A PTAS for Capacitated Sum-of-Ratios Optimization

Paat Rusmevichientong Zuo-Jun Max Shen David B. Shmoys
Cornell University UC Berkeley Cornell University

Abstract

Motivated by an application in assortment planning under the nested logit choice model, we develop a polynomial-time approximation scheme for the sum-of-ratios optimization problem with a capacity constraint and a fixed number of product groups.

Keywords: Sum-of-Ratios, Polynomial-time Approximation Scheme, and Assortment Planning.

1. Motivation and Introduction

Assortment planning is an important problem facing many retailers and has been studied extensively in the supply chain and operation management literature. Given a limited shelf capacity or inventory investment constraint, the retailer must determine the subset of products to offer that maximizes the total profit. The literature on assortment planning is broad and covers a range of operational issues. A comprehensive review of the literature in this area is given in [9]. The stream of research most closely related to our work is on demand modeling and optimization algorithms. Earlier work in assortment planning assumes that the demand of each product is independent. Recent work focuses on more complex models of customer choice behavior, allowing for more intricate substitution patterns among products, and providing more realistic models of demand. One of the most commonly used choice models in economics, marketing, and operations management is the multinomial logit (MNL) model (see, for example, [1, 2, 11, 13], and the references therein). Pioneering work in assortment optimization with complex choice models include [4, 12, 18].

Although the MNL model is analytically tractable, it exhibits the Independence of Irrelevant Alternative (IIA) property. Under this property, the likelihood of choosing between any two alternatives is independent of the assortment containing them. As discussed in [2, 13, 14] and many other work, this property is inconsistent with the actual customer choice behavior in many settings. Despite this shortcoming, the MNL model remains one of the most useful and popular choice models partly because the corresponding assortment optimization problem admits an efficient polynomial-time algorithm. An algorithm for the unconstrained optimization problem is given in [7, 10], which is later extended to the capacitated setting in [16].

The nested logit choice model [14] is one of the most popular extensions of the MNL model that alleviates the IIA property. Under this model, the products are partitioned into groups. Although products within...
the same group still exhibit the IIA property, the likelihood of choosing products from two different groups depends on the assortment containing them, providing a potentially more realistic model of customer choice behavior. Unfortunately, the assortment optimization problem under the nested logit choice model appears to be intractable. In this paper, we establish NP-completeness and demonstrate how the problem can be formulated as an integer programming problem involving a sum of ratios and a capacity constraint. As indicated in [17], there is very few work on approximation algorithms for this type of integer programming problems. To our knowledge, this paper provides the first polynomial-time approximation scheme (PTAS) for this class of problems.

2. Problem Formulation

We are given a set of $N$ products indexed by $1, \ldots, N$. For each product $\ell$, let $\pi_\ell$ and $c_\ell$ denote its marginal profit (per-unit revenue minus marginal cost) and fixed cost, respectively. The fixed cost of each product might correspond to the cost of introducing the product into the store. We denote the option of no purchase by 0 and set $\pi_0 = c_0 = 0$. Given an inventory investment or capacity constraint $C$, we wish to find the subset of products that gives the maximum expected profit.

We will model the customer demand using the nested logit choice model. Under this model, the set of products is partitioned into $G$ disjoint groups denoted by $H_1, H_2, \ldots, H_G$ where $\cup_{g=1}^{G} H_g = \{1, \ldots, N\}$ and $H_i \cap H_j = \emptyset$ for all $i \neq j$. Let $H_0 = \{0\}$ denote the set consisting of the option of no purchase. Each customer assigns a mean utility $\eta_\ell \in \mathbb{R}$ to product $\ell$, and as a convention, we set $\eta_0 = 0$. For each assortment $S \subseteq \{1, 2, \ldots, N\}$, given that a set of products $S$ is offered to the customer, the probability $P_\ell(S)$ that the customer will choose product $\ell$ is given by:

$$P_\ell(S) = \begin{cases} 
0 & \text{if } \ell \notin \{0\} \cup S \\
\frac{e^{\eta_\ell/\tau_g}}{\left(\sum_{j \in H_\ell \cap S} e^{\eta_j/\tau_g}\right) \tau_g} \times \frac{\left(\sum_{j \in H_\ell \cap S} e^{\eta_j/\tau_g}\right)^{\tau_g}}{1 + \sum_{k=1}^{G} \left(\sum_{j \in H_k \cap S} e^{\eta_j/\tau_k}\right)^{\tau_k}} & \text{if } \ell \in (\{0\} \cup S) \cap H_g \text{ for some } g
\end{cases}$$

where for each $g = 1, \ldots, G$, the parameter $0 \leq \tau_g \leq 1$ denotes the “dissimilarity factor” associated with the group $H_g$. The requirement that $0 \leq \tau_g \leq 1$ for all $g$ ensures that the nested logit choice model is consistent with utility maximization [3, 5, 14]. Note that since $H_0 = \{0\}$, the parameter $\tau_0$ is not needed and $P_0(S) = 1 / \left(1 + \sum_{g=1}^{G} \left(\sum_{j \in H_g \cap S} e^{\eta_j/\tau_g}\right)^{\tau_g}\right)$. Also, the classical MNL model is a special case of the nested logit model with $H_g = \{g\}$ for all $g$.

Our goal is to choose a subset of products that maximizes the total profit. Our optimization problem, which we refer to as the ASSORTMENT problem, can be stated as follows:

\[
\text{(ASSORTMENT)} \quad Y^* = \max \left\{ \sum_{\ell \in S} \pi_\ell P_\ell(S) \ \middle| \ S \subseteq \{1, 2, \ldots, N\} \text{ and } \sum_{\ell \in S} c_\ell \leq C \right\}
\]

Note that when $H_g = \{g\}$ for all $g$, for any assortment $S$, $\sum_{\ell \in S} \pi_\ell P_\ell(S) = \frac{\sum_{g=1}^{G} \pi_\ell \eta_\ell^{g}}{\sum_{g=1}^{G} \eta_\ell^{g}}$. In this case, the ASSORTMENT problem reduces to the knapsack problem whose objective function is a ratio, which is known to be NP-complete [8]. Thus, the following complexity result follows immediately.

