informs
http://pubsonline.informs.org/journal/trsc

TRANSPORTATION SCIENCE
Vol. 53, No. 6, November–December 2019, pp. 1591–1608
ISSN 0041-1655 (print), ISSN 1526-5447 (online)

Choice Network Revenue Management Based on New Tractable Approximations

Sumit Kunnumkal,[a] **Kalyan Talluri**[b]

[a] Indian School of Business, Gachibowli, Hyderabad 500032, India; [b] Imperial College Business School, South Kensington Campus, SW7 2AZ London, United Kingdom

Contact: sumit_kunnumkal@isb.edu, https://orcid.org/0000-0002-6832-0508 (SK); kalyan.talluri@imperial.ac.uk, https://orcid.org/0000-0003-1608-6403 (KT)

Received: May 4, 2016
Revised: December 1, 2016; October 28, 2017; April 25, 2018
Accepted: June 24, 2018
Published Online in Articles in Advance: July 5, 2019

https://doi.org/10.1287/trsc.2018.0867

Abstract. The choice network revenue management model incorporates customer purchase behavior as probability of purchase as a function of the offered products and is appropriate 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. Consequently, a linear programming approximation called choice deterministic linear program (*CDLP*) is usually used to generate controls. Tighter approximations, such as affine and piecewise-linear relaxations, have been proposed, but it was not known if they can be solved efficiently even for simple models, such as the multinomial logit (MNL) model with a single segment. We first show that the affine relaxation (and, hence, the piecewise-linear relaxation) is NP-hard even for a single-segment MNL choice model. By analyzing the affine relaxation, we derive a new linear programming approximation that admits a compact representation, implying tractability, and prove that its value falls between the *CDLP* value and the affine relaxation value. This is the first tractable relaxation for the choice network revenue management problem that is provably tighter than *CDLP*. This approximation, in turn, leads to new policies that, in our numerical experiments, show very good promise: a 2% increase in revenue on average over *CDLP* and the values typically coming very close to the affine relaxation. We extend our analysis to obtain other tractable approximations that yield even tighter bounds. We also give extensions to the case with multiple customer segments with overlapping consideration sets in which choice by each segment is according to the MNL model.

Funding: S. Kunnumkal gratefully acknowledges the financial support of the Indian School of Business and the Smith School of Business, Queen's University.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/trsc.2018.0867.

Keywords: revenue management • airline operations research

1. Introduction and Literature Review

Revenue management controls the sale of different products that share a resource to maximize revenue, and in network revenue management (NRM), the products, in addition, consume multiple resources creating network dependencies. In this paper, we consider NRM under a choice model of consumer behavior. In the canonical airline example, resources correspond to flight legs and products correspond to itineraries that span multiple flight legs; in the car rental application, resources are automobiles of a category and a product is the consecutive days of the rental; for the hotel industry, resources correspond to hotel rooms for each night and products correspond to multinight stays. The network dependencies introduce a considerable amount of additional complexity to the stochastic control problem.

The NRM problem can be formulated as a stochastic dynamic program (DP). However, solving the Bellman equation is intractable even for very small problems because of an explosion of the state space. Considering the intractability of the NRM dynamic program, Gallego et al. (2004) and Liu and van Ryzin (2008) proposed a linear programming (LP) approximation called the choice deterministic linear program (*CDLP*, similar to some earlier deterministic approximations proposed for solving NRM under the simpler perfect segmentation assumption; see Talluri and van Ryzin 2004). The optimal objective function value of *CDLP* gives an upper bound on the value function of the NRM dynamic program. Upper bounds are useful for both deriving controls from them as well as assessing the suboptimality of policies.

