New compact formulations for choice network revenue management

Sumit Kunnumkal* Kalyan Talluri†

August 13, 2015

Abstract

The choice network revenue management model incorporates customer purchase behavior as a function of the offered products, and is the appropriate model for airline and hotel network revenue management, dynamic sales of bundles, and dynamic assortment optimization. The optimization problem is a stochastic dynamic program and is intractable, and a linear program approximation called choice deterministic linear program ($CDLP$) is usually used to generate dynamic controls. Recently a compact linear programming formulation of $CDLP$ for the multinomial logit (MNL) model of customer choice has been proposed. In this paper we obtain a better approximation to the dynamic program than $CDLP$ while retaining the appealing properties of a compact linear programming representation. Our formulation is based on the affine relaxation of the dynamic program. We first show that the affine relaxation is NP-complete even for a single-segment MNL model. Nevertheless, by analyzing the affine relaxation we derive new linear programs that approximate the dynamic programming value function provably better, between the $CDLP$ value and the affine relaxation, and often coming close to the latter in our numerical experiments. We give extensions to the case with multiple customer segments where choice by each segment is according to the MNL model. Finally we perform extensive numerical comparisons on the various methods to evaluate their performance.

1 Introduction

Revenue management (RM) involves controlling the availability of products to customers who arrive over time to purchase them. RM models incorporating customer choice behavior have received much attention in recent years as the purchasing decision depends on the assortment of products made available for sale. Therefore, decisions on what products to make available for sale (the offer set) have to consider resource availabilities as well as estimates of customer purchase probabilities as a function of the offer set.

In Network Revenue Management (NRM), the products consume multiple resources, and is relevant for the airline, advertising, hotel and car rental industries. In the canonical airline example, products correspond to itineraries while resources correspond to seat capacities on the flight legs; for hotels, the products correspond to multi-night stays while resources are hotel rooms. The NRM

*Indian School of Business, Hyderabad, 500032, India, email: sumit_kunnumkal@isb.edu

†ICREA and Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain, email: kalyan.talluri@upf.edu

problem can be formulated as a stochastic dynamic program. However, solving the optimality equations and computing the value functions become intractable even for moderately sized problems. We refer the reader to Talluri and van Ryzin [19] for background on NRM.

Considering the intractability of the NRM dynamic program, Gallego, Iyengar, Phillips, and Dubey [6] and Liu and van Ryzin [11] proposed a linear-programming approximation called the $CDLP$ (Choice Deterministic Linear Program). The optimal objective function value of this linear program gives an upper bound on the value function. Upper bounds are useful both for deriving controls from them, as well as to assess the sub-optimality of policies. $CDLP$ however, has a drawback: the number of columns are exponential in the number of products, and so it has to be solved using column generation. Liu and van Ryzin [11] show that the $CDLP$ column generation procedure is tractable for the multinomial logit (MNL) choice model with multiple customer segments when the customers' consideration sets do not overlap. More recently, Gallego, Ratliff, and Shebalov [7] show that $CDLP$ has a compact linear programming formulation under the MNL model with disjoint consideration sets. On the other hand, if the consideration sets overlap, $CDLP$ is intractable even for the MNL model, as shown by Bront, Méndez-Díaz, and Vulcano [3], Rusmevichientong, Shmoys, Tong, and Topaloglu [16]. Related to this body of research, Zhang and Adelman [23] propose an affine relaxation to the NRM dynamic program and show that it obtains a tighter upper bound than $CDLP$.

This paper builds on these advances and makes the following research contributions:

1. We show that the affine relaxation of Zhang and Adelman [23] is NP-hard even for the single-segment MNL model, arguably one of the simplest possible choice models. This motivates solution methods that tighten the $CDLP$ bound and remain tractable at least for the single-segment MNL model.

2. We propose new, compact linear programming formulations that give a tighter bound on the dynamic program value function than $CDLP$, improving upon the work of Gallego et al. [7]. Compact representations are attractive from an implementation perspective since it eliminates the need for customized coding in the form of constraint-separation or column-generation techniques. To our knowledge, these are the first tractable approximation methods that are provably tighter than $CDLP$.

3. We show how our ideas can be extended to the mixture of multinomial logits (MMNL) model that can approximate any random utility choice model arbitrarily closely; McFadden and Train [12].

4. We propose control policies based on the new formulations and test their performance through an extensive numerical study. We show that our methods can yield noticeable benefits both in terms of tighter bounds and improved revenue performance.

Our computational experiments reveal that the proposed methods strike a good balance between the improved quality of the bound versus the increase in solution time. An interesting observation is that the revenue improvements from our methods typically exceed the corresponding improvements in the upper bounds. Our methods obtains sharper value function approximations towards the end of the selling horizon when capacity is relatively scarce. It turns out that the ability to make improved decisions in such capacity constrained scenarios has a significant payoff.

The remainder of the paper is organized as follows: In §2 we review the literature and in §3 we describe the choice NRM model, the notation, and the basic dynamic program. In §4 we describe the $CDLP$ and the affine relaxation of the NRM dynamic program. Next, in §5 we show that the affine relaxation is NP-hard even for the single-segment MNL model. We describe our first tractable approximation method in §6 and build on it to obtain tractable, tighter approximations in §7. §8 discusses extensions to variants of the MNL model including the MMNL model. §9 contains our computational study using the new formulations.

2 Literature review

As mentioned earlier, our paper advances the line of research on choice NRM initiated in Gallego et al. [6], Liu and van Ryzin [11] and Zhang and Adelman [23]. These papers use the well-known MNL choice model (see Ben-Akiva and Lerman [2]) which is attractive from an estimation and optimization standpoint, and has been widely used in transportation modeling, operations and marketing. For instance, van Ryzin and Mahajan [21] and Topaloglu [20] use the model in retail assortment planning, and Feldman, Liu, Topaloglu, and Ziya [5] in an application in healthcare. Dai, Ding, Kleywegt, Wang, and Zhang [4] describe a choice RM project at a major airline where they find that optimization with the MNL model gives superior policies to more complicated models.

Under the MNL choice model, Liu and van Ryzin [11] show that column generation for $CDLP$ can be efficiently carried out if the consideration sets of the segments are disjoint (consideration sets arise this way: the population is made of segments, and each segment is interested in only a subset of products, its consideration set; for instance a segment may be interested in an itinerary and its consideration set are the flights on this itinerary). Recently Gallego et al. [7] give an equivalent, compact linear programming formulation of $CDLP$ for the MNL choice model which eliminates the need for column generation. On the other hand, if the consideration sets overlap, $CDLP$ is intractable even for the MNL model as shown in Bront et al. [3] and Rusmevichientong et al. [16].

There are a number of other approaches to obtain upper bounds on the NRM value function. Zhang and Adelman [23] and Meissner and Strauss [13] use the linear programming approach to approximate dynamic programming with Zhang and Adelman [23] using affine approximations, and Meissner and Strauss [13] piecewise-linear approximations. Talluri [18] proposes a segment-based concave program and Meissner, Strauss, and Talluri [14] show how to further strengthen this formulation by adding equalities called product-cuts.

There are two important dimensions to assess the different approximation methods. One is the quality of the upper bound and the other is computational tractability. On the quality dimension, the approaches proposed by Zhang and Adelman [23] and Meissner and Strauss [13] are provably tighter than $CDLP$. However, in this paper, we show that the affine relaxation (AF) of Zhang and Adelman [23] turns out to be intractable even for the MNL model with a single segment. Since piecewise-linear value function approximations include affine functions as a special case, this implies a similar hardness result for the approach proposed by Meissner and Strauss [13].

3 Problem formulation

A product is a specification of a price and the set of resources that it consumes. For example, a product would be an itinerary-fare class combination for an airline network, where an itinerary is a combination of flight legs; in a hotel network, a product would be a multi-night stay for a particular room type at a certain price point.

Time is discrete and the sales horizon is assumed to consist of τ intervals, indexed by t. The sales horizon begins at time $t = 1$ and ends at $t = \tau$; all the resources perish instantaneously at time $\tau + 1$. We make the standard assumption that the time intervals are fine enough so that the probability of more than one customer arriving in any single time period is negligible.

We let $\mathcal{I}$ denote the set of resources and $\mathcal{J}$ the set of products. We index resources by i and products by j. We let f_j denote the revenue associated with product j and use $\mathcal{I}_j \subseteq \mathcal{I}$ to denote the set of resources used by product j. We let $\mathbb{1}_{[\cdot]}$ denote the indicator function, 1 if true and 0 if false and $\mathbb{1}_{[\mathcal{I}_j]}$ denote the vector of resources used by product j, with a 1 in the ith position if $i \in \mathcal{I}_j$ and a 0 otherwise. We use $\mathcal{J}_i \subseteq \mathcal{J}$ to denote the set of products that use resource i.

In each period the firm offers a subset S of its products for sale, called the *offer set*. We write $i \in \mathcal{I}_S$ whenever there is a $j \in S$ with $i \in \mathcal{I}_j$; that is, there is at least one product in the offer set S that uses resource i.

We use superscripts on vectors to index the vectors (for example, the resource capacity vector associated with time period t would be $\boldsymbol{r}^t$) and subscripts to indicate components (for example, the capacity on resource i in time period t would be r_i^t). Therefore, $\boldsymbol{r}^1 = [r_i^1]$ represents the initial capacity on the resources and $\boldsymbol{r}^t = [r_i^t]$ denotes the remaining capacity on the resources at the beginning of time period t. The remaining capacity r_i^t takes values in the set $\mathcal{R}_i = \{0, \dots, r_i^1\}$ and $\mathcal{R} = \prod_i \mathcal{R}_i$ represents the state space at each time t.

3.1 Demand model

We have multiple customer segments, each with distinct purchase behavior. We let $\mathcal{L}$ denote the set of customer segments. In each period a customer from segment $l \in \mathcal{L}$ arrives with probability λ_l so that $\lambda = \sum_l \lambda_l$ is the total arrival rate. Note that conditioned on a customer arrival, λ_l/λ is the probability that the customer belongs to segment l.

Customer segment l has a *consideration set* $\mathcal{C}_l \subseteq \mathcal{J}$ of products that it considers for purchase. We assume this consideration set is known to the firm (by a previous process of estimation and analysis). The choice probabilities of a segment-l customer are not affected by products not in its consideration set. Given an offer set S, an arriving customer in segment l purchases a product j in the set $S_l = \mathcal{C}_l \cap S$ or leaves without making a purchase. The no-purchase option is indexed by 0 and is always present for the customer.

Within each segment, choice is according to the MNL model. The MNL model associates a preference weight with each alternative including the no-purchase alternative. We let w_j^l denote the preference weight associated with a segment-l customer for product j. Without loss of generality, by suitably normalizing the weights, we set the no-purchase weight w_0^l to be 1. The probability that a segment-l customer purchases product j when S is the offer set is

$$P_j^l(S) = \frac{w_j^l \mathbb{1}_{[j \in S_l]}}{1 + \sum_{k \in S_l} w_k^l}. \tag{1}$$

The probability that the customer does not purchase anything is $P_0^l(S) = 1/(1 + \sum_{k \in S_l} w_k^l)$. We note that the preference weights are inputs to our model; estimating them is outside the scope of the paper. We refer the reader to Ben-Akiva and Lerman [2] for further background on this popular choice model.

Given a customer arrival, and an offer set S, the probability that the firm sells $j \in S$ is given by $P_j(S) = \sum_l \frac{\lambda_l}{\lambda} P_j^l(S)$ and makes no sale with probability $P_0(S) = 1 - \sum_{j \in S} P_j(S)$. The expected sales for product j is therefore $\lambda P_j(S) = \sum_l \lambda_l P_j^l(S)$, while $1 - \lambda + \lambda P_0(S) = 1 - \sum_{j\in S} \lambda P_j(S)$ is the probability of no sales in a time period. Given an offer set S, $Q_i^l(S) = \sum_{j\in\mathcal{J}_i} P_j^l(S)$ denotes the expected capacity consumed on resource i conditional on a segment-l customer arrival and $Q_i(S) = \sum_l \frac{\lambda_l}{\lambda} Q_i^l(S)$ denotes the expected capacity consumed on resource i conditional on a customer arrival. Note that $\lambda Q_i(S) = \sum_l \lambda_l Q_i^l(S)$ gives the expected capacity consumed on resource i in a time period. The revenue functions can be written as $R^l(S) = \sum_{j\in S_l} f_j P_j^l(S_l)$ and $R(S) = \sum_{j\in S} f_j P_j(S)$.

In what follows if we are considering a single-segment MNL model, we drop the subscript l from the probabilities.

We assume that the arrival rates and choice probabilities are stationary. This is for brevity of notation and all of our results go through with non-stationary arrival rates and choice probabilities.

3.2 Choice dynamic program

The dynamic program (DP) to determine optimal controls is as follows. Let $V_t(\boldsymbol{r}^t)$ denote the maximum expected revenue to go, given remaining capacity $\boldsymbol{r}^t$ at the beginning of period t. Then $V_t(\boldsymbol{r}^t)$ must satisfy the Bellman equation

$$V_t(\boldsymbol{r}^t) = \max_{S\subseteq\mathcal{S}(\boldsymbol{r}^t)} \left\{ \sum_{j\in S} \lambda P_j(S) \left[f_j + V_{t+1}\left(\boldsymbol{r}^t - \mathbb{1}_{[\mathcal{I}_j]}\right)\right] + \left[\lambda P_0(S) + 1 - \lambda\right] V_{t+1}\left(\boldsymbol{r}^t\right) \right\}, \tag{2}$$

where

$$\mathcal{S}(\boldsymbol{r}) = \left\{ j \mid \mathbb{1}_{[i\in\mathcal{I}_j]} \le r_i \ \forall i \right\}$$

represents the set of products that can be offered given the capacity vector $\boldsymbol{r}$. The boundary conditions are $V_{\tau+1}(\boldsymbol{r}) = V_t(\boldsymbol{0}) = 0$ for all $\boldsymbol{r}$ and for all t, where $\boldsymbol{0}$ is a vector of all zeroes. $V^{DP} = V_1(\boldsymbol{r}^1)$ denotes the optimal expected total revenue over the sales horizon, given the initial capacity vector $\boldsymbol{r}^1$.

3.3 Linear programming formulation of the dynamic program

The value functions can, alternatively, be obtained by solving a linear program (LP). The linear programming formulation of (2) has a decision variable for each state vector in each period $V_t(\boldsymbol{r})$ and is as follows:

$$\begin{aligned} V^{DPLP} = \min_{V} \quad & V_1(\boldsymbol{r}^1) \qquad (3)\\ (DPLP) \text{ s.t.} \quad & V_t(\boldsymbol{r}) \ge \sum_j \lambda P_j(S) \left[f_j + V_{t+1}\left(\boldsymbol{r} - \mathbb{1}_{[\mathcal{I}_j]}\right) - V_{t+1}\left(\boldsymbol{r}\right)\right] + V_{t+1}\left(\boldsymbol{r}\right) \\ & \qquad \forall\, \boldsymbol{r} \in \mathcal{R}, S \subseteq \mathcal{S}(\boldsymbol{r}), t. \end{aligned}$$

Both dynamic program (2) and linear program $DPLP$ are computationally intractable, but linear program $DPLP$ turns out to be useful in developing value function approximation methods, as shown in Zhang and Adelman [23].