**Lemma 2.1.** ASSORTMENT is NP-complete.
3. Connection to Sum-of-Ratios Optimization

We will now describe how an approximation algorithm for a sum-of-ratios optimization problem can be used to obtain an approximation for the ASSORTMENT problem. For each $g = 1, \ldots, G$, let $a_g = 1 - \tau_g$ and for each $\ell \in H_g$, let $v_\ell = e^{\nu_g / \tau_g}$. Define a function $J^* : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ by: for each $\lambda \in \mathbb{R}_+$,

$$J^*(\lambda) = \max \left\{ \sum_{g=1}^{G} \sum_{\ell \in S_g} \frac{(\pi_\ell - \lambda) v_\ell}{(\sum_{\ell \in S_g} v_\ell)^{\alpha_g}} \left| S_g \subseteq H_g \text{ and } \sum_{g=1}^{G} \sum_{\ell \in S_g} c_\ell \leq C \right. \right\}.$$

Using a technique pioneered by Megiddo [15], we give an alternative characterization of the optimal profit $Y^*$ associated with the ASSORTMENT problem.

Lemma 3.1 (Parametric Representation). $J^*(\cdot)$ is a continuous, convex, non-negative, and non-increasing function such that $Y^* = \max \{ \lambda : J^*(\lambda) \geq \lambda \}$.

Proof. For each $g = 1, \ldots, G$, let the functions $M_g : 2^{\{1,2,\ldots,N\}} \rightarrow \mathbb{R}_+$ and $W_g : 2^{\{1,2,\ldots,N\}} \rightarrow \mathbb{R}_+$ be defined by: if $S \cap H_g \neq \emptyset$,

$$M_g(S) = \left( \sum_{\ell \in H_g \cap S} e^{\nu_g / \tau_g} \right)^{\tau_g} \text{ and } W_g(S) = \sum_{\ell \in H_g \cap S} \left( \frac{\pi_\ell}{\sum_{j \in H_g \cap S} e^{\nu_j / \tau_g}} \right) e^{\nu_g / \tau_g},$$

and we define $W_g(S) = M_g(S) = 0$ if $S \cap H_g = \emptyset$. It follows from the definition of $P_1(S)$ that

$$\sum_{\ell \in S} \pi_\ell P_\ell(S) = \sum_{g=1}^{G} \sum_{\ell \in H_g \cap S} \pi_\ell P_\ell(S) = \sum_{g=1}^{G} \left\{ \frac{\left( \sum_{f \in H_g \cap S} e^{\nu_f / \tau_g} \right)^{\tau_g}}{1 + \sum_{k=1}^{G} \left( \sum_{f \in H_k \cap S} e^{\nu_f / \tau_k} \right)^{\tau_k}} \times \sum_{\ell \in H_g \cap S} \left( \frac{\pi_\ell}{\sum_{j \in H_g \cap S} e^{\nu_j / \tau_g}} \right) \right\} = \sum_{g=1}^{G} W_g(S) M_g(S) \left( 1 + \sum_{g=1}^{G} M_g(S) \right).$$

This result enables us to express the optimal profit $Y^*$ as follows:

$$Y^* = \max \left\{ \lambda : \exists S \subseteq \{1, \ldots, N\} \text{ with } \sum_{\ell \in S} c_\ell \leq C \text{ such that } \sum_{g=1}^{G} W_g(S) M_g(S) \frac{1}{1 + \sum_{g=1}^{G} M_g(S)} \geq \lambda \right\}$$

$$= \max \left\{ \lambda : \exists S \subseteq \{1, \ldots, N\} \text{ with } \sum_{\ell \in S} c_\ell \leq C \text{ such that } \sum_{g=1}^{G} M_g(S) (W_g(S) - \lambda) \geq \lambda \right\}$$

$$= \max \left\{ \lambda : \exists S_1 \subseteq H_1, \ldots, S_G \subseteq H_G \text{ with } \sum_{g=1}^{G} \sum_{\ell \in S_g} c_\ell \leq C \text{ and } \sum_{g=1}^{G} M_g(S_g) (W_g(S_g) - \lambda) \geq \lambda \right\},$$

where the final equality follows from the fact that for each $g$,

$$\sum_{g=1}^{G} M_g(S_g) (W_g(S_g) - \lambda) = \sum_{g=1}^{G} \sum_{\ell \in S_g} \frac{(\pi_\ell - \lambda) v_\ell}{e^{\nu_g / \tau_g}} = \sum_{g=1}^{G} \sum_{\ell \in S_g} \left( \frac{\pi_\ell - \lambda}{\sum_{\ell \in S_g} v_\ell} \right)^{\alpha_g} \left| S_g \subseteq H_g \text{ and } \sum_{g=1}^{G} \sum_{\ell \in S_g} c_\ell \leq C \right. = J^*(\lambda).$$
Lemma 3.2 (Approximation Preservation). Suppose there exists $0 < \epsilon < 1$ and a non-increasing function $\hat{J}: \mathbb{R}_+ \rightarrow \mathbb{R}_+$ such that for each $\lambda \in \mathbb{R}_+$, $(1 - \epsilon)J^*(\lambda) \leq \hat{J}(\lambda) \leq J^*(\lambda)$. If $\hat{Y} = \sup \{\lambda : \hat{J}(\lambda) \geq \lambda\}$, then $(1 - \epsilon)Y^* \leq \hat{Y} \leq Y^*$. 

Proof. By definition of $\hat{Y}$ and $\hat{J}(\cdot)$, we have that for all $\delta > 0$, $(\hat{Y} - \delta) \leq \hat{J}(\hat{Y} - \delta) \leq J^*(\hat{Y} - \delta)$. It follows from Lemma 3.1 that $\hat{Y} - \delta \leq Y^*$ for all $\delta > 0$, and thus, $\hat{Y} \leq Y^*$. We will now show that $\hat{Y} \geq (1 - \epsilon)Y^*$. The result is trivially true if $\hat{Y} = Y^*$. So, we will focus on the case where $\hat{Y} < Y^*$. Using the definition of $\hat{Y}$ and the fact that $\hat{J}$ is non-increasing, for all $0 \leq \delta \leq Y^* - \hat{Y}$, $\hat{J}(Y^*) \leq \hat{J}(\hat{Y} + \delta) \leq (\hat{Y} + \delta)$. Since this is true for all $0 < \delta \leq Y^* - \hat{Y}$, it must be the case that $\hat{J}(Y^*) \leq \hat{Y}$. Using the fact that $Y^* = J^*(Y^*)$ by Lemma 3.1, we have

$$(Y^* - \hat{Y}) = J^*(Y^*) - \hat{Y} = J^*(Y^*) - \hat{J}(Y^*) + \hat{J}(Y^*) - \hat{Y} \leq \epsilon J^*(Y^*) = \epsilon Y^*,$$

and thus, $(1 - \epsilon)Y^* \leq \hat{Y}$, which is the desired result. 

