While multi-attempt model consistently underestimates the true choice probability, re-scaled multi-attempt model may sometimes overestimate the true probability. Note that the lower bound in Corollary 1 is larger than the lower bound in Theorem 1. This suggests that re-scaled multi-attempt model improves the underestimation error while admitting the overestimation error. The important question is whether this is a good compromise overall. Our numerical results in Table 1 show that re-scaled multi-attempt model significantly improves the empirical accuracy of multi-attempt model. Theoretically, we are also able to show the exactness of re-scaled multi-attempt model when the true choice probability is MNL. This result is reminiscent of the result in Blanchet et al. (2016) that Markov chain model is exact in the case of MNL.

LEMMA 1. *Suppose that the underlying true choice model is an MNL. For any $k>0$, $S\subset\mathcal{N}$ and $i\in S_0$, the re-scaled k-attempt model is exact, i.e., $\hat{\pi}_k^R(i,S)=\pi(i,S)$ for any $k>0$.*

Proof. Let $\alpha_l(\bar{S})=\sum_{j_1,\cdots,j_l\in\bar{S}}\frac{u_{j_1}}{1-u_{j_1}}\frac{u_{j_2}}{1-u_{j_1}-u_{j_2}}\cdots\frac{u_{j_l}}{1-u_{j_1}-\cdots-u_{j_l}}$. Per our note in the proof of Theorem 1, $\alpha_l(\bar{S})\cdot u_i$ is the probability that a customer values product $j\in\{j_1,j_2,\ldots,j_l\}$ better than i and product $j'\in\mathcal{N}\setminus\{j_1,j_2,\ldots,j_l\}\cup\{0\}$ worse than i. Since a customer only purchases product i if her other favorite products (which rank higher than i) are not in the offer set, by definition of random utility model, we must have: $\pi(i,S)=\sum_{l=0}^{|\bar{S}|}\alpha_l(\bar{S})u_i$. As for k-attempt model, since customers only consider up to $k-1$ substitutes, we can write: $\pi_k(i,S)=\sum_{l=0}^{k-1}\alpha_l(\bar{S})u_i$. Putting all things together,

$$\pi(i,S)-\hat{\pi}_k^R(i,S)=\sum_{l=0}^{|\bar{S}|}\alpha_l(\bar{S})u_i-\frac{\sum_{l=0}^{k-1}\alpha_l(\bar{S})u_i}{\sum_{l=0}^{k-1}\alpha_l(\bar{S})u(S_0)}=\frac{u_i}{u(S_0)}-\frac{u_i}{u(S_0)}\;=\;0.$$

This completes the proof. ■

2.3. Numerical Experiments

We conduct numerical experiments with respect to a mixture of M MNLs to compare the performance of multi-attempt, re-scaled multi-attempt, and Markov chain (MC) models. Let n denote the number of products and M denote the number of customer segments in the MMNL model. For a fixed combination of number of products ($n = 10, 20, 50, 100$) and number of segments ($M = 3, 5, 10, 20$), we generate 100 instances. The probability distribution over different MNL segments, $\theta_1, ..., \theta_M$, are first generated using i.i.d samples of the uniform distribution in $[0, 1]$ and then normalized such that $\theta_1 + \cdots + \theta_M = 1$. For each segment $m = 1, ..., M$, the MNL parameters of segment m, $u_{0m}, ...u_{nm}$ are randomly sampled from the uniform distribution in $[0, 1]$. For each instance, we generate a random offer set of size between $n/3$ and $2n/3$, and compute the choice probabilities under the three models. We report both the average and maximum relative errors defined as: avg.Error $= \frac{1}{400}\sum_{a=1}^{400}$ Error(S_a) and max.Error $=$ max$_{1\le a\le 400}$ Error(S_a), where Error$(S) = 100\% \cdot \max_{i\in S} \frac{|\hat{\pi}(i,S)-\pi(i,S)|}{\pi(i,S)}$. The results can be seen in Table 1.