4 Approximations and upper bounds

In the following, we outline the two approximations studied in this paper. We first describe the choice deterministic linear program and then outline the affine relaxation method.

4.1 Choice deterministic linear program ($CDLP$)

The choice deterministic linear program ($CDLP$) proposed in Gallego et al. [6] and Liu and van Ryzin [11] is a certainty-equivalence approximation to (2). We write $CDLP$ as the following LP:

$$\begin{aligned} V^{CDLP} = \max_h \quad & \sum_t \sum_S \lambda R(S) h_{S,t} \\ (CDLP) \text{ s.t.} \quad & \sum_{k=1}^{t} \sum_S \lambda Q_i(S) h_{S,k} \le r_i^1 \quad \forall i, t & (4) \\ & \sum_S h_{S,t} = 1 \quad \forall t & (5) \\ & h_{S,t} \ge 0 \quad \forall S, t. \end{aligned}$$

The decision variable $h_{S,t}$ can be interpreted as the frequency with which set S (including the empty set) is offered at time period t . The first set of constraints ensure that the total expected capacity consumed on resource i up until time period t does not exceed the available capacity. Note that since $h_{S,t} \ge 0$, constraints (4) are redundant except for the last time period. Still, this expanded formulation is useful when we compare $CDLP$ with other approximation methods. The second set of constraints states that the sum of the frequencies adds up to 1.

The dual of $CDLP$ turns out to be useful in our analysis. Associating dual variables $\gamma = \{\gamma_{i,t} \,|\, \forall i, t\}$ with constraints (4) and $\beta = \{\beta_t \,|\, \forall t\}$ with constraints (5), the dual of $CDLP$ is

$$\begin{aligned} V^{dCDLP} = \min_{\beta,\gamma} \quad & \sum_t \beta_t + \sum_t \sum_i \gamma_{i,t} r_i^1 \\ (dCDLP) \text{ s.t.} \quad & \beta_t + \sum_i \left(\sum_{k=t}^{\tau} \gamma_{i,k} \right) \lambda Q_i(S) \ge \lambda R(S) \quad \forall t, S & (6) \\ & \gamma_{i,t} \ge 0 \quad \forall i, t. \end{aligned}$$

Liu and van Ryzin [11] show that the optimal objective function value of $CDLP$, V^{CDLP} is an upper bound on V^{DPLP}. Besides giving an upper bound on the value function, $CDLP$ can also be used to construct different heuristic control policies. Let $\hat{\gamma} = \{\hat{\gamma}_{i,t} \,|\, \forall i, t\}$ denote the optimal values of the dual variables associated with constraints (4), we interpret $\hat{\gamma}_{i,t}$ as giving the value of an additional unit of capacity on resource i from time period t to $t+1$. With this interpretation, $\sum_{s=t}^{\tau} \hat{\gamma}_{i,s}$ gives the marginal value of capacity on resource i at time period t. Zhang and Adelman [23] approximate the value function as

$$\hat{V}_t(\boldsymbol{r}) = \sum_i \left(\sum_{s=t}^{\tau} \hat{\gamma}_{i,s} \right) r_i \tag{7}$$

and if $\boldsymbol{r}^t$ is the vector of remaining resource capacities at time t, solve the problem

$$\max_{S \subseteq \mathcal{S}(\boldsymbol{r}^t)} \left\{ \sum_{j \in S} \lambda P_j(S) \left[f_j + \hat{V}_{t+1}\left(\boldsymbol{r}^t - \mathbb{1}_{[\mathcal{I}_j]}\right)\right] + \left[\lambda P_0(S) + 1 - \lambda\right] \hat{V}_{t+1}\left(\boldsymbol{r}^t\right) \right\}, \tag{8}$$

and use the policy of offering the set that achieves the maximum in the above optimization problem.

The number of decision variables in $CDLP$ is exponential in the number of products and so it has to be solved using column generation. The tractability of column generation depends on the underlying choice model. Liu and van Ryzin [11] show that the column generation procedure can be efficiently carried out when choice is according to the MNL model and the consideration sets of the different segments do not overlap. That is, we have $\mathcal{C}_l \cap \mathcal{C}_m = \emptyset$ for segments l and m. Under the same set of assumptions, Gallego et al. [7] further show that $CDLP$ has the following equivalent, compact formulation

$$\begin{aligned}
V^{SBLP} = \max_{x} \quad & \sum_t \sum_l \sum_{j \in \mathcal{C}_l} \lambda_l f_j x^l_{j,t} \\
(SBLP) \text{ s.t.} \quad & \sum_t \sum_l \sum_{j \in \mathcal{J}_i \cap \mathcal{C}_l} \lambda_l x^l_{j,t} \le r^1_i \quad \forall i, t \\
& x^l_{0,t} + \sum_{j \in \mathcal{C}_l} x^l_{j,t} = 1 \quad \forall l, t \\
& \frac{x^l_{j,t}}{w^l_j} - x^l_{0,t} \le 0 \quad \forall l, j \in \mathcal{C}_l, t \\
& x^l_{0,t}, x^l_{j,t} \ge 0 \quad \forall l, j, t.
\end{aligned} \tag{9}$$

In the above sales-based linear program ($SBLP$), the decision variables $x^l_{j,t}$ can be interpreted as the sales rate for product j at time t. Note that $SBLP$ is a compact formulation since the number of constraints and decision variables is polynomial in the number of products and resources. On the other hand, if the consideration sets overlap, Bront et al. [3] and Rusmevichientong et al. [16] show that the $CDLP$ column generation is NP-complete even under the MNL choice model.

4.2 Affine relaxation

The second approximation method we consider is the affine relaxation, where the value function is approximated as $V_t(\boldsymbol{r}) = \theta_t + \sum_i V_{i,t} r_i$. Note that $V_{i,t}$ can be interpreted as the marginal value of capacity on resource i at time t. Substituting this value function approximation into the formulation $DPLP$ we get the affine relaxation LP

$$\begin{aligned}
V^{AF} = \min_{\theta, V} \quad & \theta_1 + \sum_i V_{i,1} r^1_i \\
(AF) \text{ s.t.} \quad & \theta_t + \sum_i V_{i,t} r_i \ge \sum_j \lambda P_j(S) \left[f_j - \sum_{i \in \mathcal{I}_j} V_{i,t+1} \right] + \theta_{t+1} + \sum_i V_{i,t+1} r_i \\
& \qquad \forall\, \boldsymbol{r} \in \mathcal{R}, S \subseteq \mathcal{S}(\boldsymbol{r}), t \\
& \theta_t \ge 0, V_{i,t} \ge 0 \quad \forall i, t
\end{aligned}$$

with the boundary conditions $\theta_{\tau+1} = 0, V_{i,\tau+1} = 0$. Zhang and Adelman [23] show that the optimal objective function value V^{AF} is an upper bound on the value function and that there exists an optimal solution $(\hat{\theta}, \hat{V})$ of AF that satisfies $\hat{V}_{i,t} - \hat{V}_{i,t+1} \ge 0$ for all i and t.

While the number of decision variables in AF is manageable, the number of constraints is exponential both in the number of products as well as the number of resources. Vossen and Zhang [22] use Dantzig-Wolfe decomposition to derive a reduced, equivalent formulation of AF, where the number of constraints is exponential only in the number of products.

We give an alternative, simpler proof of the reduction below. The analysis we present also turns out to be useful in the development of our tractable solution methods later. We make a change of variables $\beta_t = \theta_t - \theta_{t+1}$, and $\gamma_{i,t} = V_{i,t} - V_{i,t+1}$ and write AF equivalently as

$$\begin{aligned}
\min_{\beta,\gamma} \quad & \sum_t \beta_t + \sum_t \sum_i \gamma_{i,t} r_i^1 \\
\text{s.t.} \quad & \beta_t + \sum_i \gamma_{i,t} r_i + \sum_j \lambda P_j(S) \left[\left(\sum_{i \in \mathcal{I}_j} \sum_{k=t+1}^{\tau} \gamma_{i,k} \right) - f_j \right] \geq 0 \quad \forall \boldsymbol{r} \in \mathcal{R}, S \subseteq \mathcal{S}(\boldsymbol{r}), t \qquad (10) \\
& \gamma_{i,t} \geq 0 \quad \forall i, t,
\end{aligned}$$

where we use the fact that $V_{i,t} = \sum_{k=t}^{\tau} \gamma_{i,k}$ and so $\sum_{k=t}^{\tau} \gamma_{i,k}$ can be interpreted as the marginal value of capacity on resource i at time t. Note that the nonnegativity constraint on $\gamma_{i,t}$ is without loss of generality, since there exists an optimal solution to AF that satisfies $V_{i,t} - V_{i,t+1} \geq 0$.

Now, constraints (10) can be written as

$$\min_{\boldsymbol{r} \in \mathcal{R}, S \subseteq \mathcal{S}(\boldsymbol{r})} \left\{ \beta_t + \sum_i \gamma_{i,t} r_i + \sum_j \lambda P_j(S) \left[\left(\sum_{i \in \mathcal{I}_j} \sum_{k=t+1}^{\tau} \gamma_{i,k} \right) - f_j \right] \right\} \geq 0 \qquad (11)$$

for all t. Since $\gamma_{i,t} \geq 0$, the coefficient of r_i in minimization problem (11) is nonnegative, and we can assume $r_i \in \{0, 1\}$ in the minimization (as larger values of r_i would be redundant in $S \subseteq \mathcal{S}(\boldsymbol{r})$ and would only increase the objective value). Moreover, since $\gamma_{i,t} \geq 0$, for any set S, we have $r_i = 0$ for $i \notin \mathcal{I}_S$. On the other hand, feasibility requires we have $r_i = 1$ for $i \in \mathcal{I}_S$. Therefore, (11) can be written as

$$\min_S \left\{ \beta_t + \sum_i \mathbb{1}_{[i \in \mathcal{I}_S]} \gamma_{i,t} + \sum_j \lambda P_j(S) \left[\left(\sum_{i \in \mathcal{I}_j} \sum_{k=t+1}^{\tau} \gamma_{i,k} \right) - f_j \right] \right\} \geq 0.$$

and we can write AF equivalently as

$$\begin{aligned}
V^{RAF} = \min_{\beta,\gamma} \quad & \sum_t \beta_t + \sum_t \sum_i \gamma_{i,t} r_i^1 \\
(RAF) \text{ s.t.} \quad & \beta_t + \sum_i \mathbb{1}_{[i \in \mathcal{I}_S]} \gamma_{i,t} + \sum_i \left[\left(\sum_{k=t+1}^{\tau} \gamma_{i,k} \right) \lambda Q_i(S) \right] \geq \lambda R(S) \quad \forall t, S \qquad (12) \\
& \gamma_{i,t} \geq 0 \quad \forall i, t.
\end{aligned}$$

Notice that the number of constraints in the reduced formulation RAF is an order of magnitude smaller than AF. Taking the dual of RAF by associating dual variables $h_{S,t}$ with constraints (12), we get

$$
\begin{aligned}
V^{dRAF} = \max_{h} \quad & \sum_t \sum_S \lambda R(S) h_{S,t} \\
(dRAF) \text{ s.t.} \quad & \sum_S \left(\sum_{k=1}^{t-1} \lambda Q_i(S) h_{S,k} + \mathbb{1}_{[i \in \mathcal{I}_S]} h_{S,t} \right) \leq r_i^1 \quad \forall i, t \\
& \sum_S h_{S,t} = 1 \quad \forall t \\
& h_{S,t} \geq 0 \quad \forall S, t.
\end{aligned}
$$

The above arguments imply that

Proposition 1. *(Vossen and Zhang [22])* $V^{AF} = V^{RAF} = V^{dRAF}$.

We close this section with two remarks. First, in addition to giving an upper bound on the optimal expected total revenue, the affine relaxation can also be used to construct heuristic control policies. Letting $(\hat{\beta}, \hat{\gamma})$, with $\hat{\beta} = \{\hat{\beta}_t \,|\, \forall t\}$ and $\hat{\gamma} = \{\hat{\gamma}_{i,t} \,|\, \forall i, t\}$, denote an optimal solution to RAF, we use $\sum_{s=t}^{\tau} \hat{\gamma}_{i,k}$ to approximate the marginal value of capacity on resource i at time t. We approximate $V_t(\boldsymbol{r})$ using (7) and solve problem (8) using this value function approximation to decide on the set of products to be offered at time period t. Second, Zhang and Adelman [23] show that the upper bound obtained by AF is tighter than $CDLP$. In that sense, AF is a better approximation than $CDLP$. At the same time, it is important to understand the computational effort required by AF to obtain a tighter bound. We explore this question in the following section.

5 Tractability of the affine relaxation for MNL with a single segment

In this section, we focus on the tractability of the affine relaxation for the single-segment MNL model. We restrict our attention to the single-segment MNL since it is one of the few cases where $CDLP$ is tractable. We show that the affine relaxation is NP-complete even for this simple choice model.

Since we restrict attention to the single-segment MNL model, we drop the segment superscript l for ease of notation. So we write the preference weights as w_j, and the choice probabilities, expected resource consumptions and expected revenues as

$$
P_j(S) = \frac{\mathbb{1}_{[j \in S]} w_j}{1 + \sum_{k \in S} w_k} \qquad Q_i(S) = \frac{\sum_{j \in \mathcal{J}_i \cap S} w_j}{1 + \sum_{j \in S} w_j} \qquad R(S) = \frac{\sum_{j \in S} f_j w_j}{1 + \sum_{j \in S} w_j}.
$$

Since RAF has an exponential number of constraints, we have to generate on the fly constraints (12) violated by a solution. Following the result of Grötschel, Lovász, and Schrijver [8] polynomial-solvability of a linear program is equivalent to polynomial-time generation of violated constraints.