Lemmas 3.1 and 3.2 suggest the following approach for identifying the optimal profit $Y^*$ in the Assortment problem. Since we cannot compute the function $J^*(\cdot)$ explicitly, we will perform a bisection search on the function $\hat{J}(\cdot)$, which is our $\epsilon$-approximation to the true function $J^*(\cdot)$. Note that we do not have to maintain the values of $\hat{J}(\lambda)$ for all $\lambda$. Instead, we only need to compute $\hat{J}(\lambda)$ for the specific choice of $\lambda$ that is determined during the bisection search algorithm. Let $\hat{Y} = \sup \{\lambda : \hat{J}(\lambda) \geq \lambda\}$. After $O(\log(1/\epsilon))$ iterations, the bisection search will terminate at a point $\hat{Y}$ such that $|\hat{Y} - Y^*| \leq \epsilon Y^*$. It follows from Lemma 3.2 that $|\hat{Y} - Y^*| \leq |\hat{Y} - \hat{Y}| + |\hat{Y} - Y^*| \leq 2\epsilon Y^*$, giving us the desired approximation.

The above result shows that it suffices to consider the following class of combinatorial optimization problems involving a sum of ratios with a capacity constraint. In our setting, we have a ground set indexed by $\{1, \ldots, N\}$ and a partition $H_1, \ldots, H_G$ where $\cup_{\ell=1}^G H_\ell = \{1, \ldots, N\}$ and $H_\ell \cap H_j = \emptyset$ for all $\ell \neq j$. We are interested in the following combinatorial optimization problem, which we will refer to as the Sum Of Ratios (SOR) problem:

\[
(SOR) \quad Z^* = \max \left\{ f(S_1, \ldots, S_G) := \sum_{g=1}^G \frac{\sum_{\ell \in S_g} u_\ell}{\alpha_g} \ \mid \ S_g \subseteq H_g \ \forall g \text{ and } \sum_{g=1}^G \sum_{\ell \in S_g} c_\ell \leq C \right\}
\]
\( \ell \in H_g \), the above SOR problem is the same as the optimization problem associated with \( J^*(\lambda) \) that is derived from our assortment optimization problem.

4. A PTAS for SOR when \( G \) is Fixed

In this section, we describe a PTAS for the SOR problem when the number of groups \( G \) is fixed. Let \((A_1^g, A_2^g, \ldots, A_C^g)\) denote the optimal solution associated with \( Z^* \) with \( A_g^* \subseteq H_g \) for each \( g \in \{1, 2, \ldots, G\} \) and \( \sum_{g=1}^G \sum_{\ell \in A_g^*} c_\ell \leq C \). The main result of this paper is stated in the following theorem.

**Theorem 4.1** (PTAS). For each \( \delta > 0 \) and \( k \in \mathbb{Z}_+ \) with \( k \geq 2 \), there exists an approximation algorithm \( \text{APPROX}(k, \delta) \) for the SOR problem that will generate an assortment \((A_1(k, \delta), A_2(k, \delta), \ldots, A_C(k, \delta))\) such that \( A_g(k, \delta) \subseteq H_g \) for each \( g \in \{1, 2, \ldots, G\} \), \( \sum_{g=1}^G \sum_{\ell \in A_g(k, \delta)} c_\ell \leq C \), and

\[
Z^* - f(A_1(k, \delta), A_2(k, \delta), \ldots, A_C(k, \delta)) \leq \frac{k}{k - 2} \left( \delta + \frac{2}{k} \right) Z^*. 
\]

Moreover, the \( \text{APPROX}(k, \delta) \) algorithm has a running time of

\[
O \left( G \times \text{LP} \times \left\{ N^{k+1} \times \log(N \cdot v^*) \times \log C / \log^2(1 + \delta) \right\}^G \right),
\]

where \( v^* = \max \ell v_\ell \) and \( \text{LP} \) denotes the running time for solving a linear program with \( N \) variables and \( 2 \) constraints whose coefficients and upper bound constraints are bounded above by \( \max \{ \max \ell c_\ell, \max \ell u_\ell, N v^*, C \} \).

For the rest of this section, we establish a series of auxiliary results (Lemmas 4.2 - 4.5) that are needed for the proof of Theorem 4.1, which appears in Section 4.1. Let us introduce the following notation. For each \( g \), let \( \lambda \) denote the optimal solution associated with \( Z^* \) and let \( \nu > 0 \) be defined by

\[
LP, g = \max \left\{ \sum_{\ell \in H_g} u_\ell x_\ell \middle| \sum_{\ell \in H_g} c_\ell x_\ell \leq C, \sum_{\ell \in H_g} v_\ell x_\ell \leq r_g \forall g, \text{ and } x_\ell \in \{0, 1\} \forall \ell \right\}.
\]

As the first step toward the development of the PTAS, the following lemma provides a characterization of the optimal solution \( Z^* \). The proof follows from the definition of \( Z^{IP} \) and \( Z^* \), and we omit the details.

**Lemma 4.2** (IP Representation). \( Z^* = \max \left\{ Z^{IP}(r_1, r_2, \ldots, r_G) \middle| (r_1, r_2, \ldots, r_G) \in \mathbb{R}_+^G \right\} \).

Motivated by Lemma 4.2, we will consider the following collection of linear programs, which will provide an approximation to \( Z^{IP}(r_1, r_2, \ldots, r_G) \). For each \( g \in \{1, 2, \ldots, G\} \), \( (r_g, s_g) \in \mathbb{R}_+^2 \), and \( B_g \subseteq H_g \), let \( Z^{LP,g}(r_g, s_g, B_g) \) be defined by

\[
Z^{LP,g}(r_g, s_g, B_g) = \max \left\{ \sum_{\ell \in H_g} u_\ell x_\ell \middle| \sum_{\ell \in H_g} v_\ell x_\ell \leq r_g, \sum_{\ell \in H_g} c_\ell x_\ell \leq s_g, \right. \\
\left. \forall \ell, 0 \leq x_\ell \leq 1 \text{ and } x_\ell = \begin{cases} 
1, & \text{if } \ell \in B_g, \\
0, & \text{if } \ell \notin B_g \text{ and } u_\ell > \min_{s \in B_g} u_s
\end{cases} \right\}.
\]

and let \( x^{LP,g}(r_g, s_g, B_g) \in [0, 1]^{H_g} \) denote the optimal solution to the optimization problems associated with \( Z^{LP,g}(r_g, s_g, B_g) \). As a convention, we set the objective value to \(-\infty\) if the problem is infeasible. Our
choice of the linear program is inspired by Frieze and Clarke [6] who develop a PTAS for a multi-dimensional knapsack problem. The next lemma establishes an integrality gap of the linear program.