Table 1 Comparison of approximation accuracy of various models

			k-attempt					rescaled k-attempt				
		MC	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
n=10	max.Error	11.69	72.06	48.89	30.42	16.58	7.20	21.95	15.44	9.68	5.10	1.95
	avg.Error	2.45	48.36	21.47	8.58	2.99	0.82	5.64	3.35	1.70	0.71	0.22
n=20	max.Error	8.40	68.55	45.76	29.59	18.43	10.96	17.46	11.26	6.64	4.14	2.77
	avg.Error	2.13	50.82	25.02	11.91	5.47	2.40	4.56	2.91	1.72	0.95	0.49
n=50	max.Error	8.44	63.41	39.76	24.64	15.07	9.10	13.99	10.44	7.41	5.02	3.24
	avg.Error	1.76	50.68	25.58	12.84	6.41	3.18	3.66	2.40	1.51	0.91	0.53
n=100	max.Error	3.17	61.17	37.16	22.41	13.42	7.97	6.03	4.23	2.86	1.87	1.18
	avg.Error	1.35	51.10	26.18	13.44	6.90	3.55	2.30	1.51	0.96	0.59	0.35

A number of observations can be made from Table 1. First, although the accuracy of multi-attempt model improves as k increases, its rate of improvement is not satisfactory. For example, when $n = 100$, the average relative error of 5-attempt model is 3.55%; in contrast, the average relative error of Markov chain model is only about 1.35%. Considering the fact that 5-attempt model requires much more parameters than Markov chain model, this level of performance is not acceptable. Second, re-scaled 2-attempt model significantly improves the accuracy of 2-attempt

model and its relative error is very close to the relative error of Markov chain model. Moreover, re-scaled 3-attempt consistently performs better than Markov chain model. This highlights the benefit of re-scaling.

3. Assortment Optimization

We now discuss how to use re-scaled multi-attempt model in assortment optimization. Since the approximation quality of re-scaled 2-attempt model is very close to Markov chain model, in this work, we will only focus our discussions on re-scaled 2-attempt model. (Our approach for re-scaled 2-attempt model is also generalizable to re-scaled k-attempt model.) We show that assortment optimization under re-scaled 2-attempt model is not much harder than assortment optimization under exogenous (2-attempt) demand model. In particular, it can be formulated as a Mixed Integer Fractional Linear Program (MILFP) and can be solved using the so-called Dinkelbach algorithm.

3.1. Optimization Formulation

Let r_i denote the revenue of product i and $x_i \in \{0,1\}$ be a binary decision variable for product i. We first consider unconstrained assortment optimization problem under exogenous demand model. This can be written as a Mixed Integer Linear Program (MILP) formulation below:

$$\begin{aligned} J_2 &= \max_{\vec{x}\in\{0,1\}^n} \sum_{i=1}^{n} r_i \left[\lambda_i \cdot x_i + \sum_{j\neq i} \lambda_{ji} \cdot x_i \cdot (1-x_j)\right] \\ &= \max_{\substack{\vec{x}\in\{0,1\}^n \\ \vec{y}\in[0,1]^{n(n-1)}}} \sum_{i=1}^{n} r_i \left[\lambda_i \cdot x_i + \sum_{j\neq i} \lambda_{ji} \cdot y_{ji}\right] \\ &\qquad \text{s.t. } y_{ji} \leq x_i,\ y_{ji} \leq 1-x_j,\ y_{ji} \geq x_i - x_j \ \ \forall\, i \neq j \end{aligned}$$

We next consider unconstrained assortment optimization under re-scaled 2-attempt model:

$$\begin{aligned} J_2^R &= \max_{\vec{x}\in\{0,1\}^n} \sum_{i=1}^{n} \frac{r_i \left[\lambda_i \cdot x_i + \sum_{j\neq i} \lambda_{ji} \cdot x_i \cdot (1-x_j)\right]}{\lambda_0 + \sum_{j\neq 0} \lambda_{j0} \cdot (1-x_j) + \sum_{i=1}^{n} \left[\lambda_i \cdot x_i + \sum_{j\neq i} \lambda_{ji} \cdot x_i \cdot (1-x_j)\right]} \\ &= \max_{\substack{\vec{x}\in\{0,1\}^n \\ \vec{y}\in[0,1]^{n(n-1)}}} \sum_{i=1}^{n} \frac{r_i \left[\lambda_i \cdot x_i + \sum_{j\neq i} \lambda_{ji} \cdot y_{ji}\right]}{\lambda_0 + \sum_{j\neq 0} \lambda_{j0} \cdot (1-x_j) + \sum_{i=1}^{n} \left[\lambda_i \cdot x_i + \sum_{j\neq i} \lambda_{ji} \cdot y_{ji}\right]} \\ &\qquad \text{s.t. } \ y_{ji} \leq x_i,\ y_{ji} \leq 1-x_j,\ y_{ji} \geq x_i - x_j \ \ \forall\, i \neq j \end{aligned}$$

Note that J_2^R is an MILFP. As discussed in Section 1, although MILFP is in general difficult to solve, it appears in many applications in engineering and science (Tawarmalani and Sahinidis 2002). Consequently, there is a deep and ever-growing literature on different algorithmic approaches to solve MILFP, either exactly or approximately. When it comes to large-scale MILFP, one popular approach is based on Dinkelbach algorithm, first developed in Dinkelbach (1967). In the context of our assortment problem above, Dinkelbach algorithm works as follows. First, we define $N(\vec{x},\vec{y}) = \sum_{i=1}^{n} r_i \left[\lambda_i \cdot x_i + \sum_{j\neq i} \lambda_{ji} \cdot y_{ji}\right]$ and $D(\vec{x},\vec{y}) = \lambda_0 + \sum_{j\neq 0} \lambda_{j0} \cdot (1-x_j) + \sum_{i=1}^{n} \left[\lambda_i \cdot x_i + \sum_{j\neq i} \lambda_{ji} \cdot y_{ji}\right]$, and let $F(q) = \max\{N(\vec{x},\vec{y}) - q \cdot D(\vec{x},\vec{y}) : (\vec{x},\vec{y}) \in A^*\}$, where A^* is the set of feasible $(\vec{x},\vec{y})$. Now, we proceed in three steps:

Step 1. Choose an arbitrary feasible $(\vec{x}^1,\vec{y}^1)$, set $q_2 = \frac{N(\vec{x}^1,\vec{y}^1)}{D(\vec{x}^1,\vec{y}^1)}$, and let $t = 2$

Step 2. Compute $F(q_t)$ and denote its optimal solution as $(\vec{x}^t,\vec{y}^t)$.

Step 3. If $F(q_t) \leq \epsilon$ (optimality tolerance), stop and output $(\vec{x}^t,\vec{y}^t)$;

Otherwise, let $q_{t+1} = \frac{N(\vec{x}^t,\vec{y}^t)}{D(\vec{x}^t,\vec{y}^t)}$, set $t = t+1$, and go back to Step 2.

Note that computing $F(q_t)$ in Step 2 requires solving an MILP with similar size as J^2. So, the running time of Dinkelbach algorithm approximately equals the running time for solving J^2 multiplies the number of iterations for $F(q_k)$ to be sufficiently close to 0. It has been shown that $F(q_k) \to 0$ at a super-linear rate (You et al. 2009); in fact, when all the variables are binary, in the worst case scenario, Dinkelbach algorithm only requires about log(number of variables) iterations. This highlights the practicality of Dinkelbach algorithm for solving MILFP, especially when the corresponding inner optimization can be quickly solved.

Dealing with constraints. Our optimization model can further accommodate a variety of constraints on the assortment. For example, the following types of constraint from Bertsimas and Mišic (2016) can be easily included: (1) At most U products can be chosen from a subset of size B (maximum subset, also called as cardinality constraints); (2) the number of offered products from a subset of size B cannot be greater than that from the other subset of size B (precedence type 1); (3) a specific product must be offered to include any product from a subset of size $B-1$

(precedence type 2). Any of these constraints can be formulated as a linear constraint, and adding linear constraints still results in an MILFP (under re-scaled 2-attempt model). Thus, we can still use Dinkelbach algorithm.