The *CDLP*, however, has a drawback. The number of columns are exponential in the number of products, so it has to be solved using column generation. Liu and van Ryzin (2008) 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 (2015) show that *CDLP* has a compact linear programming formulation under the MNL model with disjoint consideration sets. On the other hand, for the problem with two segments whose consideration sets overlap, *CDLP* is intractable even for the MNL model (Bront, Méndez-Díaz, and Vulcano 2009; Rusmevichientong et al. 2014). Zhang and Adelman (2009) investigate an affine relaxation (*AF*) to the NRM dynamic program and show that it obtains a tighter upper bound than *CDLP*. The hardness result for the MNL choice model with multiple customer segments and overlapping consideration sets carries over to the affine relaxation as well. Because of the negative computational complexity results for obtaining bounds on the value function, some researchers have studied methods to obtain control policies, such as bid-price controls directly; see, for example, Chaneton and Vulcano (2011); Meissner and Strauss (2012); and Hosseinalifam, Marcotte, and Savard (2016).

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 (2009) and Meissner and Strauss (2012) are provably tighter than *CDLP*. However, in this paper, we show that the *AF* of Zhang and Adelman (2009) turns out to be intractable even for the MNL model with a single segment.

On the other hand, the approximation methods proposed by Talluri (2014) and Meissner, Strauss, and Talluri (2013) are tractable provided the consideration sets are small in size (polynomial in the size of the consideration sets). However, they are not guaranteed to produce upper bounds that are provably tighter than the *CDLP* bound. This motivates the need for tractable solution methods that tighten the *CDLP* bound.

Kunnumkal and Talluri (2016) establish analytic limits on how much the *AF* bound can improve upon the *CDLP* bound and show that real improvements are possible only under low resource availabilities, which is likely to happen closer to the end of the sales horizon. Because the upper bound obtained by our approximation methods fall in between the *CDLP* and *AF* bounds, the result of Kunnumkal and Talluri (2016) applies to our formulations also with the distinction that our approximation methods are tractable and *AF* is not.

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

1. We show that the affine relaxation of NRM is NP-hard even for the single-segment MNL model (perhaps the simplest of choice models). Our result implies that stronger solution methods that obtain tighter bounds than the affine relaxation (such as the piecewise-linear approximation proposed by Meissner and Strauss 2012) are also NP-hard for the single-segment MNL model. On the other hand, our hardness result motivates solution methods that tighten the *CDLP* bound and remain tractable, at least for the single-segment MNL model.

2. We propose a new, compact, linear programming approximation that gives a tighter bound on the dynamic program value function than *CDLP*, improving upon the work of Gallego, Ratliff, and Shebalov (2015). Compact formulations are attractive from an implementation perspective for a number of reasons: they do not require customized coding for constraint separation or column generation, and they reduce the subjectivity involved in setting the stopping criterion for the constraint or column-generation process. To our knowledge, this is the first *tractable* approximation method for MNL that is also provably tighter than *CDLP*. In numerical experiments, our approximation typically produces upper bounds that are close to the affine bound (achieving nearly 75% reduction of the gap between it and the *CDLP*) and have good revenue performance (obtaining, on average, above 95% of the revenues obtained by the affine relaxation). Running times for our new approximation are typically a fraction of that of the affine relaxation (in its faster reduced form (12) described in Section 2.5).

3. We show how our ideas can be extended to the mixture-of-multinomial-logits (MMNL) model (McFadden and Train 2000) with both disjoint as well as overlapping consideration set assumptions.

4. We propose control policies based on the new approximation and test its performance through an extensive numerical study. Our method yields noticeable benefits in terms of both tighter bounds (more than 1.5% above *CDLP* on average across instances) and improved revenue performance (more than 2% above *CDLP* on average across instances). The benefits primarily come from sharper value function approximations toward the end of the selling horizon when capacity tends to be relatively scarce. So one option for practitioners is to switch to our method during the last few days of the sales horizon.

The remainder of the paper is organized as follows: In Section 2, we describe the choice NRM model, the notation, the basic dynamic program, the *CDLP*, and the affine relaxation of the NRM dynamic program. Next, in Section 3, we show that the affine relaxation is NP-hard even for the single-segment MNL model. We describe our first tractable approximation method in Section 4. Section 5 discusses extensions to the MMNL model. Section 6 contains our computational study using the new approximation.