Lemma 4.3 (Integrality Gap). For each \( g \in \{1, 2, \ldots, G\} \) with \( B_g \subseteq H_g \) and for each \((r_g, s_g) \in \mathbb{R}_+^2\),

\[
\sum_{\ell \in H_g} u_\ell \left\{ x_{\ell}^{LP,g} (r_g, s_g, B_g) - \left[ x_{\ell}^{LP,g} (r_g, s_g, B_g) \right] \right\} \leq \frac{2 Z_{LP,g} (r_g, s_g, B_g)}{|B_g|} .
\]

Proof. By our construction, if \( x_{\ell}^{LP,g} (r_g, s_g, B_g) \) is strictly between 0 and 1, then it must be the case that \( u_\ell \leq \min_{s \in B_g} u_s \). Moreover, it follows from the definition of the linear program associated with \( Z_{LP,g} (r_g, s_g, B_g) \) that their extreme points can have at most two fractional coordinates. Therefore,

\[
\sum_{\ell \in H_g} u_\ell \left\{ x_{\ell}^{LP,g} (r_g, s_g, B_g) - \left[ x_{\ell}^{LP,g} (r_g, s_g, B_g) \right] \right\} \leq 2 \min\{ u_\ell : \ell \in B_g \} \leq \frac{2 Z_{LP,g} (r_g, s_g, B_g)}{|B_g|} .
\]

When the set \( B_g \) is large, the above lemma shows that the contribution to the objective value \( Z_{LP,g} (r_g, s_g, B_g) \) from the fractional coordinates is minimal. This result suggests that for the group \( g \) where the optimal solution \( A_g^* \subseteq H_g \) is large, the rounded solution of the linear programming relaxation should serve as a good approximation to the true optimal solution. Later in this section, we will use this insight in designing our approximation algorithm. The next key ingredient for developing our PTAS is a bound on the sensitivity of the linear program as the parameters \( r_g \) and \( s_g \) vary.

Lemma 4.4 (LP Sensitivity). For each \( g \in \{1, \ldots, G\} \) and \( B_g \subseteq H_g \) and for every \( 0 \leq r_g \leq r'_g \) and \( 0 \leq s_g \leq s'_g \), \( 1 \leq \frac{Z_{LP,g} (r'_g, s'_g, B_g)}{Z_{LP,g} (r_g, s_g, B_g)} \leq \max \left\{ \frac{s'_g}{s_g}, \frac{r'_g}{r_g} \right\} .
\]

Proof. The feasible region of the dual of the linear program associated with \( Z_{LP,g} (r_g, s_g, B_g) \) is independent of \( r_g \) and \( s_g \). It follows from the strong duality theorem that there exist \( E_g (B_g) \subseteq \mathbb{R}_+^2 \) such that for each \((r_g, s_g) \in \mathbb{R}_+^2\), \( Z_{LP,g} (r_g, s_g, B_g) = \min_{(\alpha, \beta, \gamma) \in E_g (B_g) \{ \alpha r_g + \beta s_g + \gamma \}} \), and the desired result follows.

Lemma 4.4 suggests that, instead of considering all values of \( r_g \in \mathbb{R}_+ \), we can restrict our attention to a “discretized domain” with negligible loss in performance. For each \( \delta > 0 \), let \( \text{Dom}(\delta) \) be defined by:

\[
\text{Dom}(\delta) = \{0\} \cup \{ (1 + \delta)^\ell : \ell \in \{0\} \cup \mathbb{Z}_+ \} .
\]

Motivated by the results of Lemmas 4.2, 4.3, and 4.4, we will consider the following approximation to \( Z^* \). For each subset \( \mathcal{L} \subseteq \{1, 2, \ldots, G\} \) and for each assortment \( \mathcal{A} = (A_1, \ldots, A_G) \) with \( A_g \subseteq H_g \) for all \( g \), let \( V (\mathcal{L}, \mathcal{A}, \delta) \) be defined by

\[
V (\mathcal{L}, \mathcal{A}, \delta) = \max \sum_{g \in \mathcal{L}} Z_{LP,g} (r_g, s_g, A_g) \frac{r_g^{\delta_g}}{r_g^\alpha} + \sum_{g \in \{1, \ldots, G\} \setminus \mathcal{L}} \sum_{\ell \in A_g} u_\ell \frac{r_g^{\delta_g}}{r_g^\alpha} \left( \sum_{\ell \in A_g} v_\ell \right)^{\delta_g} \quad (4)
\]

\[
\text{s.t. } \{r_g, s_g\} \subseteq \text{Dom}(\delta) \ \forall g \in \mathcal{L}, \quad \text{and } \sum_{g \in \mathcal{L}} s_g + \sum_{g \in \{1, \ldots, G\} \setminus \mathcal{L}} \sum_{\ell \in A_g} c_\ell \leq C
\]

We note that the only variables in the optimization problem associated with \( V (\mathcal{L}, \mathcal{A}, \delta) \) correspond to \((r_g, s_g : g \in \mathcal{L})\). For \( g \notin \mathcal{L} \), the contribution to \( V (\mathcal{L}, \mathcal{A}, \delta) \) is determined by \( A_g \). As a convention, we set the value to \(-\infty\) if the problem is infeasible.
For each \( \mathcal{L} \subseteq \{1, 2, \ldots, G\} \) and \( \mathcal{A} = (A_1, \ldots, A_G) \), we can compute \( V(\mathcal{L}, \mathcal{A}, \delta) \) in polynomial time because it suffices to restrict \( r_g \leq N \cdot \max_{\ell} v_{\ell} \) and \( s_g \leq C \). Thus, there are only a total of \( O\left( \frac{\log(N \cdot \max_{\ell} v_{\ell})}{\log(1 + \delta)} \times \frac{\log C}{\log(1 + \delta)} \right) \) combinations of \((r_g, s_g) \in \text{Dom}(\delta) \times \text{Dom}(\delta)\) to consider. Also, for each combination, we need to solve a linear program associated with \( Z^{LP,g}(r_g, s_g, A_g) \), which can be done in polynomial time. The following lemma provides upper and lower bounds associated with the optimal value \( Z^* \) in terms of \( V(\cdot, \cdot, \cdot) \). Recall that \((A_1^*, A_2^*, \ldots, A_G^*)\) denotes the optimal solution to the SOR problem given in Equation (1).