Substituting the MNL choice probabilities, expected resource consumptions and expected revenues into constraint (12), we obtain

$$
\beta_t + \gamma_{S,t} + \sum_i \left[\left(\sum_{k=t+1}^{\tau} \gamma_{i,k} \right) \lambda \frac{\sum_{j \in \mathcal{J}_i \cap S} w_j}{1 + \sum_{j \in S} w_j} \right] \geq \lambda \frac{\sum_{j \in S} f_j w_j}{1 + \sum_{j \in S} w_j}
$$

where

$$\gamma_{S,t} = \sum_i \mathbb{1}_{[i \in \mathcal{I}_S]} \gamma_{i,t}.$$

Multiplying both sides by the positive quantity $1+\sum_{j\in S} w_j$ and simplifying, constraint (12) of RAF can be equivalently written as

$$\beta_t \geq -\gamma_{S,t}\left(1+\sum_{j\in S} w_j\right) - \sum_{j\in S} \zeta_{j,t}(\beta,\gamma), \tag{13}$$

where

$$\zeta_{j,t}(\beta,\gamma) = w_j\left[\beta_t + \lambda\left(\left(\sum_{i\in\mathcal{I}_j}\sum_{k=t+1}^{\tau}\gamma_{i,k}\right) - f_j\right)\right].$$

Since the constraint has to be satisfied for every S and t, we have $\beta_t \geq \Pi_t^{AF}(\beta,\gamma)$ for all t, where

$$\Pi_t^{AF}(\beta,\gamma) = \max_S\left\{-\gamma_{S,t}\left(1+\sum_{j\in S} w_j\right) - \sum_{j\in S}\zeta_{j,t}(\beta,\gamma)\right\} \tag{14}$$

and the affine relaxation constraint (12) can be equivalently written as

$$\beta_t \geq \Pi_t^{AF}(\beta,\gamma) \quad \forall t. \tag{15}$$

Generating constraints on the fly involves checking, given a set of values (β,γ), if constraint (13) is satisfied for all S. If not, we add the violated constraint to the LP. In other words, the RAF separation problem at time t involves solving optimization problem (14) and determining if $\beta_t \geq \Pi_t^{AF}(\beta,\gamma)$. If $\beta_t \geq \Pi_t^{AF}(\beta,\gamma)$, then constraint (13) is satisfied for all S at time t. Otherwise, the set $\hat{S}$ which attains the maximum in problem (14) violates the constraint, and we add the constraint for set $\hat{S}$ to the LP. Therefore, solving problem (14) in an efficient manner is key to separating constraints (13) efficiently. Proposition 2 below states that the affine relaxation separation problem for MNL with a single segment, as given in (13) is NP-complete.

Proposition 2. *The following problem is NP-complete:*
Input: $w_j \geq 0$, $1 \geq \lambda \geq 0$, $f_j \geq 0$, *and values* β_t *and* $\gamma_{i,t} \geq 0$.
Question: *Is there a set S that violates (13)?*

Proof
Our reduction is from the NP-complete maximum edge biclique problem (Peeters [15]). We state first the definitions and notation in the problem.

The problem is defined on an undirected, bipartite graph $G = (V_1 \cup V_2, E)$, with $|V_2| = m_2$. A (k_1, k_2)-*biclique* is a complete bipartite subgraph of G, i.e., a subgraph consisting of a pair (X, Y) of vertex subsets $X \subseteq V_1$ and $Y \subseteq V_2$, $|X| = k_1 > 1, |Y| = k_2 > 1$, such that there exists an edge $(x, y) \in E$, $\forall x \in X, y \in Y$. Note that the number of edges in the biclique is $k_1 k_2$.

Maximum edge biclique problem (MBP)
Input: A bipartite graph $G = (V_1 \cup V_2, E)$ and a positive integer p.
Question: Does G contain a biclique with at least p edges.

Consider the complement bipartite graph $\bar{G}$ of G defined on the same vertex set as G, where there is an edge $e = (u, v)$ in graph $\bar{G}$ if and only if there is no edge between u and v in G.

Define a *cover* $C_S \subseteq V_2$ of a subset $S \subseteq V_1$ in the complement graph $\bar{G}$, as $C_S = \{v \in V_2 \mid \exists e = (u, v) \in \bar{G}, u \in S\}$. By definition if C_S is a cover of some subset S, it means there is no edge from *any* $u \in S$ to *any* $v \in V_2 \backslash C_S$ in the graph $\bar{G}$. Hence, as G is a complement of $\bar{G}$, there is an edge from every $u \in S$ to every $v \in V \backslash C(S)$ in G, thus representing a biclique between S and $V \backslash C(S)$ in the graph G.

Now we set up the reduction for the separation for (13). In equation (13), for each $u \in V_1$, we associate a product j with $f_j = m_2 \frac{(p+1)}{p}$ and $w_j = m_2$. For each $v \in V_2$, we associate a resource i with weights $\gamma_{i,t} = \frac{1}{p}$ and $\gamma_{i,k} = 0, k > t$. The resource consumptions of the products j are defined from the graph $\bar{G}$: j contains all the i such that there is an edge between the associated nodes in $\bar{G}$. We let $\lambda = 1, \beta_t = m_2$.

We now claim that G has a (k_1, k_2)-biclique with $k_1 k_2 > p$ if and only if there is a set S that violates the inequality (13) for this instance.

With the above values, $S \subseteq V_1$, with $|S| = k_1, |C(S)| = m_2 - k_2$ violates (13) if and only if

$$m_2 - \frac{\sum_{j \in S} \frac{(p+1)}{p} (m_2)^2}{(1 + \sum_{j \in S} m_2)} < - \sum_{i \in C(S)} \frac{1}{p}$$

or,

$$m_2 - \frac{(p+1) m_2 k_1}{p \left(\frac{1}{m_2} + k_1 \right)} < - \frac{(m_2 - k_2)}{p}$$

or multiplying both sides by the positive number $p(\frac{1}{m_2} + k_1)$,

$$m_2 p \left(\frac{1}{m_2} + k_1 \right) - (p+1) m_2 k_1 < -(m_2 - k_2) \left(\frac{1}{m_2} + k_1 \right)$$

or,

$$p < - \frac{(m_2 - k_2)}{m_2} + k_2 k_1.$$

The term $0 < \frac{(m_2 - k_2)}{m_2} < 1$ implies, if and only if

$$p < k_2 k_1.$$

□

Therefore, even though the affine relaxation tightens the $CDLP$ bound, it comes at a significant cost. This motivates the solution method that we propose in the following section, which tightens the $CDLP$ bound while remaining tractable.

6 Weak affine relaxation

In this section we propose our first tractable approximation method that tightens the $CDLP$ bound. We also show that our approximation method can, in fact, be formulated as a compact LP. In our initial development, we restrict attention to the single-segment MNL choice model. We emphasize that this is only for clarity of exposition. In §8 we show how the ideas can be readily extended to more realistic variants of the MNL model that consider multiple customer segments. Moreover, all of the test problems in our computational experiments involve multiple customer segments with and without overlapping consideration sets.

6.1 Preliminaries

All of our approximation methods involve solving an optimization problem of the form $\min_{\beta,\gamma} \sum_t \beta_t + \sum_t \sum_i \gamma_{i,t} r_i^1$ subject to the constraints $\beta_t \geq \Pi_t(\beta,\gamma)$, where $\Pi_t(\cdot,\cdot)$ is a scalar function of $\beta = \{\beta_t \,|\, \forall t\}$ and $\gamma = \{\gamma_{i,t} \,|\, \forall i,t\}$. The following observation is useful in comparing the upper bounds obtained by the different approximation methods.

Lemma 1. *Let*

$$\begin{aligned} V^I = \min_{\beta,\gamma} \quad & \sum_t \beta_t + \sum_t \sum_i \gamma_{i,t} r_i^1 \\ (I) \;\; s.t. \quad & \beta_t \geq \Pi_t^I(\beta,\gamma) \quad \forall t \\ & \gamma_{i,t} \geq 0 \quad \forall i,t, \end{aligned}$$

and let

$$\begin{aligned} V^{II} = \min_{\beta,\gamma} \quad & \sum_t \beta_t + \sum_t \sum_i \gamma_{i,t} r_i^1 \\ (II) \;\; s.t. \quad & \beta_t \geq \Pi_t^{II}(\beta,\gamma) \quad \forall t \\ & \gamma_{i,t} \geq 0 \quad \forall i,t. \end{aligned}$$

If $\Pi_t^I(\beta,\gamma) \leq \Pi_t^{II}(\beta,\gamma)$ for all t, then $V^I \leq V^{II}$.

Proof

The proof follows by noting that a feasible solution to problem (II) is also feasible to problem (I) and both optimization problems have the same objective function.

□

6.2 *CDLP* vs. *AF* for single-segment MNL

We begin by comparing the *CDLP* and *AF* separation problems for the single-segment MNL model. For this choice model, the *CDLP* constraints can be separated efficiently, while the *AF* separation problem is intractable. Comparing the *CDLP* and *AF* separation problems helps us identify the difficult term in the affine relaxation. Replacing this difficult term in the *AF* separation problem with something more tractable yields our approximation method.

Using the single-segment MNL formulas for the expected resource consumptions and expected revenues, the *CDLP* dual constraint (6) can be written as

$$\beta_t \geq -\sum_{j \in S} w_j \left[\beta_t + \lambda \left(\left(\sum_{i \in \mathcal{I}_j} \sum_{k=t}^{\tau} \gamma_{i,k} \right) - f_j \right) \right] \quad \forall t, S$$

which looks similar to the right-hand-side of (13) except that the inner summation over k runs from t instead of $t+1$. To make the comparison with AF easier, we rewrite the above constraint as

$$\beta_t \geq \Pi_t^{CDLP}(\beta,\gamma) \quad \forall t \tag{16}$$

where

$$\Pi_t^{CDLP}(\beta,\gamma) = \max_S \left\{ -\lambda \sum_{j\in S} w_j \left(\sum_{i\in\mathcal{I}_j} \gamma_{i,t} \right) - \sum_{j\in S} \zeta_{j,t}(\beta,\gamma) \right\}. \tag{17}$$

Since $0 \leq \lambda \leq 1$, and $\gamma_{S,t} = \sum_i \mathbb{1}_{[i\in\mathcal{I}_S]}\gamma_{i,t} \geq \sum_{i\in\mathcal{I}_j} \gamma_{i,t} \geq 0$ for all $j \in S$, we have

$$\gamma_{S,t} \left(1 + \sum_{j\in S} w_j \right) \geq \lambda \sum_{j\in S} w_j \left(\sum_{i\in\mathcal{I}_j} \gamma_{i,t} \right).$$

Therefore $\Pi_t^{AF}(\beta,\gamma) \leq \Pi_t^{CDLP}(\beta,\gamma)$ and by Lemma 1, $V^{AF} \leq V^{CDLP}$, which gives an alternative proof of the AF bound being tighter than the $CDLP$ bound. More importantly, the comparison hints at how we can obtain tractable relaxations that are tighter than $CDLP$.

6.3 A tractable approximation

We are now ready to describe our first tractable approximation method, which we refer to as weak affine relaxation (wAR). The difficult term in (14) is the $\gamma_{S,t}(1+\sum_{j\in S} w_j)$, and $CDLP$ is tractable as it replaces this by $\lambda \sum_{j\in S} w_j(\sum_{i\in\mathcal{I}_j} \gamma_{i,t})$. We instead replace the $\gamma_{S,t}(1+\sum_{j\in S} w_j)$ term in (14) with $\gamma_{S,t} + \sum_{j\in S} w_j(\sum_{i\in\mathcal{I}_j} \gamma_{i,t})$ and solve the linear program

$$\begin{aligned} V^{wAR} = \min_{\beta,\gamma} \quad & \sum_t \beta_t + \sum_t \sum_i \gamma_{i,t} r_i^1 \\ (wAR) \text{ s.t.} \quad & \beta_t \geq \Pi_t^{wAR}(\beta,\gamma) \quad \forall t \\ & \gamma_{i,t} \geq 0 \quad \forall i,t, \end{aligned} \tag{18}$$

where

$$\Pi_t^{wAR} = \max_S \left\{ -\gamma_{S,t} - \sum_{j\in S} w_j \left(\sum_{i\in\mathcal{I}_j} \gamma_{i,t} \right) - \sum_{j\in S} \zeta_{j,t}(\beta,\gamma) \right\}. \tag{19}$$

Proposition 3 below shows that wAR obtains an upper bound on the value function that is weaker than AF but stronger than $CDLP$. Kunnumkal and Talluri [9] show that it also gives a tighter upper bound than by working with a continuous relaxation of $\Pi_t^{AF}(\beta,\gamma)$.

Proposition 3. $V^{AF} \leq V^{wAR} \leq V^{CDLP}$.

Proof
The proof follows by noting that

$$\gamma_{S,t} \left(1 + \sum_{j\in S} w_j \right) \geq \gamma_{S,t} + \sum_{j\in S} w_j \left(\sum_{i\in\mathcal{I}_j} \gamma_{i,t} \right) \geq \lambda \sum_{j\in S} w_j \left(\sum_{i\in\mathcal{I}_j} \gamma_{i,t} \right).$$

Therefore $\Pi_t^{AF}(\beta,\gamma) \le \Pi_t^{wAR}(\beta,\gamma) \le \Pi_t^{CDLP}(\beta,\gamma)$ and the result now follows from Lemma 1.

□

In the remainder of this section, we show that the weak affine relaxation upper bound, V^{wAR}, can be obtained in a tractable manner; moreover we show that the weak affine relaxation LP can, in fact, be reformulated as a compact linear program where the number of variables and constraints is polynomial in the number of products and resources.

Observe that solving problem (19) in an efficient manner is key to separating the weak affine relaxation constraints efficiently. Therefore, we focus on solving optimization problem (19). Introducing decision variables $q_{i,t}$ and $u_{j,t}$, respectively, to indicate if resource i and product j are open at time t, problem (19) can be formulated as the integer program

$$\begin{aligned}
\Pi_t^{wAR}(\beta,\gamma) = \max_{q,u} \quad & -\sum_i \gamma_{i,t} q_{i,t} - \sum_j \left[\zeta_{j,t}(\beta,\gamma) + w_j \left(\sum_{i\in\mathcal{I}_j} \gamma_{i,t}\right)\right] u_{j,t} && (20)\\
\text{s.t.} \quad & u_{j,t} - q_{i,t} \le 0 \quad \forall i \in \mathcal{I}_j, \forall j && (21)\\
& q_{i,t} \le 1 \quad \forall i && (22)\\
& u_{j,t} \ge 0, \text{ integer} \quad \forall j. && (23)
\end{aligned}$$

Note that the first constraint ensures that a product is open only if all the resources it uses are open.

Now, observe that the constraint matrix of the above integer program has exactly one $+1$ and one -1 coefficient in each row, and hence is totally unimodular. So we can ignore the integer restriction and solve (20)–(23) exactly as a linear program. In fact, problem (20)–(23) can also be solved combinatorially as a flow problem: the dual of the LP can be transformed to be a network flow problem on a bipartite graph with one set of nodes representing products and the other side resources and edges representing product-resource incidence, and flow from a source to a sink node, each connected to the product and resource nodes respectively; fast algorithms of Ahuja, Orlin, Stein, and Tarjan [1] can then be used to solve the problem in time $O(|\mathcal{I}||E| + \min(|\mathcal{I}|^3, |\mathcal{I}|^2\sqrt{|E|}))$ where $|\mathcal{I}|$ is the number of resources and $|E|$ is the number of edges in this graph. Therefore, problem (20)–(23) can be solved efficiently and separating the wAR constraints is tractable.