2. Problem Formulation

We are interested in controlling the sale of products over a finite sales horizon. A product is a specification of a price and the set of resources that it consumes. Time is discrete, and the sales horizon consists 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, one if true and zero if false, and $\mathbb{1}_{[\mathcal{I}_j]}$ denote the vector of resources used by product j with a one in the ith position if $i \in \mathcal{I}_j$ and a zero 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.

2.1. Demand Model

We have multiple customer segments, each with distinct purchase behavior. The segmentation of the customers could be according to different criteria—for example, price sensitivities, demographics, or even geographic locations. 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 zero 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 one. The probability that a segment-l customer purchases product j when S is the offer set is

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

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 (1985) 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, $\lambda P_j(S) = \sum_l \lambda_l P_j^l(S)$, and $1 - \lambda + \lambda P_0(S) = 1 - \sum_{j\in S} \lambda P_j(S)$ 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 $R^l(S) = \sum_{j\in S} f_j P_j^l(S)$ and $R(S) = \sum_{j\in S} f_j P_j(S)$.

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

2.2. Choice Dynamic Program

The 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

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

where

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

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. We let $V^{DP} = V_1(\boldsymbol{r}^1)$ denote the optimal expected revenue over the sales horizon given the initial capacity vector $\boldsymbol{r}^1$.

2.3. Linear Programming Formulation of the Dynamic Program

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

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

Both dynamic program (2) and *DPLP* are computationally intractable, but *DPLP* turns out to be useful in developing value function approximation methods as shown in Zhang and Adelman (2009). In the following, we describe two approximation methods, namely the choice deterministic linear program and the affine relaxation. Carefully analyzing the differences between the two formulations leads to our new tractable approximation.

2.4. Choice Deterministic LP

The (*CDLP*) proposed in Gallego et al. (2004) and Liu and van Ryzin (2008) 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 ensures that the total expected capacity consumed on resource i up until time period t does not exceed the available capacity. Note that, because $h_{S,t} \ge 0$, 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 one.

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

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

Liu and van Ryzin (2008) 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. We describe one heuristic control proposed by Zhang and Adelman (2009): Letting $\hat{\gamma} = \{\hat{\gamma}_{i,t} \mid \forall i, t\}$ 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 (2009) approximate the value function $V_t(\boldsymbol{r}^t)$ as

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

The heuristic control replaces the value function by its approximation in optimality Equation (2) to determine the offer set. That is, if $\boldsymbol{r}^t$ is the vector of remaining resource capacities at time t, the heuristic control solves the problem

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

and offers the set that achieves the maximum in this 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 (2008) 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 $\mathscr{C}_l \cap \mathscr{C}_m = \emptyset$ for segments l and m. Under the same set of assumptions, Gallego, Ratliff, and Shebalov (2015) further show that *CDLP* has the following equivalent, compact formulation

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

In this sales-based linear program (*SBLP*), the decision variables $x_{j,t}^l$ can be interpreted as the sales rate for product j at time t. Note that *SBLP* is a compact formulation because 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, Méndez-Díaz, and Vulcano (2009) and Rusmevichientong et al. (2014) show that the *CDLP* column generation is NP-hard even under the MNL choice model.

2.5. Affine Relaxation

The second approximation method we consider is the affine relaxation, with which 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

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

with the boundary conditions $\theta_{\tau+1} = 0$, $V_{i,\tau+1} = 0$. Zhang and Adelman (2009) 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.

Although the number of decision variables in *AF* is manageable, the number of constraints is exponential in both the number of products as well as the number of resources. Vossen and Zhang (2015) use Dantzig–Wolfe decomposition to derive a reduced, equivalent formulation of *AF*, in which the number of constraints is exponential only in the number of products.

We give an alternative, simpler proof of the reduction here. 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

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

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 because there exists an optimal solution to *AF* that satisfies $V_{i,t} - V_{i,t+1} \ge 0$.