**Lemma 4.5 (Bounds on \( Z^* \)).** For each \( \mathcal{L} \subseteq \{1, 2, \ldots, G\} \) and \( \mathcal{A} = (A_1, \ldots, A_g) \) such that \( A_g \subseteq A_g^* \) for all \( g \), if \( \min_{l \in A_g} u_{\ell} \geq \max_{l \in A_g^* \setminus A_g} u_{\ell} \) for each \( g \in \mathcal{L} \) and \( A_g = A_g^* \) for each \( g \notin \mathcal{L} \), then for all \( \delta > 0 \),

\[
\left(1 - \frac{2}{\min(|A_g| : g \in \mathcal{L})}\right) V(\mathcal{L}, \mathcal{A}, \delta) \leq Z^* \leq (1 + \delta) V(\mathcal{L}, \mathcal{A}, \delta).
\]

The proof of Lemma 4.5 is straightforward, but somewhat lengthy. To facilitate our exposition, the proof is deferred until Section 4.2. The results of Lemma 4.5, however, offer two key insights. First, the loss in performance from restricting our attention to a discretized domain \( \text{Dom}(\delta) \) is negligible. Second, we can divide the optimal solution \((A_1^*, A_2^*, \ldots, A_G^*)\) into two categories. The first category corresponds to those groups whose \( A_g^* \) has large cardinality (denote by the set \( \mathcal{L} \)), while the second category corresponds to groups whose corresponding optimal solution has smaller cardinality. From the above lemma, we know that for groups \( g \in \mathcal{L} \), our approximation based on linear programs should be close to the true optimal value. On the other hand, for the group \( g \) whose \( A_g^* \) has a small cardinality, we should be able to use enumeration to identify the optimal solution. This is exactly the key intuition underlying our proposed approximation scheme, which we refer to as \text{APPROX}(k, \delta)\). In this case, the parameter \( k \) controls the cardinality of subsets that we will explore. And the parameter \( \delta \) calibrates the granularity of the domain \( \text{Dom}(\delta) \). The description of the algorithm is given below.

**APPROX\((k, \delta)\):**

**Input:** Parameters \( k \in \mathbb{Z}_+ \) and \( \delta > 0 \)

**Description:** Solve the following optimization problem by enumeration:

\[
\max V(\mathcal{L}, \mathcal{A}, \delta)
\]

s.t. \( \mathcal{L} \subseteq \{1, 2, \ldots, G\} \) and \( \mathcal{A} = (A_1, A_2, \ldots, A_G) \) where \( A_g \subseteq H_g \) for all \( g \), and \( |A_g| \leq k \) for \( g \notin \mathcal{L} \) and \( |A_g| = k \) for all \( g \in \mathcal{L} \)

where \( V(\mathcal{L}, \mathcal{A}) \) is defined in Equation (4). Let \( \bar{\mathcal{L}} \subseteq \{1, \ldots, G\} \) and \( \bar{\mathcal{A}} = (\bar{A}_1, \ldots, \bar{A}_g) \) denote the optimal solution to the above optimization problem. Note that by our construction, \( |ar{A}_g| = k \) for all \( g \in \bar{\mathcal{L}} \). Let \( r(\bar{\mathcal{L}}, \bar{\mathcal{A}}) = (r_g(\bar{\mathcal{L}}, \bar{\mathcal{A}}) : g \in \bar{\mathcal{L}}) \) and \( s(\bar{\mathcal{L}}, \bar{\mathcal{A}}) = (s_g(\bar{\mathcal{L}}, \bar{\mathcal{A}}) : g \in \bar{\mathcal{L}}) \) denote the optimal solution associated with \( V(\bar{\mathcal{L}}, \bar{\mathcal{A}}, \delta) \). Also, for each \( g \in \bar{\mathcal{L}} \), let

\[
x^{LP,g}(\bar{\mathcal{L}}, \bar{\mathcal{A}}) := x^{LP,g}(r_g(\bar{\mathcal{L}}, \bar{\mathcal{A}}), s_g(\bar{\mathcal{L}}, \bar{\mathcal{A}}), \bar{A}_g) \in [0, 1]^{|H_g|}
\]

denote the optimal solution of the linear program associated with \( Z^{LP,g}(r_g(\bar{\mathcal{L}}, \bar{\mathcal{A}}), s_g(\bar{\mathcal{L}}, \bar{\mathcal{A}}), \bar{A}_g) \).
Output: The output of the APPOX\((k, \delta)\) algorithm is the assortment \(A(k, \delta) = (A_1(k, \delta), \ldots, A_G(k, \delta))\) defined by: for each \(g \in \{1, \ldots, G\}, \)
\[
A_g(k, \delta) = \begin{cases} 
\bar{A}_g, & \text{if } g \notin \bar{L}, \\
\ell \in H_g : x_{\ell}^{LP, g}(\bar{L}, \bar{A}) = 1, & \text{if } g \in \bar{L}, 
\end{cases}
\]
and let \(\hat{Z}(k, \delta) = f(A(k, \delta))\) denote the corresponding profit.