3.2. Numerical Experiments

To compare the performance of the multi-attempt choice model in assortment optimization, we conduct numerical experiments using the same random instances of the mixture of M MNLs as in Section 2.3. In addition, we also generate a random number between 0 and 1 for the revenue of each product(i.e., r_i for product i). We then compute the optimal assortment under the Markov chain, 2-attempt, and re-scaled 2-attempt models, and calculate the expected revenue of each solution under the true choice model. Table 2 summarizes the average relative gap in expected revenue from the true optimal revenue, including the average running time, for each model. We note that all the computational experiments are carried out on a Mac with Intel Core i5 @ 2.7 GHz and 16-GB RAM. All models and solution procedures are coded in Matlab 2011 and the MILP problems in the proposed algorithm are solved using CPLEX 12 with optimality tolerance of 10^{-5}.

Table 2 **Average relative gap in expected revenue for Markovian model and multi-attempt models with its computing time in second.**

		Markov Chain		2-attempt		rescaled 2-attempt	
		gap(%)	time(s)	gap(%)	time (s)	gap(%)	time (s)
n=10	$M=3$	0.0243	0.0011	9.5785	0.0062	0.0723	0.0193
	$M=5$	0.1077	0.0009	9.5440	0.0061	0.1246	0.0196
	$M=10$	0.0601	0.0009	9.4490	0.0064	0.0998	0.0201
n=20	$M=6$	0.0558	0.0028	14.6167	0.0260	0.0751	0.0859
	$M=10$	0.0217	0.0027	14.5592	0.0252	0.0217	0.0834
	$M=20$	0.0478	0.0029	14.6049	0.0250	0.0492	0.0859
n=50	$M=10$	0.0380	0.0134	21.2468	0.2997	0.0498	1.2327
	$M=20$	0.0380	0.0134	21.2468	0.2997	0.0498	1.2327
	$M=50$	0.0069	0.0169	20.7945	0.2882	0.0080	1.2406
n=100	$M=10$	0.0218	0.0407	24.5002	3.3212	0.0234	14.8783
	$M=20$	0.0179	0.0439	23.8145	3.2690	0.0186	14.6535
	$M=50$	0.0044	0.0555	24.4001	3.2699	0.0045	14.5374

Observe that re-scaling significantly improves the performance of 2-attempt model. Moreover,

the difference between the relative gap of Markov chain model and re-scaled 2-attempt model is negligible. As expected, assortment optimization under Markov chain model can be solved extremely quickly. Although the running time of assortment optimization under re-scaled 2-attempt model is not as short as the running time under Markov chain model, it is nevertheless still quite tractable. Note that the running time under 2-attempt model is only about 3 seconds for $n = 100$. In the case of re-scaled 2-attempt, we use about 5 iterations in the Dinkelbach algorithm, which explains the approximate running time of 15 seconds for $n = 100$. The number of iterations in Dinkelbach algorithm is dictated by the optimality tolerance ϵ (see Step 3). Practically, by adjusting the desired optimality level, one can further reduce the running time under re-scaled 2-attempt model.

Table 3 Average relative gap in expected revenue for constrained (non-scaled and rescaled) 2-attempt models with its computing time in second.

		2-attempt		rescaled 2-attempt	
		gap(%)	time(s)	gap(%)	time(s)
n = 10, M = 5	No constraints	9.5541	0.0068	0.1218	0.0191
	Max.subset, C = 2, B = 5, U = 3	7.8698	0.0077	0.1250	0.0203
	Prec.type 1, C = 1, B = 5	9.6111	0.0076	0.1392	0.0210
	Prec.type 2, C = 2, B = 5	7.1255	0.0073	0.0846	0.0218
n = 20, M =10	No constraints	14.8914	0.0268	0.0440	0.0839
	Max.subset, C = 4, B = 5, U = 3	12.2535	0.0387	0.0420	0.0944
	Prec.type 1, C = 3, B = 5	13.7742	0.0373	0.0534	0.0959
	Prec.type 2, C = 4, B = 5	11.2007	0.0334	0.0668	0.1110
n = 50, M =20	No constraints	21.3102	0.2994	0.0229	1.2814
	Max.subset, C = 5, B = 10, U = 5	14.6982	0.4824	0.0209	1.3596
	Prec.type 1, C = 4, B = 10	20.3282	0.3591	0.0154	1.3101
	Prec.type 2, C = 5, B = 10	17.1013	0.3500	0.0148	1.5295
n =100, M =20	No constraints	24.3847	3.3763	0.0111	15.2983
	Max.subset, C =10, B = 10, U = 5	18.2043	5.5751	0.0119	16.9237
	Prec.type 1, C = 9, B = 10	24.7584	4.4081	0.0125	16.8627
	Prec.type 2, C =10, B = 10	20.0725	3.8482	0.0143	18.3797