Now, constraints (10) can be written as

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

for all t. Because $\gamma_{i,t} \ge 0$, 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 $S \subseteq \mathcal{S}(r)$ and would only increase the objective value). Moreover, because $\gamma_{i,t} \ge 0$ 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

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

And we can write *AF* equivalently as

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

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

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

These arguments imply the following:

Proposition 1. (Vossen and Zhang 2015). $V^{AF} = V^{RAF} = V^{dRAF}$.

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 $\hat{\beta} = \{\hat{\beta}_t \mid \forall t\}$ and $\hat{\gamma} = \{\hat{\gamma}_{i,t} \mid \forall i, t\}$, denote an optimal solution to *RAF*, we use $\sum_{k=t}^{\tau} \hat{\gamma}_{i,k}$ to approximate the marginal value of capacity on resource i at time t. We approximate $V_t(\boldsymbol{r}^t)$ 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 (2009) 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.

3. 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 because it is one of the few cases in which *CDLP* is tractable. We show that the affine relaxation is NP-hard even for this simple choice model.

Let the preference weights be w_j (as mentioned earlier, we drop the segment index l when we are analyzing a single-segment problem). The choice probabilities, expected resource consumptions, and expected revenues are then given by

$$P_j(S) = \frac{\mathbb{1}_{[j\in S]} w_j}{1 + \sum_{k\in S} w_k} \qquad Q_i(S) = \frac{\sum_{j\in \mathcal{I}_i \cap S} w_j}{1 + \sum_{j\in S} w_j}$$
$$R(S) = \frac{\sum_{j\in S} f_j w_j}{1 + \sum_{j\in S} w_j}. \tag{13}$$

Because *RAF* has an exponential number of constraints, we have to use constraint separation and generate constraints (12) violated by a solution on the fly. Following the result of Grötschel, Lovász, and Schrijver (1988), polynomial solvability of an LP is equivalent to polynomial-time generation of violated constraints, and so we focus on separating constraints (12).

Substituting (13) into constraint (12), we obtain

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

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

$$\beta_t \geq -\gamma_{S,t} \left(1 + \sum_{j\in S} w_j \right) - \sum_{j\in S} \zeta_{j,t}(\beta, \gamma), \tag{14}$$

where

$$\zeta_{j,t}(\beta, \gamma) = w_j \left[\beta_t + \lambda \left(\left(\sum_{i\in \mathcal{I}_j} \sum_{k=t+1}^{\tau} \gamma_{i,k} \right) - f_j \right) \right]. \tag{15}$$

Because 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

$$\Pi_t^{AF}(\beta, \gamma) = \max_S \left\{ -\gamma_{S,t} \left(1 + \sum_{j\in S} w_j \right) - \sum_{j\in S} \zeta_{j,t}(\beta, \gamma) \right\} \tag{16}$$

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

$$\beta_t \geq \Pi_t^{AF}(\beta, \gamma) \quad \forall t. \tag{17}$$

Generating constraints on the fly involves checking, given a set of values (β, γ), if constraint (14) 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 (16) and determining if $\beta_t \geq \Pi_t^{AF}(\beta, \gamma)$. If $\beta_t \geq \Pi_t^{AF}(\beta, \gamma)$, then constraint (14) is satisfied for all S at time t. Otherwise, the set $\hat{S}$ that attains the maximum in problem (16) violates the constraint, and we add the constraint for set $\hat{S}$ to the LP.

Proposition 2 states that the affine relaxation separation problem for MNL with a single segment, as given in (14), is NP-hard.

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* (14)*?*

Proof. Our reduction is from the NP-complete maximum edge biclique problem (Peeters 2003). 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, that is, 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 $k_1 k_2$.

3.1. Maximum Edge Biclique Problem

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 $C_S = \{v \in V_2 | \exists e = (u, v) \in \bar{G}, u \in S\}$. 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 (14). In Equation (14), 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 (14) for this instance.

With these values, $S \subseteq V_1$ with $|S| = k_1, |C(S)| = m_2 - k_2$ violates (14) 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

$$p < k_2 k_1. \quad \square$$

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 retaining tractability.