4.1 Proof of the Main Result (Theorem 4.1)

We will first show that the output \(A(k, \delta)\) satisfies the capacity constraint. By definition,
\[
\sum_{g=1}^{G} \sum_{\ell \in A_g(k, \delta)} c_\ell \leq \sum_{g \notin \bar{L}} \sum_{\ell \in H_g} c_\ell x_{\ell}^{LP, g}(\bar{L}, \bar{A}) + \sum_{g \notin \bar{L}} \sum_{\ell \in \bar{A}_g} c_\ell \leq \sum_{g \notin \bar{L}} \sum_{\ell \in \bar{A}_g} \bar{s}_g(\bar{L}, \bar{A}) + \sum_{g \notin \bar{L}} \sum_{\ell \in \bar{A}_g} \bar{c}_\ell \leq C,
\]
where the second inequality follows from the fact that \(x^{LP, g}(\bar{L}, \bar{A})\) is an optimal solution of the linear program associated with \(Z^{LP, g}(r_g(\bar{L}, \bar{A}), s_g(\bar{L}, \bar{A}), \bar{A}_g)\). The final inequality follows from the definition of \(V(\bar{L}, \bar{A}, \delta)\). Thus, \(A(k, \delta)\) is a valid assortment. We will establish the approximation guarantee by showing that

\[
Z^* - \hat{Z}(k, \delta) \leq \frac{k}{k - 2} \left(\delta + \frac{2}{k}\right) Z^*.
\]

Recall that \(A_1^*, \ldots, A_G^*\) denote the optimal solution associated with \(Z^*\). Let \(\text{Small} = \{g : |A_g^*| \leq k\}\) and \(\text{Large} = \{g : |A_g^*| > k\}\). For each \(g \in \text{Large}\), let \(T_g^* \subseteq A_g^*\) consists of the \(k\) elements in \(A_g^*\) with the highest values of \(u_\ell\)’s. Note that for each \(g \in \text{Large}\), \(\min_{\ell \in T_g^*} u_\ell \geq \max_{\ell \in A_g^* \setminus T_g^*} u_\ell\). Also, let \(T = (T_1, \ldots, T_G)\) be defined by: for each \(g \in \{1, \ldots, G\}, \)
\[
T_g = \begin{cases} 
T_g^*, & \text{if } g \in \text{Large}, \\
A_g^*, & \text{if } g \in \text{Small}.
\end{cases}
\]

Note that \(T_g \subseteq A_g^*\) for all \(g\) and \(T\) satisfies the hypothesis of Lemma 4.5. Therefore,
\[
Z^* - \hat{Z}(k, \delta) = Z^* - f(A(k, \delta)) \geq (1 + \delta) V(\text{Large}, T, \delta) - f(A(k, \delta))
\]
We will now establish a lower bound on \(f(A(k, \delta))\). By definition of \(A(k, \delta)\) and the fact that \(x^{LP, g}(\bar{L}, \bar{A})\) is an optimal solution of the linear program associated with \(Z^{LP, g}(r_g(\bar{L}, \bar{A}), s_g(\bar{L}, \bar{A}), \bar{A}_g)\), we have
\[
f(A(k, \delta)) = \sum_{g \notin \bar{L}} \sum_{\ell \in H_g} u_\ell \frac{x_{\ell}^{LP, g}(\bar{L}, \bar{A})}{(r_g(\bar{L}, \bar{A}))^{\alpha_g}} + \sum_{g \notin \bar{L}} \sum_{\ell \in \bar{A}_g} u_\ell \frac{x_{\ell}^{LP, g}(\bar{L}, \bar{A})}{(r_g(\bar{L}, \bar{A}))^{\alpha_g}}
\geq \sum_{g \notin \bar{L}} \sum_{\ell \in H_g} u_\ell \frac{x_{\ell}^{LP, g}(\bar{L}, \bar{A}) - x_{\ell}^{LP, g}(\bar{L}, \bar{A})}{(r_g(\bar{L}, \bar{A}))^{\alpha_g}} + \sum_{g \notin \bar{L}} \sum_{\ell \in \bar{A}_g} u_\ell \frac{x_{\ell}^{LP, g}(\bar{L}, \bar{A})}{(r_g(\bar{L}, \bar{A}))^{\alpha_g}} + \sum_{g \notin \bar{L}} \sum_{\ell \in \bar{A}_g} u_\ell \frac{x_{\ell}^{LP, g}(\bar{L}, \bar{A})}{(r_g(\bar{L}, \bar{A}))^{\alpha_g}}
\geq -\sum_{g \notin \bar{L}} \frac{2}{|A_g|} \frac{Z^{LP, g}(r_g(\bar{L}, \bar{A}), s_g(\bar{L}, \bar{A}), \bar{A}_g)}{(r_g(\bar{L}, \bar{A}))^{\alpha_g}} + \sum_{g \notin \bar{L}} \frac{Z^{LP, g}(r_g(\bar{L}, \bar{A}), s_g(\bar{L}, \bar{A}), \bar{A}_g)}{(r_g(\bar{L}, \bar{A}))^{\alpha_g}} + \sum_{g \notin \bar{L}} \sum_{\ell \in \bar{A}_g} u_\ell \frac{x_{\ell}^{LP, g}(\bar{L}, \bar{A})}{(r_g(\bar{L}, \bar{A}))^{\alpha_g}},
where the last inequality follows from Lemma 4.3. Our construction of \( \text{APPROX}(k, \delta) \) ensures that \( |\tilde{A}_g| = k \) for all \( g \in \tilde{L} \). Therefore,

\[
 f(A(k, \delta)) \\
\geq \frac{2}{k} \sum_{g \in \tilde{L}} \frac{Z^{LP,g}(r_g(\tilde{L}, \tilde{A}), s_g(\tilde{L}, \tilde{A}), \tilde{A}_g)}{(r_g(\tilde{L}, \tilde{A}))^{u_g}} + \sum_{g \in \tilde{L}} \frac{Z^{LP,g}(r_g(\tilde{L}, \tilde{A}), s_g(\tilde{L}, \tilde{A}), \tilde{A}_g)}{(r_g(\tilde{L}, \tilde{A}))^{u_g}} + \sum_{\ell \in \tilde{A}_g} \frac{\sum_{\ell \in \tilde{A}_g} u_{\ell}}{(\sum_{\ell \in \tilde{A}_g} u_{\ell})^{a_g}}
\]

\[
= \left(1 - \frac{2}{k}\right) \sum_{g \in \tilde{L}} \frac{Z^{LP,g}(r_g(\tilde{L}, \tilde{A}), s_g(\tilde{L}, \tilde{A}), \tilde{A}_g)}{(r_g(\tilde{L}, \tilde{A}))^{u_g}} + \sum_{g \not\in \text{Large}} \frac{\sum_{\ell \in \tilde{A}_g} u_{\ell}}{(\sum_{\ell \in \tilde{A}_g} u_{\ell})^{a_g}}
\]