We next show that wAR can be formulated as a compact LP eliminating the need for generating constraints on the fly. Since the separation problem can be solved as a LP where all the fixed values (β,γ) appear in the objective function only, we can fold it into the original LP as follows: First take the dual of (20)–(23) with dual variables $\pi_{i,j,t}$ corresponding to (21), and $\psi_{i,t}$ to (22):

$$\begin{aligned}
\Pi_t^{wAR}(\beta,\gamma) = \min_{\pi,\psi} \quad & \sum_i \psi_{i,t}\\
\text{s.t.} \quad & \sum_{i\in\mathcal{I}_j} \pi_{i,j,t} \ge -\left[\zeta_{j,t}(\beta,\gamma) + w_j\left(\sum_{i\in\mathcal{I}_j}\gamma_{i,t}\right)\right] \quad \forall j\\
& -\sum_{j\in\mathcal{J}_i} \pi_{i,j,t} + \psi_{i,t} = -\gamma_{i,t} \quad \forall i\\
& \pi_{i,j,t}, \psi_{i,t} \ge 0 \quad \forall i, j \in \mathcal{J}_i.
\end{aligned}$$

Then use the second constraint in the above LP to eliminate the variable $\psi_{i,t}$ to write the dual as

$$
\begin{aligned}
\Pi_t^{wAR}(\beta,\gamma) = \min_{\pi} \quad & \sum_i \left[\sum_{j\in\mathcal{J}_i} \pi_{i,j,t} - \gamma_{i,t} \right] \\
\text{s.t.} \quad & \sum_{i\in\mathcal{I}_j} \pi_{i,j,t} \geq - \left[\zeta_{j,t}(\beta,\gamma) + w_j \left(\sum_{i\in\mathcal{I}_j} \gamma_{i,t} \right) \right] \quad \forall j && (24) \\
& \sum_{j\in\mathcal{J}_i} \pi_{i,j,t} \geq \gamma_{i,t} \quad \forall i && (25) \\
& \pi_{i,j,t} \geq 0 \quad \forall i, j \in \mathcal{J}_i.
\end{aligned}
$$

Now we fold in the above LP formulation of $\Pi_t^{wAR}(\beta,\gamma)$ into constraints (18) and write wAR equivalently as

$$
\begin{aligned}
V^{wAR} = \min_{\beta,\gamma,\pi} \quad & \sum_t \beta_t + \sum_t \sum_i \gamma_{i,t} r_i^1 \\
\text{s.t.} \quad & \beta_t \geq \sum_i \left[\sum_{j\in\mathcal{J}_i} \pi_{i,j,t} - \gamma_{i,t} \right] \quad \forall t \\
& (24), (25)\, \forall t \\
& \gamma_{i,t}, \pi_{i,j,t} \geq 0 \quad \forall i, j \in \mathcal{J}_i, t.
\end{aligned}
$$

The size of the above LP is polynomial in the number of resources and products. Hence, not only is wAR stronger than $CDLP$, it is also tractable and has a compact formulation. Notice that this formulation would have been hard to derive and justify without the line of reasoning starting from AF.

The dual of the above LP gives more insight into the weak affine relaxation. We get the dual LP as

$$
\begin{aligned}
V^{wAR} = \max_{x,\rho} \quad & \sum_t \sum_j \lambda f_j x_{j,t} \\
(dwAR)\ \text{s.t.} \quad & x_{0,t} + \sum_{s=1}^{t-1} \sum_{j\in\mathcal{J}_i} \lambda x_{j,s} + \sum_{j\in\mathcal{J}_i} x_{j,t} - \rho_{i,t} \leq r_i^1 \quad \forall i, t \\
& x_{0,t} + \sum_j x_{j,t} = 1 \quad \forall t \\
& \frac{x_{j,t}}{w_j} - x_{0,t} + \rho_{i,t} \leq 0 \quad \forall i, j \in \mathcal{J}_i, t \\
& x_{0,t}, x_{j,t}, \rho_{i,t} \geq 0 \quad \forall i, j, t.
\end{aligned}
$$

If we interpret $x_{j,t}$ as the sales rate for product j at time t and $x_{0,t} - \rho_{i,t}$ as the resource level no-purchase rate at time t, then we can view wAR as a refinement of $SBLP$ of Gallego et al. [7], where the sales rates at each time period are modulated by the expected remaining resource capacities.

7 Tighter, tractable relaxations

The weak affine relaxation is based on isolating the difficult term in the affine relaxation and replacing it with a simpler, more tractable term. In this section, we build on this idea and propose two tractable approximation methods that further tighten the wAR bound. We again restrict attention to the single-segment MNL model to reduce notational overhead. In §8, we describe extensions to multi-segment variants of the MNL model.

7.1 Weak affine relaxation$^+$ (wAR^+)

We describe a simple way to tighten the wAR bound, while retaining the compact formulation. Associating decision variables $q_{i,t}$ and $u_{j,t}$, respectively, to indicate if resource i and product j are open, the AF separation problem (14) can be written as

$$\begin{aligned} \Pi_t^{AF}(\beta,\gamma) = \max_{q,u} \quad & -\sum_i \gamma_{i,t} q_{i,t}\left(1+\sum_j w_j u_{j,t}\right) - \sum_j \zeta_{j,t}(\beta,\gamma) u_{j,t} \\ \text{s.t.} \quad & (21), (22), (23). \end{aligned}$$

Now wAR replaces the product term $q_{i,t}u_{j,t}$ for $j \notin \mathcal{J}_i$ in the first summation with 0 and since $q_{i,t}u_{j,t} \geq 0$, we have $\Pi_t^{AF}(\beta,\gamma) \leq \Pi_t^{wAR}(\beta,\gamma)$. Noting that $q_{i,t}u_{j,t} \geq q_{i,t} + u_{j,t} - 1$, we propose replacing the right hand side of constraints (15) with

$$\begin{aligned} \Pi_t^{wAR^+}(\beta,\gamma) = \max_{q,u} \quad & -\sum_i \gamma_{i,t} q_{i,t} - \sum_i \sum_{j \notin i} \gamma_{i,t} w_j \chi_{i,j,t} - \sum_j \left[\zeta_{j,t}(\beta,\gamma) + w_j\left(\sum_{i \in \mathcal{I}_j} \gamma_{i,t}\right)\right] u_{j,t} \\ \text{s.t.} \quad & (21), (22) \\ & \chi_{i,j,t} \geq q_{i,t} + u_{j,t} - 1 \quad \forall i, j \notin i \\ & u_{j,t}, \chi_{i,j,t} \geq 0 \quad \forall j, i \notin \mathcal{I}_j. \end{aligned}$$

The following lemma is immediate.

Lemma 2. $\Pi_t^{AF}(\beta,\gamma) \leq \Pi_t^{wAR^+}(\beta,\gamma) \leq \Pi_t^{wAR}(\beta,\gamma)$.

Therefore, we replace the right hand side of constraints (15) with $\Pi_t^{wAR^+}(\beta,\gamma)$ and solve the LP

$$\begin{aligned} V^{wAR^+} = \min_{\beta,\gamma} \quad & \sum_t \beta_t + \sum_t \sum_i \gamma_{i,t} r_i^1 \\ (wAR^+) \text{ s.t.} \quad & \beta_t \geq \Pi_t^{wAR^+}(\beta,\gamma) \quad \forall t \\ & \gamma_{i,t} \geq 0 \quad \forall i, t. \end{aligned} \tag{26}$$

We refer to this method as weak affine relaxation$^+$ (wAR^+). Lemma 2 together with Lemma 1 implies that $V^{AF} \leq V^{wAR^+} \leq V^{wAR}$. Therefore, wAR^+ further tightens the wAR bound. Note however that the wAR^+ separation problem can have as many as $|\mathcal{I}||\mathcal{J}|$ additional constraints compared to wAR. Still, the wAR^+ separation problem involves solving a linear program and hence is tractable. Moreover, it is possible to obtain a compact formulation of wAR^+ by following the steps in §6.3; we omit the details.

7.2 A hierarchical family of relaxations

In this section we show how to construct a hierarchical family of relaxations that at the highest level (level-n, the number of products) gives us the affine relaxation. Naturally, because of the NP-hardness of solving the affine relaxation, we cannot expect tractability, and so we concentrate on small levels. The level-1 relaxation already turns out to be a tighter relaxation than wAR. While the level-1 relaxation separation problem can be solved in a tractable manner, a potential drawback is that, unlike wAR and wAR^+, it cannot be folded into the original problem to yield a compact formulation.

For simplicity we describe the level-1 formulation and remark on how it extends to a hierarchy of relaxations. In the level-1 relaxation, which we refer to as hierarchical affine relaxation (hAR), we replace the $\gamma_{S,t}(1+\sum_{j\in S} w_j)$ term in (14) with $\gamma_{S,t} + (\sum_{j\in S} w_j)(\max_{j'\in S}\sum_{i\in\mathcal{I}_{j'}}\gamma_{i,t})$ and solve the LP

$$\begin{aligned} V^{hAR} = \min_{\beta,\gamma} \quad & \sum_t \beta_t + \sum_t \sum_i \gamma_{i,t} r_i^1 \\ (hAR)\ \text{s.t.} \quad & \beta_t \geq \Pi_t^{hAR}(\beta,\gamma) \quad \forall t \\ & \gamma_{i,t} \geq 0 \quad \forall i,t, \end{aligned} \tag{27}$$

where

$$\Pi_t^{hAR} = \max_S \left\{ -\gamma_{S,t} - \left(\sum_{j\in S} w_j\right)\left(\max_{j'\in S}\sum_{i\in\mathcal{I}_{j'}}\gamma_{i,t}\right) - \sum_{j\in S}\zeta_{j,t}(\beta,\gamma)\right\}. \tag{28}$$

We have the following lemma.

Lemma 3. $\Pi_t^{AF}(\beta,\gamma) \leq \Pi_t^{hAR}(\beta,\gamma) \leq \Pi_t^{wAR}(\beta,\gamma)$.

Proof

By definition, we have $\mathcal{I}_j \subseteq \mathcal{I}_S$ for all $j \in S$. Therefore, $\gamma_{S,t} = \sum_{i\in\mathcal{I}_S}\gamma_{i,t} \geq \sum_{i\in\mathcal{I}_j}\gamma_{i,t}$ for all $j\in S$ and so $\gamma_{S,t} \geq \max_{j\in S}\sum_{i\in\mathcal{I}_j}\gamma_{i,t}$. The proof now follows by noting that $\gamma_{S,t}\left(1+\sum_{j\in S}w_j\right) \geq \gamma_{S,t} + \left(\sum_{j\in S}w_j\right)\left(\max_{j'\in S}\sum_{i\in\mathcal{I}_{j'}}\gamma_{i,t}\right) \geq \gamma_{S,t} + \sum_{j\in S} w_j\left(\sum_{i\in\mathcal{I}_j}\gamma_{i,t}\right)$.

□

Lemma 3 together with Lemma 1 implies that $V^{AF} \leq V^{hAR} \leq V^{wAR}$. Therefore, hAR obtains a tighter bound than wAR.

Next, we show that hAR separation problem (28) can be solved in a tractable manner. Associating binary decision variables $q_{i,t}$ and $u_{j,t}$, respectively, to indicate if resource i and product j are open, problem (28) can be written as

$$\begin{aligned} \Pi_t^{hAR}(\beta,\gamma) = \max_{q,u} \quad & -\sum_i \gamma_{i,t} q_{i,t} - \left(\sum_j w_j u_{j,t}\right)\left(\max_{j'}\sum_{i\in\mathcal{I}_{j'}}\gamma_{i,t}u_{j',t}\right) - \sum_j \zeta_{j,t}(\beta,\gamma)u_{j,t} \\ \text{s.t.} \quad & (21)-(23). \end{aligned}$$

Although the above optimization problem has a nonlinear objective function, we can solve it through a sequence of linear programs in the following manner. We fix a product $\hat{j}$ as the one achieving the maximum value of $\max_{j'} \gamma_{i,t} u_{j',t}$. Since $\hat{j}$ achieves the maximum value, we must have $u_{\hat{j},t} = 1$ and $u_{j,t} = 0$ for j with $\sum_{i \in \mathcal{I}_j} \gamma_{i,t} > \sum_{i \in \mathcal{I}_{\hat{j}}} \gamma_{i,t}$. Letting $\hat{\mathcal{J}}_{\hat{j}} = \left\{ j \mid \sum_{i \in \mathcal{I}_j} \gamma_{i,t} > \sum_{i \in \mathcal{I}_{\hat{j}}} \gamma_{i,t} \right\}$, we solve the following linear integer program for product $\hat{j}$:

$$\begin{aligned} \Pi_t^{hAR,\hat{j}}(\beta,\gamma) = \max_{q,u} \quad & -\sum_i \gamma_{i,t} q_{i,t} - \sum_j \left[\zeta_{j,t}(\beta,\gamma) + w_j \left(\sum_{i \in \mathcal{I}_{\hat{j}}} \gamma_{i,t} \right) \right] u_{j,t} \\ \text{s.t.} \quad & (21), (22) \\ & u_{\hat{j},t} = 1 \\ & u_{j,t} = 0 \quad \forall j \in \hat{\mathcal{J}}_{\hat{j}} \\ & u_{j,t} \geq 0 \text{ integer} \quad \forall j \in \mathcal{J} \backslash \hat{\mathcal{J}}_{\hat{j}}. \end{aligned}$$

Since the constraint matrix is totally unimodular, we can solve the above linear integer program equivalently as a linear program. So we solve the linear program for each product $\hat{j} \in \mathcal{J}$ and obtain $\Pi_t^{hAR}(\beta,\gamma) = \max_{\hat{j} \in \mathcal{J}} \Pi_t^{hAR,\hat{j}}(\beta,\gamma)$.

Since problem (28) can be solved in a tractable manner, separating the hAR constraints is tractable, and hAR can be solved in polynomial time by the ellipsoid method. However, unlike wAR^+, hAR does not seem to have a compact linear programming formulation. This is because the set $\hat{\mathcal{J}}_{\hat{j}}$ depends on the values of the γ's in a nonlinear fashion and the duality argument in §6.3 that we used to fold the separation problem back into the original LP does not hold. On the other hand, an appealing feature of hAR is that its separation problem has fewer number of decision variables and constraints than wAR^+.

Remark: One can get further relaxations by considering pairs of elements j', j'' for a level-2 relaxation (or triples for level-3, and so on) such that we find the offer set S that maximizes

$$-\left(1 + \sum_{j \in S} w_j\right) \left[\max_{\{j', j'' \in S\}} \sum_{i \in \mathcal{I}_{\{j', j''\}}} \gamma_{i,t} \right].$$

In this way, we can control the degree of approximation to the affine relaxation. We limit our numerical results to fixing a single element j'.

8 Extensions

In this section we describe how to extend the weak affine relaxation of §6 to variants of the MNL model (the development for wAR^+ and hAR is similar). In §8.1 we consider the MNL choice model with multiple customer segments and disjoint consideration sets. In §8.2 we consider the case where the consideration sets of the different segments may overlap. This model, also referred to as the mixture of multinomial logits (MMNL), is a rich choice model that can approximate any random utility choice model arbitrarily closely; McFadden and Train [12]. It is also possible to extend the weak affine relaxation idea to the general attraction model of Gallego et al. [7] in a transparent manner. Kunnumkal and Talluri [9] show how the same ideas can be extended to the nested-logit choice model.