4. Weak Affine Relaxation

In this section, we propose our 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. Although this is primarily for clarity of exposition, we note that the single-segment results may be of independent interest, especially in the context of optimization of personalized assortments; see for example Golrezaei, Nazerzadeh, and Rusmevichientong (2014) and Gallego et al. (2016). In Section 5, we show how the ideas can be readily extended to the MNL model with multiple customer segments.

4.1. 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*

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

and

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

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. □

4.2. *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, and 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 a more tractable term 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 (14) except that the inner summation over k runs from t instead of $t+1$. To make the comparison with *AF* easier, we rewrite the constraint as

$$\beta_t \geq \Pi_t^{CDLP}(\beta, \gamma) \quad \forall t, \tag{18}$$

where

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

and $\zeta_{j,t}(\beta,\gamma)$ is defined in (15). Because $0 \leq \lambda \leq 1$, 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*.

4.3. A New Tractable Approximation

We are now ready to describe our tractable approximation method, which we refer to as weak affine relaxation (wAR). The difficult term in (16) 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 (16) with $\gamma_{S,t} + \sum_{j\in S} w_j(\sum_{i\in\mathcal{I}_j} \gamma_{i,t})$ and solve the LP

$$\begin{aligned} V^{wAR} = \min_{\beta,\gamma} \quad & \sum_t \beta_t + \sum_t \sum_i \gamma_{i,t} r_i^1 \\ (wAR) \text{ s.t.} \quad & \beta_t \geq \Pi_t^{wAR}(\beta,\gamma) \quad \forall t \\ & \gamma_{i,t} \geq 0 \quad \forall i, t, \end{aligned} \tag{20}$$

where

$$\Pi_t^{wAR} = \max_S \left\{ -\gamma_{S,t} - \sum_{j\in S} w_j \left(\sum_{i\in\mathcal{I}_j} \gamma_{i,t}\right) - \sum_{j\in S} \zeta_{j,t}(\beta,\gamma)\right\}. \tag{21}$$

Proposition 3 shows that wAR obtains an upper bound on the value function that is weaker than *AF* but stronger than *CDLP*.

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, $\Pi_t^{AF}(\beta,\gamma) \leq \Pi_t^{wAR}(\beta,\gamma) \leq \Pi_t^{CDLP}(\beta,\gamma)$, 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 LP in which the number of variables and constraints is polynomial in the number of products and resources.

Observe that solving problem (21) in an efficient manner is key to separating the weak affine relaxation constraints efficiently. Therefore, we focus on solving optimization problem (21). 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 (21) can be formulated as the integer program

$$\Pi_t^{wAR}(\beta,\gamma) = \max_{q,u} \quad -\sum_i \gamma_{i,t} q_{i,t} - \sum_j \left[\zeta_{j,t}(\beta,\gamma) + w_j \left(\sum_{i\in\mathcal{I}_j} \gamma_{i,t}\right)\right] u_{j,t} \tag{22}$$

$$\text{s.t.} \quad u_{j,t} - q_{i,t} \leq 0 \quad \forall i \in \mathcal{I}_j, \forall j \tag{23}$$

$$q_{i,t} \leq 1 \quad \forall i \tag{24}$$

$$u_{j,t} \geq 0, \ \text{integer} \quad \forall j. \tag{25}$$

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 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 (22)–(25) exactly as an LP. In fact, problems (22)–(25) 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 et al. (1994) 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, problems (22)–(25) 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. Because the separation problem can be solved as an LP in which 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 (22)–(25)

with dual variables $\pi_{i,j,t}$ corresponding to (23) and $\psi_{i,t}$ to (24):