\[
\geq \left(1 - \frac{2}{k}\right) V(\tilde{L}, \tilde{A}, \delta) \geq \left(1 - \frac{2}{k}\right) V(\text{Large}, T, \delta),
\]

where the final inequality follows from the definition of \( V(\tilde{L}, \tilde{A}, \delta) \) and the fact that \( |T_g| \leq k \) for all \( g \not\in \text{Large} \) and \( |T_g| = k \) for all \( g \in \text{Large} \). Since \( Z^* - \tilde{Z}(k, \delta) \leq (1 + \delta) V(\text{Large}, T, \delta) - f(A(k, \delta)) \), we have that

\[
Z^* - \tilde{Z}(k, \delta) \leq \left(\delta + \frac{2}{k}\right) V(\text{Large}, T, \delta) \leq \frac{k}{k - 2}\left(\delta + \frac{2}{k}\right) Z^*,
\]

where the final inequality follows from the lower bound in Lemma 4.5 and the fact that \( |T_g| = k \) for all \( g \in \text{Large} \).

It remains to establish the running time of the \( \text{APPROX}(k, \delta) \) algorithm. We need to solve an optimization associated with \( V(\mathcal{L}, \mathcal{A}, \delta) \) for each \( \mathcal{L} \subseteq \{1, 2, \ldots, G\} \) and \( \mathcal{A} = (A_1, A_2, \ldots, A_G) \) such that \( A_g \subseteq H_g \) and \( |A_g| \leq k \) for all \( g \). This represents a total of \( O(2^G \times N^{Gk}) = O(N^{G(k+1)}) \) optimization problems. Let \( v^* = \max_{\ell} v_{\ell} \). Each \( V(\mathcal{L}, \mathcal{A}, \delta) \) can be determined by enumerating \( (r_g, s_g) \in \text{Dom}(\delta) \times \text{Dom}(\delta) \) for all \( g \in \mathcal{L} \). It suffices to restrict \( r_g \) to be at most \( Nv^* \) and \( s_g \) to be at most \( C \). For each pair \( (r_g, s_g) \), we need solve a linear program associated with \( Z^{LP,g}(r_g, s_g, A_g) \) that has at most \( N \) variables and \( 2 \) constraints, and the coefficients and upper bound constraints are bounded above by \( \max\{\max_{\ell} c_{\ell}, \max_{\ell} u_{\ell}, Nv^*, C\} \). Let \( \text{LP} \) denote an upper bond on the complexity of solving each of these linear programs. Since we have at most \( G \) such linear programs for each combination \( (r_g, s_g : g \in \mathcal{L}) \), computing \( V(\mathcal{L}, \mathcal{A}, \delta) \) requires a total of \( O(G \times \text{LP} \times \frac{\log(N \cdot v^*)}{\log(1 + \delta)} \times \frac{\log C}{\log(1 + \delta)}) \) evaluations. Thus, the total computation time for \( \text{APPROX}(k, \delta) \) algorithm is of order

\[
O \left( G \times \text{LP} \times N^{G(k+1)} \times \left(\frac{\log(N \cdot v^*)}{\log(1 + \delta)} \times \frac{\log C}{\log(1 + \delta)}\right)^G \right) = O \left( G \times \text{LP} \times \left(\frac{N^{k+1} \times \log(N \cdot v^*) \times \log C}{\log^2(1 + \delta)}\right)^G \right).
\]

### 4.2 Proof of Lemma 4.5

Let \( \mathcal{A} = (A_1, \ldots, A_G) \) with \( A_g \subseteq \hat{A}_g \) for all \( g \) and \( \mathcal{L} \subseteq \{1, \ldots, G\} \) be given, along with \( \delta > 0 \). Also, suppose that \( \min_{\ell \in A_g u_{\ell}} \geq \max_{\ell \in \hat{A}_g \setminus A_g} u_{\ell} \) for each \( g \in \mathcal{L} \) and \( A_g = \hat{A}_g \) for each \( g \not\in \mathcal{L} \).

We will first establish the upper bound. For each \( g \in \mathcal{L} \), let \( r_g^* = \sum_{\ell \in A_g} v_{\ell} \) and \( s_g^* = \sum_{\ell \in \hat{A}_g} c_{\ell} \). Define \( r_g(\delta) \in \text{Dom}(\delta) \) and \( s_g(\delta) \in \text{Dom}(\delta) \) as follows:

\[
r_g(\delta) = \max \left\{ r \in \text{Dom}(\delta) : r \leq r_g^* \right\} \quad \text{and} \quad s_g(\delta) = \max \left\{ s \in \text{Dom}(\delta) : s \leq s_g^* \right\}.
\]

By definition, we have that \( r_g(\delta) \leq \sum_{\ell \in A_g} v_{\ell} < (1 + \delta)r_g(\delta) \) and \( s_g(\delta) \leq \sum_{\ell \in \hat{A}_g} c_{\ell} < (1 + \delta)s_g(\delta) \). Moreover, for each \( g \in \mathcal{L} \), since \( A_g \subseteq \hat{A}_g \) and \( \min_{\ell \in A_g u_{\ell}} \geq \max_{\ell \in \hat{A}_g \setminus A_g} u_{\ell} \), it follows from our definition that \( A_g \) is a
feasible solution for the linear program associated with $Z^{L^P,g} ((1 + \delta)r_g(\delta), (1 + \delta)s_g(\delta), A_g)$, which implies that $\sum_{\ell \in A_g^*} u_\ell \leq Z^{L^P,g} ((1 + \delta)r_g(\delta), (1 + \delta)s_g(\delta), A_g)$. It then follows from the definition that

$$\frac{\sum_{\ell \in A_g^*} u_\ell}{Z^{L^P,g}(r_g(\delta), s_g(\delta), A_g)} \leq \frac{Z^{L^P,g}((1 + \delta)r_g(\delta), (1 + \delta)s_g(\delta), A_g)}{Z^{L^P,g}(r_g(\delta), s_g(\delta), A_g)} \leq 1 + \delta,$$