8.1 Multiple segments with disjoint consideration sets

Now we consider the case where the total demand is comprised of demand from multiple customer segments. The consideration sets of the different segments are disjoint and so we have $\mathcal{C}_l \cap \mathcal{C}_m = \emptyset$ for segments l and m. We note that the case of disjoint consideration sets for the segments is one of the few known cases where the $CDLP$ formulation is tractable. We describe below how wAR can be extended to tighten the $CDLP$ bound in a tractable manner. The key idea is to look at the AF separation problem for each customer segment, which again turns out to be intractable. We apply the ideas from the single-segment case to get a tractable relaxation.

Let $\mathcal{I}_l = \{i \in \mathcal{I} \,|\, \exists j \in \mathcal{C}_l \text{ and } j \in \mathcal{J}_i\}$ and $\mathcal{L}_i = \{l \in \mathcal{L} \,|\, i \in \mathcal{I}_l\}$. We can interpret $\mathcal{I}_l$ as the set of resources that are used by segment l and $\mathcal{L}_i$ as the set of segments that use resource i.

Now consider the separation problem for AF. Using $\lambda Q_i(S) = \sum_l \lambda_l Q_i^l(S_l)$ and $\lambda R(S) = \sum_l \lambda_l R^l(S_l)$, where $S_l = S \cap \mathcal{C}_l$, constraint (12) can be written as

$$\beta_t + \sum_i \mathbb{1}_{[i\in\mathcal{I}_S]}\gamma_{i,t} + \sum_i \left[\left(\sum_{k=t+1}^{\tau} \gamma_{i,k} \right) \sum_l \lambda_l Q_i^l(S) \right] \geq \sum_l \lambda_l R^l(S). \tag{29}$$

We first split this constraint into l separate constraints, one for each segment, by introducing variables $\beta_{l,t}$. The constraint for segment l at time t is that

$$\beta_{l,t} + \sum_{i\in\mathcal{I}_l} \mathbb{1}_{\left[i\in\mathcal{I}_{S_l}\right]}\gamma_{i,t}\frac{\lambda_l}{\sum_{l'\in\mathcal{L}_i}\lambda_{l'}} + \sum_i \left[\left(\sum_{k=t+1}^{\tau} \gamma_{i,k} \right) \lambda_l Q_i^l(S_l) \right] \geq \lambda_l R^l(S_l) \tag{30}$$

for each $S_l = S \cap \mathcal{C}_l$. The proof of Proposition 4 below shows that the segment level constraints (30) imply (29) and that we obtain a looser upper bound by separating over (30) instead of (29).

We observe that the segment level constraints (30) have the same form as constraints (12) in the single-segment case, and are therefore hard to separate. So we use the same relaxation as we did for the single-segment case to obtain a tractable separation problem at the segment level:

$$\begin{aligned}
\Pi_{l,t}^{swAR}(\beta,\gamma) = \max_{q,u} \quad & -\sum_{i\in\mathcal{I}_l} \frac{\lambda_l \gamma_{i,t}}{\sum_{l'\in\mathcal{L}_i}\lambda_{l'}} q_{i,t} \\
& - \sum_{j\in\mathcal{C}_l} w_j^l \left[\beta_{l,t} + \lambda_l \left(\left(\sum_{i\in\mathcal{I}_j} \sum_{k=t+1}^{\tau} \gamma_{i,k} \right) - f_j + \sum_{i\in\mathcal{I}_j} \frac{\gamma_{i,t}}{\sum_{l'\in\mathcal{L}_i}\lambda_{l'}} \right) \right] u_{j,t} \\
\text{s.t.} \quad & (21) - (23).
\end{aligned}$$

We replace constraint (30) with $\beta_{l,t} \geq \Pi_{l,t}^{swAR}(\beta,\gamma)$ to obtain a segment-based weak affine relaxation ($swAR$):

$$\begin{aligned}
V^{swAR} = \min_{\beta,\gamma} \quad & \sum_t \sum_l \beta_{l,t} + \sum_t \sum_i \gamma_{i,t} r_i^1 \\
\text{s.t.} \quad & \beta_{l,t} \geq \Pi_{l,t}^{swAR}(\beta,\gamma) \quad \forall l,t \\
& \gamma_{i,t} \geq 0 \quad \forall i,t.
\end{aligned}$$

Moreover, by following the same steps as for the single-segment case, it is possible to show that

$swAR$ can be formulated as the compact LP

$$
\begin{aligned}
V^{swAR} = \min_{\gamma,\beta,\pi} \quad & \sum_t \sum_l \beta_{l,t} + \sum_i \sum_t \gamma_{i,t} r_i^1 \\
(swAR) \text{ s.t.} \quad & \beta_{l,t} \geq \sum_{i \in \mathcal{I}_l} \left[\sum_{j \in \mathcal{J}_i, j \in \mathcal{C}_l} \pi_{i,j,t} - \frac{\lambda_l}{\sum_{l' \in \mathcal{L}_i} \lambda_{l'}} \gamma_{i,t} \right] \quad \forall l, t \\
& \sum_{i \in \mathcal{I}_j} \pi_{i,j,t} \geq \lambda_{\ell_j} w_j^{\ell_j} \left[f_j - \sum_{i \in \mathcal{I}_j} \left(\sum_{k=t+1}^{\tau} \gamma_{i,k} + \frac{\gamma_{i,t}}{\sum_{l' \in \mathcal{L}_i} \lambda_{l'}} \right) - \frac{\beta_{\ell_j,t}}{\lambda_{\ell_j}} \right] \quad \forall j, t \\
& \sum_{j \in \mathcal{J}_i, j \in \mathcal{C}_l} \pi_{i,j,t} - \frac{\lambda_l}{\sum_{l' \in \mathcal{L}_i} \lambda_{l'}} \gamma_{i,t} \geq 0 \quad \forall i, l \in \mathcal{L}_i, t \\
& \gamma_{i,t}, \pi_{i,j,t} \geq 0 \quad \forall, i, j \in \mathcal{J}_i, t,
\end{aligned}
$$

where ℓ_j denotes the segment to which product j belongs. $swAR$ can be viewed as an extension of wAR to the MNL model with multiple segments and disjoint consideration sets. Note that $swAR$ is again tractable as it is a compact LP. Proposition 4 below shows that it also obtains an upper bound on the value function that is tighter than $CDLP$.

Proposition 4. $V^{AF} \leq V^{swAR} \leq V^{CDLP}$.

Proof
Using the MNL choice probability (1) and rearranging terms, the $swAR$ constraint $\beta_{l,t} \geq \Pi_{l,t}^{swAR}(\beta, \gamma)$ can be equivalently written as

$$
\beta_{l,t} \geq \lambda_l \left[R^l(S_l) - \sum_{i \in \mathcal{I}_l} \sum_{k=t+1}^{\tau} Q_i^l(S_l) \gamma_{i,k} \right] - \sum_{i \in \mathcal{I}_l} \mathbb{1}_{[i \in \mathcal{I}_{S_l}]} \gamma_{i,t} \frac{\lambda_l}{\sum_{l' \in \mathcal{L}_i} \lambda_{l'}} \left(\sum_{j \in \mathcal{J}_i} P_j^l(S_l) + P_0^l(S_l) \right) \tag{31}
$$

for all $S_l \subseteq \mathcal{C}_l$.

Consider now two intermediate problems:

$$
\begin{aligned}
\underline{V} = \min_{\beta,\gamma} \quad & \sum_t \sum_l \beta_{l,t} + \sum_t \sum_i \gamma_{i,t} r_i^1 \\
\text{s.t.} \quad & (30) \quad \forall l, S_l \subseteq \mathcal{C}_l, t \\
& \gamma_{i,t} \geq 0 \quad \forall i, t,
\end{aligned}
$$

and

$$
\begin{aligned}
\bar{V} = \min_{\beta,\gamma} \quad & \sum_t \sum_l \beta_{l,t} + \sum_t \sum_i \gamma_{i,t} r_i^1 \\
\text{s.t.} \quad & \beta_{l,t} \geq \lambda_l \left[R^l(S_l) - \sum_{i \in \mathcal{I}_l} \sum_{k=t}^{\tau} Q_i^l(S_l) \gamma_{i,k} \right] \quad \forall l, S_l \subseteq \mathcal{C}_l, t \qquad (32) \\
& \gamma_{i,t} \geq 0 \quad \forall i, t.
\end{aligned}
$$

We can interpret the first problem as a segment based relaxation of AF, while the second problem can be viewed as a segment based relaxation of $CDLP$.

We next show that $V^{AF} \leq \underline{V} \leq V^{swAR} \leq \bar{V} = V^{CDLP}$, which completes the proof of the proposition.

(i) $\underline{V} \le V^{swAR} \le \bar{V}$.

Since the objective functions of all the problems are the same, we only need to compare the corresponding constraints. Since $\sum_{j\in\mathcal{J}_i} P_j^l(S_l) + P_0^l(S_l) \le 1$, it follows that constraint (31) implies constraint (30) and we have $\underline{V} \le V^{swAR}$.

On the other hand, the right hand side of constraint (32) can be written as

$$\lambda_l \left[R^l(S_l) - \sum_{i\in\mathcal{I}_l} \sum_{k=t+1}^{\tau} Q_i^l(S_l)\gamma_{i,k} \right] - \sum_{i\in\mathcal{I}_l} \lambda_l Q_i^l(S_l)\gamma_{i,t}.$$

Now note that

$$\begin{aligned}
\lambda_l Q_i^l(S_l)\gamma_{i,t} &= \lambda_l \mathbb{1}_{[i\in\mathcal{I}_{S_l}]} Q_i^l(S_l)\gamma_{i,t} = \lambda_l \mathbb{1}_{[i\in\mathcal{I}_{S_l}]} \left[\sum_{j\in\mathcal{J}_i} P_j^l(S_l) \right] \gamma_{i,t} \\
&\le \frac{\lambda_l}{\sum_{l'\in\mathcal{L}_i} \lambda_{l'}} \mathbb{1}_{[i\in\mathcal{I}_{S_l}]} \left[\sum_{j\in\mathcal{J}_i} P_j^l(S_l) \right] \gamma_{i,t} \le \frac{\lambda_l}{\sum_{l'\in\mathcal{L}_i} \lambda_{l'}} \mathbb{1}_{[i\in\mathcal{I}_{S_l}]} \left[\sum_{j\in\mathcal{J}_i} P_j^l(S_l) + P_0^l(S_l) \right] \gamma_{i,t}
\end{aligned}$$

where the first equality holds since if $\mathbb{1}_{[i\in\mathcal{I}_{S_l}]} = 0$, then $Q_i^l(S_l) = 0$ and the first inequality holds since $\sum_{l'\in\mathcal{L}_i} \lambda_{l'} \le 1$. Therefore constraint (32) implies constraint (31) and we have $V^{swAR} \le \bar{V}$.

(ii) $V^{AF} \le \underline{V}$.

Suppose that $(\hat{\beta}, \hat{\gamma})$ satisfies constraints (30). We show that it satisfies constraints (29) as well. Fix a set S and let $S_l = S \cap \mathcal{C}_l$. Adding up constraints (30) for all the segments

$$\begin{aligned}
\sum_l \hat{\beta}_{l,t} &\ge \sum_l \left\{ \lambda_l \left[R^l(S_l) - \sum_{i\in\mathcal{I}_l} \sum_{k=t+1}^{\tau} Q_i^l(S_l)\hat{\gamma}_{i,k} \right] - \sum_{i\in\mathcal{I}_l} \mathbb{1}_{[i\in\mathcal{I}_{S_l}]} \hat{\gamma}_{i,t} \frac{\lambda_l}{\sum_{l'\in\mathcal{L}_i} \lambda_l'} \right\} \\
&= \lambda \left[R(S) - \sum_i \sum_{k=t+1}^{\tau} Q_i(S)\hat{\gamma}_{i,k} \right] - \sum_i \hat{\gamma}_{i,t} \sum_{l\in\mathcal{L}_i} \mathbb{1}_{[i\in\mathcal{I}_{S_l}]} \frac{\lambda_l}{\sum_{l'\in\mathcal{L}_i} \lambda_l'} \\
&\ge \lambda \left[R(S) - \sum_i \sum_{k=t+1}^{\tau} Q_i(S)\hat{\gamma}_{i,k} \right] - \sum_i \hat{\gamma}_{i,t} \sum_{l\in\mathcal{L}_i} \mathbb{1}_{[i\in\mathcal{I}_S]} \frac{\lambda_l}{\sum_{l'\in\mathcal{L}_i} \lambda_l'} \\
&= \lambda \left[R(S) - \sum_i \sum_{k=t+1}^{\tau} Q_i(S)\hat{\gamma}_{i,k} \right] - \sum_i \hat{\gamma}_{i,t} \mathbb{1}_{[i\in\mathcal{I}_S]},
\end{aligned}$$

where the first equality uses the fact that $Q_i^l(S_l) = 0$ for $l \notin \mathcal{L}_i$ and hence $\lambda Q_i(S) = \sum_l \lambda_l Q_i^l(S_l) = \sum_{l\in\mathcal{L}_i} \lambda_l Q_i^l(S_l)$. The second inequality holds since $\mathbb{1}_{[i\in\mathcal{I}_{S_l}]} \le \mathbb{1}_{[i\in\mathcal{I}_S]}$. Letting $\tilde{\beta} = \{\tilde{\beta}_t = \sum_l \hat{\beta}_{l,t} \,|\, \forall t\}$, it follows that $(\tilde{\beta}, \hat{\gamma})$ satisfies constraints (29). Therefore $V^{AF} \le \sum_t \tilde{\beta}_t + \sum_t \sum_i \hat{\gamma}_{i,t} = \underline{V}$.

Meissner et al. [14] prove the following that we include for completeness.

(iii) $\bar{V} = V^{CDLP}$. (Meissner et al. [14])

Constraints (6) in $dCDLP$ are equivalent to

$$\begin{aligned}\beta_t &= \max_S \left\{ \lambda \left[R(S) - \sum_i \sum_{k=t}^{\tau} Q_i(S)\gamma_{i,k} \right] \right\} \\ &= \max_S \left\{ \sum_l \lambda_l \left[R^l(S \cap \mathcal{C}_l) - \sum_{i \in \mathcal{I}_l} \sum_{k=t}^{\tau} Q_i^l(S \cap \mathcal{C}_l)\gamma_{i,k} \right] \right\} \\ &= \sum_l \max_{S_l} \left\{ \lambda_l \left[R^l(S_l) - \sum_{i \in \mathcal{I}_l} \sum_{k=t}^{\tau} Q_i^l(S_l)\gamma_{i,k} \right] \right\}\end{aligned}$$

where the last inequality uses the fact that the consideration sets are disjoint. Therefore, the $dCDLP$ constraint is equivalent to the constraints $\beta_t = \sum_l \beta_{l,t}$ and

$$\beta_{l,t} = \max_{S_l} \left\{ \lambda_l \left[R^l(S_l) - \sum_{i \in \mathcal{I}_l} \sum_{k=t}^{\tau} Q_i^l(S_l)\gamma_{i,k} \right] \right\},$$

which is exactly constraint (32).