Constrained problem. To see the effect of constraints in optimization performance, we solve the optimization instances that we used in Table 2 with a combination of constraints that we discussed in Section 3.1. For each constraint set, we create C constraints by randomly partitioning a set of n products into (mutually exclusive) subsets of size B. The average relative gap and computing time for each constraint are summarized in Table 3. We confirm that the constrained

problems also can be solved quickly (less than 20 seconds when $n = 100$) and its performance is very close to the true optimal performance (less than 0.2% of relative average gap).

4. Concluding Remarks

In this work, we provide a new approach to approximate a general mixed-logit-based choice model. We show that the classic exogenous demand model can be significantly improved by re-scaling. The resulting approximation is exact for MNL and has an empirical performance that is very close to the performance of the recently developed Markov chain model. Moreover, since the proposed approximation model has an explicit mathematical expression, it can be immediately used in an assortment optimization with a variety real-world constraints. Our numerical experiments show that our model is quite tractable for a reasonable sized problem.

References

Bertsimas, Dimitris, Velibor V Mišic. 2016. Data-driven assortment optimization. *submitted* .

Blanchet, Jose H, Guillermo Gallego, Vineet Goyal. 2016. A markov chain approximation to choice modeling. *Operations Research, Forthcoming* .

Désir, Antoine, Vineet Goyal, Danny Segev, Chun Ye. 2015. Capacity constrained assortment optimization under the markov chain based choice model. *Operations Research, Forthcoming* .

Dinkelbach, Werner. 1967. On nonlinear fractional programming. *Management Science* **13**(7) 492–498.

Feldman, Jacob B, Huseyin Topaloglu. 2014. Revenue management under the markov chain choice model .

Hess, Stephane, Andrew Daly. 2014. *Handbook of choice modelling*. Edward Elgar Publishing.

Kochenberger, Gary, Jin-Kao Hao, Fred Glover, Mark Lewis, Zhipeng Lü, Haibo Wang, Yang Wang. 2014. The unconstrained binary quadratic programming problem: a survey. *Journal of Combinatorial Optimization* **28**(1) 58–81.

Kök, A Gürhan, Marshall L Fisher. 2007. Demand estimation and assortment optimization under substitution: Methodology and application. *Operations Research* **55**(6) 1001–1021.

McFadden, Daniel, Kenneth Train, et al. 2000. Mixed mnl models for discrete response. *Journal of applied Econometrics* **15**(5) 447–470.

Rusmevichientong, Paat, Zuo-Jun Max Shen, David B Shmoys. 2010. Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations research* **58**(6) 1666–1680.

Tawarmalani, Mohit, Nikolaos V Sahinidis. 2002. *Convexification and global optimization in continuous and mixed-integer nonlinear programming: theory, algorithms, software, and applications*, vol. 65. Springer Science & Business Media.

You, Fengqi, Pedro M Castro, Ignacio E Grossmann. 2009. Dinkelbach's algorithm as an efficient method to solve a class of minlp models for large-scale cyclic scheduling problems. *Computers & Chemical Engineering* **33**(11) 1879–1889.

Yue, Dajun, Gonzalo Guillén-Gosálbez, Fengqi You. 2013. Global optimization of large-scale mixed-integer linear fractional programming problems: A reformulation-linearization method and process scheduling applications. *AIChE Journal* **59**(11) 4255–4272.