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

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

$$\begin{aligned} \Pi_t^{wAR}(\beta,\gamma) = \min_{\pi} \quad & \sum_i \left[\sum_{j \in \mathcal{J}_i} \pi_{i,j,t} - \gamma_{i,t}\right] \\ \text{s.t.} \quad & \sum_{i \in \mathcal{I}_j} \pi_{i,j,t} \geq -\left[\zeta_{j,t}(\beta,\gamma) + w_j\left(\sum_{i \in \mathcal{I}_j} \gamma_{i,t}\right)\right] \\ & \qquad\qquad \forall j \qquad\qquad (26) \\ & \sum_{j \in \mathcal{J}_i} \pi_{i,j,t} \geq \gamma_{i,t} \quad \forall i \qquad\qquad (27) \\ & \pi_{i,j,t} \geq 0 \quad \forall i, j \in \mathcal{J}_i. \end{aligned}$$

Now we fold in the LP formulation of $\Pi_t^{wAR}(\beta,\gamma)$ into constraints (20) and write *wAR* equivalently as

$$\begin{aligned} V^{wAR} = \min_{\beta,\gamma,\pi} \quad & \sum_t \beta_t + \sum_t \sum_i \gamma_{i,t} r_i^1 \\ \text{s.t.} \quad & \beta_t \geq \sum_i \left[\sum_{j \in \mathcal{J}_i} \pi_{i,j,t} - \gamma_{i,t}\right] \quad \forall t \\ & (26), (27)\ \forall t \\ & \gamma_{i,t}, \pi_{i,j,t} \geq 0 \quad \forall i, j \in \mathcal{J}_i, t. \end{aligned}$$

The size of the 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 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, Ratliff, and Shebalov (2015), in which the sales rates at each time period are modulated by the expected remaining resource capacities.

The weak affine relaxation is based on isolating the difficult term in the affine relaxation and replacing it with a simpler, more tractable term. The separation problem involving the simpler, more tractable term can be formulated as an LP. Taking the dual of the LP formulation of the separation problem yields the compact formulation of the weak affine relaxation. One advantage of having a compact formulation is that it eliminates the overhead associated with optimizing the constraint-separation code and memory management. Another benefit is that it reduces the subjectivity involved in setting the stopping criterion for the constraint-separation process. It is possible to build on these ideas and obtain other tractable approximation methods that further tighten the *wAR* bound. We describe two such approximations in the online appendix.

5. MNL with Multiple Customer Segments

In this section, we describe how to extend the weak affine relaxation of Section 4 to the MMNL model. The MMNL model is a rich choice model that can approximate any random utility choice model arbitrarily closely (McFadden and Train 2000). In Section 5.1, we consider the MMNL choice model with disjoint consideration sets. In Section 5.2, we consider the case in which the consideration sets of the different segments overlap. It is also possible to extend the weak affine relaxation idea to the general attraction model of Gallego, Ratliff, and Shebalov (2015) in a transparent manner.

5.1. Disjoint Consideration Sets

We consider the case in which 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 in which the *CDLP* formulation is tractable. We describe 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 $\mathcal{I}_l = \{i \in \mathcal{I} \mid \exists j \in \mathcal{C}_l \text{ and } j \in \mathcal{J}_i\}$ and $\mathcal{L}_i = \{l \in \mathcal{L} \mid i \in \mathcal{I}_l\}$. 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. Letting λ_l denote the arrival rate for segment l, we can interpret $\sum_{l \in \mathscr{L}_i} \lambda_l$ as the effective arrival rate for 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 $S_l = S \cap \mathscr{C}_l$, constraint (12) can be written as

$$\beta_t + \sum_i \mathbb{1}_{[i \in \mathscr{I}_S]} \gamma_{i,t} + \sum_i \left[\left(\sum_{k=t+1}^{\tau} \gamma_{i,k} \right) \sum_l \lambda_l Q_i^l(S) \right] \geq \sum_l \lambda_l R^l(S). \tag{28}$$

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

$$\beta_{l,t} + \sum_{i \in \mathscr{I}_l} \mathbb{1}_{[i \in \mathscr{I}_{S_l}]} \gamma_{i,t} \lambda_i^l + \sum_i \left[\left(\sum_{k=t+1}^{\tau} \gamma_{i,k} \right) \lambda_l Q_i^l(S_l) \right] \geq \lambda_l R^l(S_l) \tag{29}$$

for each $S_l = S \cap \mathscr{C}_l$, where $\lambda_i^l = \lambda_l / \sum_{l' \in \mathscr{L}_i} \lambda_{l'}$ can be interpreted as the probability of a segment-l arrival given the arrival of a segment that uses resource i. The proof of Proposition 4 shows that the segment level constraints (29) imply (28) and that we obtain a looser upper bound by separating over (29) instead of (28).