□

As we show in the next section, it is possible to extend the $swAR$ formulation to the MNL model with multiple segments when the consideration sets overlap. The dual of $swAR$, which we give below, turns out to be useful for this purpose.

$$V^{dswAR} = \max_{x,\rho} \quad \sum_t \sum_l \sum_{j \in \mathcal{C}_l} \lambda_l f_j x_{j,t}^l$$

$$(dswAR) \text{ s.t.} \quad \sum_{l \in \mathcal{L}_i} \lambda_l \left[\frac{x_{0,t}^l}{\sum_{l' \in \mathcal{L}_i} \lambda_{l'}} + \sum_{s=1}^{t-1} \sum_{j \in \mathcal{J}_i \cap \mathcal{C}_l} x_{j,s}^l + \frac{\sum_{j \in \mathcal{J}_i \cap \mathcal{C}_l} x_{j,t}^l}{\sum_{l' \in \mathcal{L}_i} \lambda_{l'}} - \frac{\rho_{it}^l}{\sum_{l' \in \mathcal{L}_i} \lambda_{l'}} \right] \leq r_i^1 \quad \forall i, t \tag{33}$$

$$x_{0,t}^l + \sum_{j \in \mathcal{C}_l} x_{j,t}^l = 1 \quad \forall l, t$$

$$\frac{x_{j,t}^l}{w_j^l} - x_{0,t}^l + \rho_{i,t}^l \leq 0 \quad \forall l, i, j \in \mathcal{J}_i \cap \mathcal{C}_l, t \tag{34}$$

$$x_{0,t}^l, x_{j,t}^l, \rho_{i,t}^l \geq 0 \quad \forall l, i, j \in \mathcal{J}_i \cap \mathcal{C}_l, t.$$

8.2 Multiple segments with overlapping consideration sets

When the segment consideration sets overlap, the $CDLP$ formulation is difficult to solve, even for MNL with just two segments. So one would imagine that it is difficult to find a tractable bound tighter than $CDLP$ in this case. One strategy, pursued in Meissner et al. [14] is to formulate the problem by segments and then add a set of consistency conditions called *product-cut* equalities (PC-equalities). These equalities apply to any general discrete-choice model and appear to be quite powerful in numerical experiments, often bringing the solution close to $CDLP$ value. Strauss and

Talluri [17] subsequently show that when the consideration set structure has a certain tree structure, the cuts in fact achieve the $CDLP$ value. Talluri [18] shows how to specialize the PC-equalities to the MNL choice model. In this section we describe how the PC-equalities, specialized for MNL, can be added to $dswAR$ to tighten the formulation.

We begin with a brief description of the PC-equalities: Meissner et al. [14] allow different sets to be offered to different segments. However, to ensure consistency, they require that for any product $j \in \mathcal{C}_l \cap \mathcal{C}_m$, the length of time it is offered to segment l must be the equal to the length of time it is offered to segment m. This leads to a set of consistency constraints which they term as PC-equalities. Talluri [18] uses choice probabilities (1) to specialize the PC-equalities to the MNL model as:

$$\frac{x^l_{j,t}}{w^l_j} = \sum_{\{S \subseteq (\mathcal{C}_l \cap \mathcal{C}_m) \,|\, j \in S\}} y^{l,m}_S \quad \forall l, m, j \in \mathcal{C}_l \cap \mathcal{C}_m \tag{35}$$

$$y^{l,m}_{S,j} \leq y^{l,m}_S \quad \forall l, m, S \subseteq \mathcal{C}_l \cap \mathcal{C}_m, j \in \mathcal{C}_l \setminus \mathcal{C}_m \tag{36}$$

$$\begin{aligned} &\sum_{\{T \subseteq (\mathcal{C}_l \cap \mathcal{C}_m) \,|\, T \supseteq S\}} \left\{ \sum_{j \in \mathcal{C}_l \setminus \mathcal{C}_m} w^l_j y^{l,m}_{T,j} + (1 + W^l_T) y^{l,m}_T \right\} = \\ &\sum_{\{T' \subseteq (\mathcal{C}_m \cap \mathcal{C}_l) \,|\, T' \supseteq S\}} \left\{ \sum_{j \in \mathcal{C}_m \setminus \mathcal{C}_l} w^m_j y^{m,l}_{T',k} + (1 + W^m_{T'}) y^{m,l}_{T'} \right\} \forall l, m, S \subseteq \mathcal{C}_l \cap \mathcal{C}_m, \end{aligned} \tag{37}$$

where $W^l_S = \sum_{j \in S} w^l_j$ and we have new variables of the form $y^{l,m}_S$ defined for all pairs of segments l, m and for all $S \subseteq \mathcal{C}_l \cap \mathcal{C}_m$; see Talluri [18]. If the overlap in the consideration sets of the different segments is not too large, then the number of PC-equalities is manageable.

Talluri [18] shows that adding PC-equalities (35)-(37) to the sales-based linear program ($SBLP$) of Gallego et al. [7] further tightens the $SBLP$ bound. We are also able to tighten the $dswAR$ bound by doing the same thing. Moreover, comparing $dswAR$ with $SBLP$, it is easy to see that a feasible solution to $dswAR$ is also feasible to $SBLP$. Therefore, $dswAR$ is tighter than $SBLP$. It follows that $dswAR$ augmented with the PC-equalities, continues to be tighter than $SBLP$ with the same PC-equalities. So in conclusion, when segment consideration sets overlap, we also have

Proposition 5. *The objective function value of $dswAR$ with (35–37) added, is less than or equal to the objective function value of $SBLP$ with (35–37) added.*

In closing, we note that $dswAR$ augmented with the PC-equalities is not guaranteed to be tighter than $CDLP$. We numerically compare the performance of $dswAR$ with $CDLP$ in our computational experiments that we present next.

9 Computational experiments

In this section, we compare the upper bounds and the revenue performance of the policies obtained by the different benchmark solution methods. We test the performance of our benchmark solution methods on a hub-and-spoke network, with a single hub serving multiple spokes. While comparing the revenue performance of the benchmark methods, we divide the booking period into five equal intervals. At the beginning of each interval, we re-solve the benchmark solution methods to get fresh estimates for the marginal value of capacity on the resources. All of the benchmark methods give a solution of the form $(\hat{\beta}, \hat{\gamma})$ with $\sum_{s=t}^{\tau} \hat{\gamma}_{i,s}$ being an estimate for the marginal value of capacity on resource i at time t. We use these marginal values to construct a value function approximation $\hat{V}_t(\boldsymbol{r}) = \sum_i (\sum_{s=t}^{\tau} \hat{\gamma}_{i,s}) r_i$ and solve problem (8) to decide on the offer set. We continue to use this decision rule until the beginning of the next interval where we re-solve the benchmark solution methods. In all of our test problems, we have multiple customer segments and choice within each segment is governed by the MNL model. We begin by describing the different benchmark solution methods and the experimental setup.

9.1 Benchmark solution methods

Choice deterministic linear program (CDLP) This is the solution method described in §4.1.

Weak affine relaxation (wAR) This is the version of weak affine relaxation that applies to multiple segments and described in §8 ($swAR$).

Weak affine relaxation$^+$ (wAR$^+$) This is the version of wAR^+ that applies to multiple segments. As mentioned, it is possible to extend the wAR^+ formulation described in §7.1 to the setting with multiple segments by following the steps in §8.

Hierarchical affine relaxation (hAR) This is the version of the level-1 hierarchical affine relaxation that applies to multiple segments. As mentioned, it is possible to extend the hierarchical affine relaxation method described in §7.2 to the setting with multiple segments by following the steps in §8. Since hAR does not admit a compact formulation, we solve hAR by generating constraints on the fly and stop when we are within 1% of optimality.

Affine relaxation (AF) This is the solution method described in §4.2. We work with the reduced formulation RAF of Vossen and Zhang [22]. While the number of decision variables in RAF is manageable, it has a large number of constraints. We solve RAF by generating constraints on the fly (using integer programming) and stop when we are within 1% of optimality.

9.2 Hub-and-spoke network

We consider a hub-and-spoke network with a single hub that serves N spokes. Half of the spokes have two flights to the hub, while the remaining half have two flights from the hub so that the total number of flights is $2N$. Figure 1 shows the structure of the network with $N = 8$. We note that the flight legs correspond to the resources in our NRM formulation.

The total number of fare-products is $2N(N+2)$. There are $4N$ fare products connecting spoke-to-hub and hub-to-spoke origin-destination pairs, of which half are high fare-products and the remaining half are low-fare products. The high fare-product is 50% more expensive than the corresponding low fare-product. The remaining $4N^2$ fare-products connect spoke-to-spoke origin-destination pairs. Half of the $4N^2$ fare-products are high fare-products and the rest are low fare-products, with the high fare-product being 50% more expensive than the corresponding low fare-product.

Each origin-destination pair is associated with a customer segment and each segment is only interested in the fare-products connecting that origin-destination pair. Therefore, the consideration sets are disjoint. Within each segment choice is governed by the MNL model. We sample the preference weights of the fare-products from a poisson distribution with a mean of 100 and set the no-purchase preference weight to be $0.5 \sum_{j \in \mathcal{C}_l} w_j^l$. So the probability that a customer does not purchase anything when all the products in the consideration set are offered is around 33%.

We measure the tightness of the leg capacities using the nominal load factor, which is defined in the following manner. Letting $\hat{S}_t = \text{ argmax }_S R(S)$ denote the optimal set of products offered at time period t when there is ample capacity on all flight legs, we define the nominal load factor

$$\alpha = \frac{\sum_t \sum_i \lambda_t Q_i(\hat{S}_t)}{\sum_i r_i^1},$$

where λ_t denotes the arrival rate at time period t. Initially, we assume stationary arrivals and set the arrival rate to be 0.9 in each time period. We have $\tau = 200$ in all of our test problems. We label our test problems by (N, α) where $N \in \{4, 6, 8\}$ and $\alpha \in \{0.8, 1.0, 1.2, 1.6\}$, which gives us 24 test problems in total.

Table 1 gives the upper bounds obtained by the benchmark solution methods. The first column in the table gives the problem characteristics. The second to sixth columns, respectively, give the upper bounds obtained by $CDLP$, wAR, wAR^+, hAR, and AF. The last four columns give the percentage gap between the upper bounds obtained by $CDLP$ and wAR, $CDLP$ and wAR^+, $CDLP$ and hAR, and $CDLP$ and AF, respectively. AF generates the tightest upper bound and $CDLP$ the weakest, with the remaining upper bounds sandwiched in between. The average percentage gap between wAR and $CDLP$ is 1.59%, although we observe instances where the gap is as high as 2.73%. The percentage gap between wAR and $CDLP$ seems to increase with the nominal load factor and the number of spokes in the network. wAR^+ and hAR tighten the wAR bound and obtain bounds that are on average 1.81% and 1.63% tighter than $CDLP$. AF obtains bounds that are on average 2.16% tighter than $CDLP$.

Table 2 gives the CPU seconds required by the different solution methods for different numbers of spokes in the network and different numbers of time periods in the booking horizon. All of our computational experiments are carried out on a Pentium Core 2 Duo desktop with 3-GHz CPU and 4-GB RAM. We use CPLEX 11.2 to solve all linear programs. Since we have disjoint consideration sets, $CDLP$ has a compact sales-based formulation, $SBLP$, which can be solved in a matter of seconds. The solution times of the other methods are generally in minutes. wAR typically runs faster than AF and the savings can be significant especially for relatively large networks. In light of the hardness result in Proposition 2, we only expect the savings in run times to increase with the problem size. wAR^+ and hAR have additional computational overheads associated with them and can take longer than wAR.

If we compare the improvement in the upper bound relative to what AF achieves over $CDLP$ with the corresponding increase in solution time (relative to the overhead incurred by AF over $CDLP$), we find that wAR closes around 70% of the gap between the AF and $CDLP$ bounds by incurring a computational overhead of around 15% of that of AF. In contrast, wAR^+ closes around 80% of the gap by incurring a 45% computational overhead, while hAR closes around 75% of the gap by incurring roughly a 22% computational overhead. Overall, wAR seems to achieve a good balance between the quality of the solution and the computational effort.

Table 3 gives the expected revenues obtained by the different benchmark methods. The columns have a similar interpretation as in Table 1 except that they give the expected total revenues. We evaluate the revenue performance by simulation and use common random numbers in our simulations. In the last four columns, we use ✓ to indicate that the corresponding benchmark method generates higher revenues than $CDLP$ at the 95% level, an $\odot$ if the difference in the revenue performance of the benchmark method and $CDLP$ is not significant at the 95% level and a $\times$ if the benchmark method generates lower revenues than $CDLP$ at the 95% level. wAR on average generates revenues that are 2.28% higher than $CDLP$, although we observe instances where the gap is as high as 7%. wAR^+, hAR and AF, respectively, generate revenues that are on average 3.02%, 2.94% and 2.34%, higher than the $CDLP$ revenues.

It is interesting to note that the improvements in the revenue performance of wAR (over $CDLP$) tend to be larger than the corresponding improvements in the upper bound. We understand this as follows. Recall that wAR, $CDLP$ and AF all yield a solution of the form $(\hat{\beta}, \hat{\gamma})$ with $\sum_{s=t}^{\tau} \hat{\gamma}_{i,s}$ being an estimate of the marginal value of capacity of resource i at time t. Figure 2 shows a representative plot of how the marginal values of capacity obtained by wAR, $CDLP$ and AF vary with time. We note that the wAR and AF marginal values change with time, while $CDLP$ yields static marginal values. We observe that wAR does a much better job of tracking the AF marginal values compared to $CDLP$. This implies that wAR obtains a sharper value function approximation (7) and consequently is able to make better decisions on the set of products to offer by solving problem (8). Figure 2 also reveals that the marginal values of capacity obtained by AF are constant for the most part and start varying only towards the end of the sales horizon. This suggests that much of the benefits of wAR better tracking the AF marginal values are likely to be accrued in a short time window before the end of the sales horizon. Below, we describe an additional set of computational experiments that further supports these observations.

Fluid scaling: We consider a fluid scaling of the hub-and-spoke test problems where we scale the flight leg capacities and the length of the sales horizon by a factor $0 < \theta \leq 1$. That is, in the θ-scaled problem, the initial capacities given by $\theta \mathbf{r}^1$ and the length of the selling season is $\theta\tau$. Note that if we set $\theta = 1$, then we get back the original set of test problems, and smaller values of θ correspond to test problems with lower initial capacities and shorter sales horizons. Table 4 compares the upper bounds obtained by $CDLP$, wAR and AF as we vary the scaling parameter θ. We consider the hub-and-spoke network with $N = 6$ spokes and vary $\alpha \in \{0.8, 1.0, 1.2, 1.6\}$ and $\theta \in \{1.0, 0.8, 0.6, 0.4, 0.2\}$. The first column in Table 4 gives the problem characteristics (N, α, θ). The remaining columns have a similar interpretation as in Table 1. Kunnumkal and Talluri [10] show that the gap between the AF and $CDLP$ bounds is inversely proportional to the initial capacities of the resources. This explains the larger gaps between the $CDLP$ and AF bounds as the scaling parameter θ becomes smaller. The gaps between $CDLP$ and wAR also get noticeably larger as θ becomes smaller and we observe gaps as large as 10%. Furthermore, wAR is able to close roughly 75% of the gap between $CDLP$ and AF in such cases. Table 5 reports the CPU seconds required by the three solution methods as we vary the scaling parameter θ. We observe that wAR can be solved in a matter of seconds when the number of time periods and resource capacities are relatively small. On the other hand the running time of AF is still in minutes.