where the last inequality follows from Lemma 4.4. The above result implies that

$$Z^* = \sum_{g \in \mathcal{L}} \left( \frac{\sum_{\ell \in A_g^*} u_\ell}{(\sum_{\ell \in A_g^*} v_\ell)^{\alpha_g}} \right) \frac{Z^{L^P,g}(\bar{r}_g, \bar{s}_g, A_g)}{(\bar{r}_g)^{\alpha_g}} \leq (1 + \delta) \sum_{g \in \mathcal{L}} \left( \frac{\sum_{\ell \in \mathcal{L}} u_\ell}{(\sum_{\ell \in \mathcal{L}} v_\ell)^{\alpha}} \right) \leq (1 + \delta)V(\mathcal{L}, A, \delta),$$

where the last inequality follows because, by definition of $s_g(\delta)$ and the fact that $A_g = A_g^*$ for all $g \notin \mathcal{L}$,

$$\sum_{g \in \mathcal{L}} s_g(\delta) \leq \sum_{g \in \mathcal{L}} \sum_{\ell \in A_g^*} c_\ell \leq C - \sum_{g \notin \mathcal{L}} \sum_{\ell \in \mathcal{L}} c_\ell = C - \sum_{g \notin \mathcal{L}} \sum_{\ell \in A_g^*} c_\ell.$$

Therefore, $(r_g(\delta), s_g(\delta) : g \in \mathcal{L})$ is a feasible solution to the optimization problem associated with $V(\mathcal{L}, A, \delta)$, giving us the desired upper bound.

We will now proceed to establish a lower bound. Suppose that $V(\mathcal{L}, A, \delta) = \sum_{g \in \mathcal{L}} Z^{L^P}(\bar{r}_g, \bar{s}_g, A_g)$ and an assortment $\mathcal{A}' = (A_1', A_2', ..., A_G')$ as follows: for each $g \in \{1, 2, ..., G\}$,

$$r_g' = \begin{cases} r_g^*, & \text{if } g \in \{1, ..., G\} \setminus \mathcal{L} \\ \bar{r}_g, & \text{if } g \in \mathcal{L} \end{cases} \quad \text{and} \quad A_g' = \begin{cases} A_g^*, & \text{if } \{1, ..., G\} \setminus \mathcal{L} \\ \left\{ \ell : x^{L^P,g}_\ell (\bar{r}_g, \bar{s}_g, A_g) = 1 \right\}, & \text{if } g \in \mathcal{L} \end{cases}.$$

By our definition of these linear programs (see Equation (3)), observe that for each $g \in \mathcal{L}$,

$$\sum_{\ell \in A_g} v_\ell = \sum_{\ell \in H_g} v_\ell \left[ x^{L^P,g}_\ell (\bar{r}_g, \bar{s}_g, A_g) \right] \leq \bar{r}_g,$$

and

$$\sum_{g \notin \mathcal{L}} \sum_{\ell \in \mathcal{L}} c_\ell \leq \sum_{g \notin \mathcal{L}} \sum_{\ell \in \mathcal{L}} c_\ell \left[ x^{L^P,g}_\ell (\bar{r}_g, \bar{s}_g, A_g) \right] \leq \sum_{g \in \mathcal{L}} \bar{s}_g \leq C - \sum_{g \notin \{1, ..., G\} \setminus \mathcal{L}} \sum_{\ell \in A_g^*} c_\ell.$$

The above inequalities imply that $(A_1', A_2', ..., A_G')$ is a feasible solution to the binary integer program associated with $Z^{IP}(r_1^*, r_2^*, ..., r_G^*)$ (see Equation (2)). Therefore, it follows from Lemma 4.2 that

$$Z^* = Z^{IP}(r_1^*, r_2^*, ..., r_G^*) \geq Z^{IP}(r_1', r_2', ..., r_G') \geq \frac{\sum_{g=1}^{G} \sum_{\ell \in A_g^*} u_\ell}{(r_g')^{\alpha_g}}.$$
Since \( r'_g = r^*_g = \sum_{\ell \in A_g} v_{\ell} \) and \( A'_g = A^*_g \) for all \( g \in \{1, \ldots, G\} \setminus L \), we have that

\[
\frac{\sum_{g \in L} \frac{\sum_{\ell \in A^*_g} u_{\ell}}{(r^'_g)^{a_g}}}{\sum_{g \in L} \left( \sum_{\ell \in A^*_g} v_{\ell} \right)^{a_g}} \geq \frac{\sum_{g \in L} \sum_{\ell \in A'_g} u_{\ell}}{(r^*_g)^{a_g}} = \sum_{g \in L} \frac{\sum_{\ell \in H_g} u_{\ell} \left| x_{\ell}^{LP,g} (\bar{r}_g, \bar{s}_g, A_g) \right|}{(\bar{r}_g)^{a_g}}
\]

Finally, putting everything together and using Lemma 4.3, we have that

\[
V(\mathcal{L}, A, \delta) - Z^* = \frac{\sum_{g \in L} Z^{LP,g} (\bar{r}_g, \bar{s}_g, A_g)}{(\bar{r}_g)^{a_g}} - \sum_{g \in L} \frac{\sum_{\ell \in A^*_g} u_{\ell}}{(r^*_g)^{a_g}} \leq \sum_{g \in L} \frac{Z^{LP,g}(\bar{r}_g, \bar{s}_g, A_g)}{(\bar{r}_g)^{a_g}} - \sum_{g \in L} \frac{\sum_{\ell \in H_g} u_{\ell} \left| x_{\ell}^{LP,g} (\bar{r}_g, \bar{s}_g, A_g) \right|}{(\bar{r}_g)^{a_g}} = \sum_{g \in L} \frac{\sum_{\ell \in H_g} u_{\ell} \left\{ x_{\ell}^{LP,g} (\bar{r}_g, \bar{s}_g, A_g) - \left| x_{\ell}^{LP,g} (\bar{r}_g, \bar{s}_g, A_g) \right| \right\}}{(\bar{r}_g)^{a_g}} \leq \frac{2}{\min\{|A_g| : g \in L\}} \sum_{g \in L} \frac{Z^{LP,g}(\bar{r}_g, \bar{s}_g, A_g)}{(\bar{r}_g)^{a_g}} \leq \frac{2}{\min\{|A_g| : g \in L\}} V(\mathcal{L}, A, \delta),
\]

which is the desired result.

**Acknowledgement**

The first author would like to thank Mike Todd for insightful and stimulating discussions on the structure of the linear programs associated with this problem. This research is supported in part by the National Science Foundation through grants DMS-0732196, CMMI-0746844, CMMI-0621433, CCR-0635121, and CMMI-0727640.

**References**