We observe that the segment level constraints (29) 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:

$$\begin{aligned} \Pi_{l,t}^{swAR}(\beta, \gamma) = \max_{q,u} \quad & -\sum_{i \in \mathscr{I}_l} \lambda_i^l \gamma_{i,t} q_{i,t} \\ & - \sum_{j \in \mathscr{C}_l} \left[\zeta_{j,t}^l(\beta, \gamma) + w_j^l \sum_{i \in \mathscr{I}_j} \lambda_i^l \gamma_{i,t} \right] u_{j,t} \\ \text{s.t.} \quad & (23) - (25), \end{aligned}$$

where

$$\zeta_{j,t}^l(\beta, \gamma) = w_j^l \left[\beta_{l,t} + \lambda_l \left(\left(\sum_{i \in \mathscr{I}_j} \sum_{k=t+1}^{\tau} \gamma_{i,k} \right) - f_j \right) \right]. \tag{30}$$

We replace constraint (29) 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 following compact LP:

$$\begin{aligned} V^{swAR} = \min_{\gamma, \beta, \pi} \quad & \sum_t \sum_l \beta_{l,t} + \sum_i \sum_t \gamma_{i,t} r_i^1 \\ (swAR)\ \text{s.t.} \quad & \beta_{l,t} \geq \sum_{i \in \mathscr{I}_l} \left[\sum_{j \in \mathscr{J}_i, j \in \mathscr{C}_l} \pi_{i,j,t} - \lambda_i^l \gamma_{i,t} \right] \quad \forall l, t \\ & \sum_{i \in \mathscr{I}_j} \pi_{i,j,t} \geq - \left[\zeta_{j,t}^{\ell_j}(\beta, \gamma) + w_j^{\ell_j} \left(\sum_{i \in \mathscr{I}_j} \lambda_i^{\ell_j} \gamma_{i,t} \right) \right] \\ & \qquad \forall j, t \\ & \sum_{j \in \mathscr{J}_i, j \in \mathscr{C}_l} \pi_{i,j,t} - \lambda_i^l \gamma_{i,t} \geq 0 \quad \forall i, l \in \mathscr{L}_i, t \\ & \gamma_{i,t}, \pi_{i,j,t} \geq 0 \quad \forall, i, j \in \mathscr{J}_i, t, \end{aligned}$$

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. In particular, *swAR* coincides with *wAR* if there is only a single segment. Note that *swAR* is again tractable as it is a compact LP. Proposition 4 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), (30), and rearranging terms, the *swAR* constraint $\beta_{l,t} \geq \Pi_{l,t}^{swAR}(\beta, \gamma)$ can be equivalently written as

$$\begin{aligned} \beta_{l,t} \geq \lambda_l & \left[R^l(S_l) - \sum_{i \in \mathscr{I}_l} \sum_{k=t+1}^{\tau} Q_i^l(S_l) \gamma_{i,k} \right] \\ & - \sum_{i \in \mathscr{I}_l} \mathbb{1}_{[i \in \mathscr{I}_{S_l}]} \gamma_{i,t} \lambda_i^l \left(\sum_{j \in \mathscr{J}_i} P_j^l(S_l) + P_0^l(S_l) \right) \end{aligned} \tag{31}$$

for all $S_l \subseteq \mathscr{C}_l$, where $\lambda_i^l = \frac{\lambda_l}{\sum_{l' \in \mathscr{L}_i} \lambda_{l'}}$.