Hybrid control: The results in tables 4 and 5 indicate that wAR can be much more beneficial towards the end of the sales horizon when the remaining capacities of the resources tend to be low. This suggests a hybrid control where we use $CDLP$-based controls early on in the booking period and switch to wAR-based controls only towards the end of the sales horizon. Recall that for revenue estimation, we divide the booking period into five equal intervals. We consider solving $CDLP$ and using the associated marginal values to make the control decisions for the first $k \in \{0, 1, \ldots, 5\}$ intervals. We then switch to wAR and use the wAR marginal values to make the control decisions for the remaining $5 - k$ intervals. We refer to this hybrid control as $Hyb(k)$. Note that $Hyb(0)$ corresponds to wAR, while $Hyb(5)$ corresponds to $CDLP$. Table 6 gives the revenue performance of the hybrid control for the hub-and-spoke network test problems. We observe that $Hyb(4)$, which involves solving wAR only in the last interval can obtain around 30% of the revenue gains of $Hyb(0)$ over $Hyb(5)$. $Hyb(2)$ which involves using wAR-based controls for roughly half of the sales horizon, is able to obtain about 75% of the revenue boost obtained by $Hyb(0)$.

Next we test the performance of a dynamic programming decomposition approach based on wAR. After that we test the impact of introducing nonstationarity in the problem parameters and overlap in the consideration sets of the segments on the performance of wAR.

Dynamic programming decomposition: wAR emerges as a solution method which strikes a good balance between the quality of the solution and the computational cost. We test the effectiveness of a dynamic programming decomposition scheme based on wAR. Liu and van Ryzin [11] show how the $CDLP$ dual solution can be used to decompose the network problem into a number of single resource problems. Letting $\hat{\gamma} = \{\hat{\gamma}_{i,t} | \forall i, t\}$ denote the optimal values of the dual variables associated with constraints (4) in $CDLP$, the dynamic programming decomposition idea described in Liu and van Ryzin [11] is to solve the optimality equation

$$v_{i,t}^{\hat{\gamma}}(r_i^t) = \max_{S \subseteq \mathcal{S}_i(r_i^t)} \Bigg\{ \sum_{j \in S} \lambda P_j(S) \Big[f_j - \sum_{k \neq i, k \in \mathcal{I}_j} \big(\sum_{s=t+1}^{\tau} \hat{\gamma}_{k,s} \big) + v_{i,t+1}^{\hat{\gamma}} \left(r_i^t - \mathbb{1}_{[j \in \mathcal{J}_i]} \right) \Big] + [\lambda P_0(S) + 1 - \lambda] \, v_{i,t+1}^{\hat{\gamma}} \left(r_i^t \right) \Bigg\},$$

where $\mathcal{S}_i(r_i) = \{j | \mathbb{1}_{[i \in \mathcal{I}_j]} \leq r_i\}$ and the boundary conditions are $v_{i,\tau+1}^{\hat{\gamma}}(r_i) = 0$ for all r_i and $v_{i,t}^{\hat{\gamma}}(0) = 0$ for all t. Zhang and Adelman [23] show that $\min_i \{ v_{i,1}^{\hat{\gamma}}(r_i^1) + \sum_{k \neq i} (\sum_{s=t}^{\tau} \hat{\gamma}_{k,s}) r_k^1 \}$ is an upper bound on the value function and that this bound is tighter than the $CDLP$ bound.

It is possible to apply a similar decomposition idea to wAR as well by using the optimal dual variables associated with constraints (33). Furthermore, it is possible to show that this dynamic programming decomposition approach obtains an upper bound that is tighter than the wAR bound; we omit the details. Table 7 gives the upper bounds obtained by dynamic programming decomposition approaches based on $CDLP$ and wAR, referred to as $DP - CDLP$ and $DP - wAR$, respectively. The second and third columns in Table 7 give the upper bounds obtained by $DP - CDLP$ and $DP - wAR$, respectively, while the last column gives the percentage gap in the upper bounds obtained by $DP - CDLP$ and $DP - wAR$. Although $DP - wAR$ is not uniformly tighter than $DP - CDLP$, it generates bounds that are on average 1.24% tighter, and we observe gaps as high as 4.4%. It is also worthwhile noting that in many cases the wAR bound (from Table 1) is itself tighter than $DP - CDLP$.

Nonstationary arrivals: So far, all of our test problems have involved stationary problem parameters. We investigate the impact of nonstationarity in the arrival rates. We divide the booking period into three intervals of equal length. The arrival rates remain the same within each interval, but increase from the first interval to the third. The total arrival rate in the first, second, and third intervals are 0.3, 0.6 and 0.9, respectively.

Table 8 compares the upper bounds obtained by the benchmark solution methods for the hub-and-spoke network with nonstationary arrivals, while Table 9 gives the expected revenues. The results display the same trends as before. wAR obtains upper bounds that are on average 2.75% tighter than $CDLP$ and closes roughly 80% of the gap between the $CDLP$ and AF bounds. In terms of the revenue improvement, wAR generates revenues that are on average 3.8% higher than $CDLP$. For both the upper bounds and the expected revenues, the percentage gaps between $CDLP$ and wAR increase on average, compared to the stationary arrivals case. Figure 3 shows how the marginal values of capacity obtained by $CDLP$, wAR and AF vary with time in the case of nonstationary arrivals. Compared to the stationary arrivals case, $CDLP$ is a poorer approximation to AF. On the other hand, wAR continues to do a good job of tracking the AF marginal values.

Overlapping consideration sets: Finally, we consider test problems where the consideration sets of the segments overlap. We continue to work with the hub-and-spoke network test problems, except that now each origin-destination pair is associated with two customer segments. The first segment considers only the low fare products connecting the origin-destination pair, while the second segment considers the high fare products as well. Table 10 gives the upper bounds obtained by $CDLP$, wAR and AF. Note that by wAR, we mean the segment-based weak affine relaxation augmented with product-cut equalities described in §8.2. As mentioned, when the consideration sets overlap, wAR is not provably tighter than $CDLP$. However, we observe that overall wAR tends to obtain tighter bounds than $CDLP$. Table 12 gives the expected revenues obtained by the different solution methods. We find that wAR continues to provide noticeable revenue boosts over $CDLP$. Table 11 reports the CPU seconds required by $CDLP$, wAR and AF for the hub-and-spoke network test problems with overlapping consideration sets. We note that $CDLP$ does not have a compact formulation anymore and has to be solved using column generation. When the consideration sets overlap, the $CDLP$ column generation problem is intractable. Hence the $CDLP$ solution time increases considerably compared to the case with disjoint consideration sets. The solution time for wAR (with product-cut equalities) is roughly comparable to that of $CDLP$. Both wAR and $CDLP$ can be solved in a matter of minutes while AF can take hours.

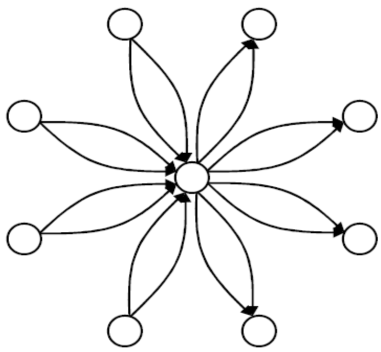

Figure 1: Structure of the airline network with a single hub and eight spokes.

Problem	Upper Bound					% Gap with $CDLP$			
(N, α)	$CDLP$	wAR	wAR^+	hAR	AF	wAR	wAR^+	hAR	AF
(4, 0.8)	7,180	7,176	7,175	7,176	7,155	0.06	0.08	0.06	0.35
(4, 1.0)	6,462	6,377	6,368	6,377	6,352	1.31	1.45	1.32	1.70
(4, 1.2)	6,138	6,053	6,044	6,053	6,027	1.38	1.53	1.39	1.81
(4, 1.6)	5,389	5,304	5,298	5,303	5,277	1.57	1.70	1.60	2.08
(6, 0.8)	6,918	6,891	6,873	6,882	6,860	0.39	0.65	0.52	0.84
(6, 1.0)	6,357	6,241	6,224	6,234	6,205	1.83	2.08	1.93	2.39
(6, 1.2)	5,799	5,683	5,666	5,676	5,654	2.00	2.28	2.12	2.50
(6, 1.6)	4,796	4,704	4,697	4,703	4,672	1.91	2.05	1.94	2.57
(8, 0.8)	6,040	5,992	5,975	5,992	5,959	0.79	1.07	0.79	1.33
(8, 1.0)	5,460	5,328	5,307	5,328	5,288	2.43	2.81	2.43	3.15
(8, 1.2)	4,993	4,857	4,839	4,857	4,817	2.73	3.10	2.73	3.52
(8, 1.6)	4,243	4,129	4,119	4,129	4,089	2.70	2.92	2.70	3.63
					avg.	1.59	1.81	1.63	2.16

Table 1: Comparison of the upper bounds for the hub-and-spoke test problems.

No. of spokes	CPU secs. $CDLP$	wAR	wAR^+	hAR	$awAR$	AF
6	0.4	16	46	23	30	98
8	0.8	46	135	59	109	405
10	1.2	105	347	127	213	1,595
12	1.9	191	698	335	407	5,204

No. of time periods	CPU secs. $CDLP$	wAR	wAR^+	hAR	$awAR$	AF
100	0.2	6	15	20	10	73
200	0.4	16	46	23	30	98
300	0.6	21	62	41	49	132
400	0.8	31	104	59	77	169

Table 2: CPU seconds for the benchmark solution methods as a function of the number of spokes in the airline network and the number of time periods in the booking horizon.

Problem	Expected Revenue					% Gap with $CDLP$			
(N, α)	$CDLP$	wAR	wAR^+	hAR	AF	wAR	wAR^+	hAR	AF
(4, 0.8)	5,755	5,748	5,814	5,819	5,744	-0.13 ⊙	1.03 ✓	1.11 ✓	-0.19 ⊙
(4, 1.0)	5,263	5,242	5,349	5,310	5,305	-0.39 ⊙	1.64 ✓	0.90 ✓	0.80 ✓
(4, 1.2)	5,056	5,080	5,107	5,095	5,136	0.47 ⊙	1.01 ✓	0.77 ✓	1.57 ✓
(4, 1.6)	4,413	4,570	4,601	4,569	4,580	3.56 ✓	4.26 ✓	3.55 ✓	3.78 ✓
(6, 0.8)	5,487	5,531	5,582	5,591	5,473	0.81 ✓	1.73 ✓	1.89 ✓	-0.25 ⊙
(6, 1.0)	5,047	5,127	5,155	5,179	5,098	1.58 ✓	2.13 ✓	2.62 ✓	1.00 ✓
(6, 1.2)	4,665	4,764	4,796	4,797	4,760	2.12 ✓	2.80 ✓	2.82 ✓	2.02 ✓
(6, 1.6)	3,824	4,101	4,109	4,104	4,075	7.23 ✓	7.46 ✓	7.34 ✓	6.56 ✓
(8, 0.8)	4,829	4,888	4,926	4,928	4,862	1.22 ✓	2.01 ✓	2.06 ✓	0.69 ✓
(8, 1.0)	4,343	4,434	4,493	4,497	4,456	2.09 ✓	3.46 ✓	3.55 ✓	2.61 ✓
(8, 1.2)	3,969	4,091	4,106	4,097	4,125	3.08 ✓	3.46 ✓	3.24 ✓	3.93 ✓
(8, 1.6)	3,384	3,579	3,561	3,570	3,570	5.77 ✓	5.21 ✓	5.50 ✓	5.49 ✓
					avg.	2.28	3.02	2.94	2.34

Table 3: Comparison of the expected revenues for the hub-and-spoke test problems.

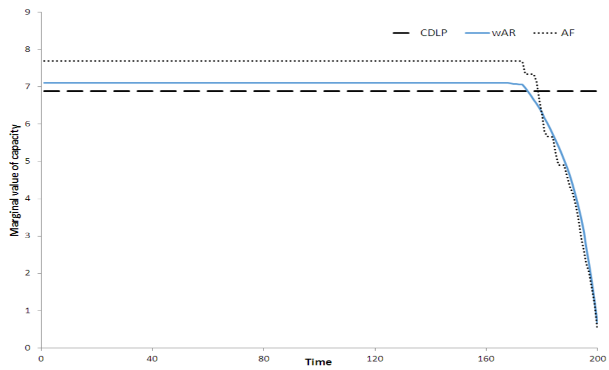

Figure 2: Marginal values of capacity obtained by $CDLP$, wAR and AF as a function of time for stationary arrivals. The plots are for the hub-and-spoke test problem with parameters (6, 1.6).

Problem	Upper Bound			% Gap with $CDLP$	
(N, α, θ)	$CDLP$	wAR	AF	wAR	AF
(6, 0.8, 1.0)	6,918	6,891	6,860	0.39	0.84
(6, 1.0, 1.0)	6,357	6,241	6,205	1.83	2.39
(6, 1.2, 1.0)	5,799	5,683	5,654	2.00	2.50
(6, 1.6, 1.0)	4,796	4,704	4,672	1.91	2.57
(6, 0.8, 0.8)	5,586	5,558	5,531	0.49	0.98
(6, 1.0, 0.8)	5,142	5,026	4,993	2.26	2.89
(6, 1.2, 0.8)	4,750	4,634	4,603	2.45	3.09
(6, 1.6, 0.8)	4,030	3,896	3,863	3.31	4.14
(6, 0.8, 0.6)	4,102	4,038	4,006	1.56	2.33
(6, 1.0, 0.6)	3,918	3,804	3,774	2.92	3.70
(6, 1.2, 0.6)	3,605	3,488	3,451	3.24	4.26
(6, 1.6, 0.6)	2,934	2,823	2,794	3.77	4.76
(6, 0.8, 0.4)	2,793	2,741	2,711	1.85	2.95
(6, 1.0, 0.4)	2,644	2,531	2,504	4.26	5.28
(6, 1.2, 0.4)	2,401	2,281	2,250	4.98	6.29
(6, 1.6, 0.4)	2,086	1,929	1,901	7.52	8.84
(6, 0.8, 0.2)	1,375	1,316	1,286	4.26	6.45
(6, 1.0, 0.2)	1,280	1,162	1,135	9.28	11.31
(6, 1.2, 0.2)	1,197	1,076	1,049	10.12	12.38
(6, 1.6, 0.2)	1,076	959	921	10.92	14.41

Table 4: Comparison of the upper bounds for the hub-and-spoke test problems with $N = 6$ under a fluid scaling.

Scaling param.	CPU secs.		
θ	$CDLP$	wAR	AF
1	0.4	16	98
0.8	0.2	13	97
0.6	0.2	10	94
0.4	0.1	4	86
0.2	0.1	1	65

Table 5: CPU seconds for the benchmark solution methods as a function of the scaling parameter θ for the hub-and-spoke test problems with $N = 6$.

Problem	Expected Revenue						% Gap with $Hyb(5)$				
(N, α)	$Hyb(0)$	$Hyb(1)$	$Hyb(2)$	$Hyb(3)$	$Hyb(4)$	$Hyb(5)$	$Hyb(0)$	$Hyb(1)$	$Hyb(2)$	$Hyb(3)$	$Hyb(4)$
(4, 0.8)	5,748	5,748	5,749	5,752	5,756	5,755	-0.13 ⊙	-0.13 ⊙	-0.10 ⊙	-0.04 ⊙	0.02 ⊙
(4, 1.0)	5,242	5,265	5,310	5,305	5,298	5,263	-0.39 ⊙	0.04 ⊙	0.89 ✓	0.79 ✓	0.67 ✓
(4, 1.2)	5,080	5,112	5,098	5,063	5,071	5,056	0.47 ✓	1.10 ✓	0.83 ✓	0.14 ⊙	0.29 ⊙
(4, 1.6)	4,570	4,535	4,535	4,485	4,452	4,413	3.56 ✓	2.76 ✓	2.77 ✓	1.64 ✓	0.88 ✓
(6, 0.8)	5,531	5,530	5,524	5,519	5,502	5,487	0.81 ✓	0.79 ✓	0.69 ✓	0.58 ✓	0.27 ⊙
(6, 1.0)	5,127	5,105	5,096	5,082	5,069	5,047	1.58 ✓	1.15 ✓	0.97 ✓	0.69 ✓	0.43 ⊙
(6, 1.2)	4,764	4,744	4,726	4,703	4,711	4,665	2.12 ✓	1.69 ✓	1.29 ✓	0.82 ✓	0.98 ✓
(6, 1.6)	4,101	4,041	4,007	3,960	3,880	3,824	7.23 ✓	5.68 ✓	4.78 ✓	3.57 ✓	1.47 ✓
(8, 0.8)	4,888	4,854	4,859	4,864	4,837	4,829	1.22 ✓	0.52 ✓	0.63 ✓	0.73 ✓	0.17 ⊙
(8, 1.0)	4,434	4,443	4,406	4,384	4,374	4,343	2.09 ✓	2.30 ✓	1.44 ✓	0.95 ✓	0.71 ✓
(8, 1.2)	4,091	4,099	4,048	4,028	4,009	3,969	3.08 ✓	3.28 ✓	2.01 ✓	1.51 ✓	1.01 ✓
(8, 1.6)	3,579	3,559	3,527	3,478	3,433	3,384	5.77 ✓	5.16 ✓	4.22 ✓	2.78 ✓	1.44 ✓
						avg.	2.28	2.03	1.70	1.18	0.69

Table 6: Comparison of the expected revenues obtained by the hybrid controls for the hub-and-spoke test problems.

Problem	Upper Bound		% Gap between $DP-CDLP$
(N,α)	$DP-CDLP$	$DP-wAR$	and $DP-wAR$
(4, 0.8)	7,146	7,158	-0.17
(4, 1.0)	6,415	6,363	0.82
(4, 1.2)	6,091	6,038	0.88
(4, 1.6)	5,323	5,266	1.08
(6, 0.8)	6,838	6,857	-0.28
(6, 1.0)	6,306	6,225	1.28
(6, 1.2)	5,750	5,667	1.43
(6, 1.6)	4,749	4,675	1.55
(8, 0.8)	5,961	5,969	-0.13
(8, 1.0)	5,408	5,310	1.80
(8, 1.2)	4,941	4,835	2.15
(8, 1.6)	4,200	4,015	4.41
		avg.	1.24

Table 7: Comparison of the upper bounds obtained by the dynamic programming decomposition approaches for the hub-and-spoke test problems.

Problem	Upper Bound			% Gap with $CDLP$	
(N,α)	$CDLP$	wAR	AF	wAR	AF
(4, 0.8)	4,400	4,396	4,380	0.09	0.45
(4, 1.0)	4,138	4,053	4,036	2.05	2.45
(4, 1.2)	3,796	3,711	3,689	2.23	2.81
(4, 1.6)	3,100	3,037	3,024	2.03	2.47
(6, 0.8)	4,311	4,256	4,236	1.28	1.72
(6, 1.0)	4,015	3,900	3,868	2.87	3.68
(6, 1.2)	3,628	3,508	3,481	3.31	4.04
(6, 1.6)	2,855	2,769	2,751	3.00	3.64
(8, 0.8)	3,802	3,678	3,650	3.25	3.99
(8, 1.0)	3,440	3,308	3,273	3.85	4.86
(8, 1.2)	3,082	2,940	2,909	4.61	5.63
(8, 1.6)	2,475	2,364	2,341	4.46	5.41
			avg.	2.75	3.43

Table 8: Comparison of the upper bounds for the hub-and-spoke test problems with nonstationary arrival rates.

Problem	Expected Revenue			% Gap with $CDLP$	
(N,α)	$CDLP$	wAR	AF	wAR	AF
(4, 0.8)	3,556	3,553	3,545	-0.08 ⊙	-0.33 ⊙
(4, 1.0)	3,290	3,334	3,329	1.33 ✓	1.19 ✓
(4, 1.2)	3,050	3,089	3,109	1.28 ✓	1.93 ✓
(4, 1.6)	2,468	2,630	2,616	6.57 ✓	6.00 ✓
(6, 0.8)	3,399	3,430	3,419	0.92 ✓	0.59 ✓
(6, 1.0)	3,134	3,181	3,173	1.48 ✓	1.23 ✓
(6, 1.2)	2,800	2,861	2,899	2.20 ✓	3.53 ✓
(6, 1.6)	2,154	2,409	2,417	11.83 ✓	12.21 ✓
(8, 0.8)	2,963	3,030	3,022	2.26 ✓	1.96 ✓
(8, 1.0)	2,642	2,733	2,703	3.41 ✓	2.29 ✓
(8, 1.2)	2,330	2,424	2,391	4.06 ✓	2.63 ✓
(8, 1.6)	1,810	1,996	1,996	10.29 ✓	10.29 ✓
			avg.	3.80	3.63

Table 9: Comparison of the expected revenues for the hub-and-spoke test problems with nonstationary arrival rates.

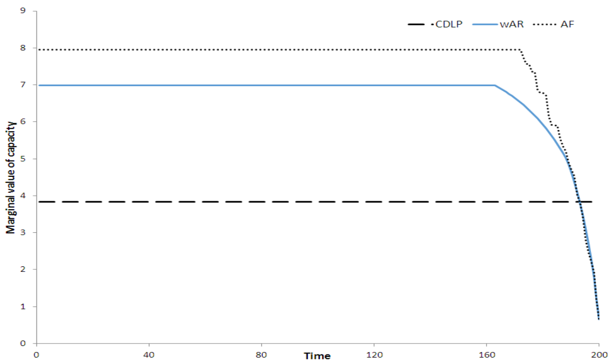

Figure 3: Marginal values of capacity obtained by $CDLP$, wAR and AF as a function of time for nonstationary arrivals. The plots are for the hub-and-spoke test problem with parameters (6, 1.6).

Problem	Upper Bound			% Gap with $CDLP$	
(N, α)	$CDLP$	wAR	AF	wAR	AF
(4, 0.8)	7,069	7,094	7,060	-0.35	0.13
(4, 1.0)	6,309	6,266	6,241	0.69	1.08
(4, 1.2)	5,975	5,907	5,879	1.14	1.60
(4, 1.6)	5,207	5,140	5,098	1.30	2.10
(6, 0.8)	6,783	6,807	6,773	-0.35	0.14
(6, 1.0)	6,240	6,149	6,109	1.46	2.10
(6, 1.2)	5,789	5,683	5,645	1.84	2.48
(6, 1.6)	4,770	4,704	4,675	1.38	2.01
(8, 0.8)	5,921	5,916	5,883	0.08	0.63
(8, 1.0)	5,342	5,233	5,193	2.04	2.79
(8, 1.2)	4,848	4,719	4,684	2.67	3.37
(8, 1.6)	4,170	4,044	3,998	3.03	4.14
			avg.	1.24	1.88

Table 10: Comparison of the upper bounds for the hub-and-spoke test problems with overlapping consideration sets.

No. of spokes	CPU secs. $CDLP$	wAR	AF
6	43	26	171
8	122	90	867
10	335	286	2,782
12	613	673	12,228

No. of time periods	CPU secs. $CDLP$	wAR	AF
100	22	7	71
200	43	26	171
300	60	101	295
400	100	301	349

Table 11: CPU seconds for the benchmark solution methods as a function of the number of spokes in the airline network and the number of time periods in the booking horizon for overlapping consideration sets.

Problem	Expected Revenue			% Gap with $CDLP$	
(N, α)	$CDLP$	wAR	AF	wAR	AF
(4, 0.8)	6,862	6,828	6,835	-0.49 ×	-0.39 ⊙
(4, 1.0)	5,827	5,887	5,913	1.04 ✓	1.48 ✓
(4, 1.2)	5,515	5,584	5,650	1.24 ✓	2.45 ✓
(4, 1.6)	4,592	4,774	4,750	3.98 ✓	3.44 ✓
(6, 0.8)	6,337	6,439	6,291	1.61 ✓	-0.73 ×
(6, 1.0)	5,799	5,738	5,730	-1.04 ×	-1.19 ×
(6, 1.2)	5,147	5,367	5,236	4.26 ✓	1.71 ✓
(6, 1.6)	4,109	4,357	4,390	6.06 ✓	6.85 ✓
(8, 0.8)	5,554	5,591	5,557	0.67 ✓	0.05 ⊙
(8, 1.0)	4,803	4,894	4,887	1.90 ✓	1.74 ✓
(8, 1.2)	4,267	4,384	4,370	2.73 ✓	2.41 ✓
(8, 1.6)	3,528	3,674	3,641	4.13 ✓	3.19 ✓
			avg.	2.17	1.75

Table 12: Comparison of the expected revenues for the hub-and-spoke test problems with overlapping consideration sets.

10 Contribution

$CDLP$ and the affine relaxation are two methods in the literature that give upper bounds on the value function for choice network revenue management. While $CDLP$ is known to be tractable for the MNL model with disjoint consideration sets, we show that the affine relaxation is NP-hard even for the single-segment MNL model. Nevertheless, our analysis helps to isolate the term in the affine relaxation which makes it hard to solve. By relaxing this difficult term, we obtain weaker, but tractable approximations. We show that our approximations yield upper bounds that are in between the $CDLP$ and affine bounds. Our relaxations retain the appeal of the formulation discovered in Gallego et al. [7] in that they involve solving compact linear programs, eliminating the need for constraint or column generation. We extend our approximations to the mixture of multinomial logit models. Our computational study indicates that our approximations often produce upper bounds that are close to the affine bound, have good revenue performance and are tractable alternatives to solving the affine relaxation.

References

[1] Ahuja, R. K., J. B. Orlin, C. Stein, R. E. Tarjan. 1994. Improved algorithms for bipartite network flow. *SIAM Journal on Computing* **23**(5) 906–933.

[2] Ben-Akiva, M., S. Lerman. 1985. *Discrete-Choice Analysis: Theory and Application to Travel Demand*. MIT Press, Cambridge, MA.

[3] Bront, J. J. M., I. Méndez-Díaz, G. Vulcano. 2009. A column generation algorithm for choice-based network revenue management. *Operations Research* **57**(3) 769–784.

[4] Dai, J., W. Ding, A. J. Kleywegt, X. Wang, Y. Zhang. 2014. Choice based revenue management for parallel flights. Working Paper, Cornell University University, NY.

[5] Feldman, J., N. Liu, H. Topaloglu, S. Ziya. 2014. Appointment scheduling under patient preference and no-show behavior. *Operations Research (forthcoming)* .

[6] Gallego, G., G. Iyengar, R. Phillips, A. Dubey. 2004. Managing flexible products on a network. Tech. Rep. TR-2004-01, Dept of Industrial Engineering, Columbia University, NY, NY.

[7] Gallego, G., R. Ratliff, S. Shebalov. 2014. A general attraction model and sales-based linear program for the network revenue management under customer choice. *Operations Research (forthcoming)* .

[8] Grötschel, M., L. Lovász, A. Schrijver. 1988. *Geometric Algorithms and Combinatorial Optimization*, vol. 2. Springer.

[9] Kunnumkal, S., K. Talluri. 2013. New compact linear programming formulations for choice network revenue management. Technical Report, Department of Economics, Universitat Pompeu Fabra, Barcelona, Spain.

[10] Kunnumkal, S., K. Talluri. 2015. A note on network revenue management dynamic programming relaxations. Working Paper, Department of Economics, Universitat Pompeu Fabra, Barcelona, Spain.

[11] Liu, Q., G. J. van Ryzin. 2008. On the choice-based linear programming model for network revenue management. *Manufacturing and Service Operations Management* **10**(2) 288–310.

[12] McFadden, D., K. Train. 2000. Mixed mnl models for discrete response. *Journal of Applied Economics* **15**.

[13] Meissner, J., A. K. Strauss. 2012. Network revenue management with inventory-sensitive bid prices and customer choice. *European Journal of Operational Research* **216**(2) 459–468.

[14] Meissner, J., A. K. Strauss, K. T. Talluri. 2013. An enhanced concave programming method for choice network revenue management. *Production and Operations Management* **22**(1) 71–87.

[15] Peeters, R. 2003. The maximum edge biclique problem is NP-complete. *Discrete Applied Mathematics* **131** 651–654.

[16] Rusmevichientong, P., D. Shmoys, C. Tong, H. Topaloglu. 2014. Assortment optimization under the multinomial logit model with random choice parameters. *Production and Operations Management (forthcoming)* .

[17] Strauss, A.K, K. T. Talluri. 2012. A tractable consideration set structure for network revenue management. Tech. rep., Working Paper 1303, Department of Economics, Universitat Pompeu Fabra, available at `http://www.econ.upf.edu/en/research/onepaper.php?id=1215`.

[18] Talluri, K. T. 2014. New formulations for choice network revenue management. *INFORMS Journal of Computing (to appear)* .

[19] Talluri, K. T., G. J. van Ryzin. 2004. *The Theory and Practice of Revenue Management*. Kluwer, New York, NY.

[20] Topaloglu, H. 2013. Joint stocking and product offer decisions under the multinomial logit model. *Production and Operations Management* **22** 1182–1199.

[21] van Ryzin, G. J., S. Mahajan. 1999. On the relationship between inventory costs and variety benefits in retail assortments. *Management Science* **45** 1496–1509.

[22] Vossen, T. W. M., D Zhang. 2012. Reductions of approximate linear program for network revenue management. Tech. rep., Leeds School of Business, University of Colorado at Boulder, Boulder, CO.

[23] Zhang, D., D. Adelman. 2009. An approximate dynamic programming approach to network revenue management with customer choice. *Transportation Science* **43**(3) 381–394.