Consider now two intermediate problems:

$$\begin{aligned} \underline{V} = \min_{\beta, \gamma} \quad & \sum_t \sum_l \beta_{l,t} + \sum_t \sum_i \gamma_{i,t} r_i^1 \\ \text{s.t.} \quad & (29) \quad \forall l, S_l \subseteq \mathscr{C}_l, t \\ & \gamma_{i,t} \geq 0 \quad \forall i, t, \end{aligned}$$

and

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

We can interpret the first problem as a segment-based relaxation of *AF*, and the second problem can be viewed as a segment-based relaxation of *CDLP*.

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

(i) $\underline{V} \le V^{swAR} \le \bar{V}$. Because the objective functions of all the problems are the same, we only need to compare the corresponding constraints. Because $\sum_{j \in \mathcal{J}_i} P^l_j(S_l) + P^l_0(S_l) \le 1$, it follows that constraint (31) implies constraint (29), and we have $\underline{V} \le V^{swAR}$.

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

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

Now note that

$$\begin{aligned}
\lambda_l Q^l_i(S_l)\gamma_{i,t} &= \lambda_l \mathbb{1}_{[i \in \mathcal{I}_{S_l}]} Q^l_i(S_l)\gamma_{i,t} \\
&= \lambda_l \mathbb{1}_{[i \in \mathcal{I}_{S_l}]} \left[\sum_{j \in \mathcal{J}_i} P^l_j(S_l) \right] \gamma_{i,t} \\
&\le \lambda^l_i \mathbb{1}_{[i \in \mathcal{I}_{S_l}]} \left[\sum_{j \in \mathcal{J}_i} P^l_j(S_l) \right] \gamma_{i,t} \\
&\le \lambda^l_i \mathbb{1}_{[i \in \mathcal{I}_{S_l}]} \left[\sum_{j \in \mathcal{J}_i} P^l_j(S_l) + P^l_0(S_l) \right] \gamma_{i,t},
\end{aligned}$$

where the first equality holds because, if $\mathbb{1}_{[i \in \mathcal{I}_{S_l}]} = 0$, then $Q^l_i(S_l) = 0$, and the first inequality holds because $\sum_{l' \in \mathcal{L}_i} \lambda_{l'} \le 1$, and so $\lambda_l \le \frac{\lambda_l}{\sum_{l' \in \mathcal{L}_i} \lambda_{l'}} = \lambda^l_i$. 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 (29). We show that it satisfies constraints (28) as well. Fix a set S and let $S_l = S \cap \mathcal{C}_l$. Adding up constraints (29) for all the segments,

$$\begin{aligned}
\sum_l \hat{\beta}_{l,t} &\ge \sum_l \Bigg\{ \lambda_l \left[R^l(S_l) - \sum_{i \in \mathcal{I}_l} \sum_{k=t+1}^{\tau} Q^l_i(S_l) \hat{\gamma}_{i,k} \right] \\
&\qquad - \sum_{i \in \mathcal{I}_l} \mathbb{1}_{[i \in \mathcal{I}_{S_l}]} \hat{\gamma}_{i,t} \lambda^l_i \Bigg\} \\
&= \lambda \left[R(S) - \sum_i \sum_{k=t+1}^{\tau} Q_i(S) \hat{\gamma}_{i,k} \right] - \sum_i \hat{\gamma}_{i,t} \sum_{l \in \mathcal{L}_i} \mathbb{1}_{[i \in \mathcal{I}_{S_l}]} \lambda^l_i \\
&\ge \lambda \left[R(S) - \sum_i \sum_{k=t+1}^{\tau} Q_i(S) \hat{\gamma}_{i,k} \right] - \sum_i \hat{\gamma}_{i,t} \sum_{l \in \mathcal{L}_i} \mathbb{1}_{[i \in \mathcal{I}_{S}]} \lambda^l_i \\
&= \lambda \left[R(S) - \sum_i \sum_{k=t+1}^{\tau} Q_i(S) \hat{\gamma}_{i,k} \right] - \sum_i \hat{\gamma}_{i,t} \mathbb{1}_{[i \in \mathcal{I}_{S}]},
\end{aligned}$$

where the first equality uses the fact that $Q^l_i(S_l) = 0$ for $l \notin \mathcal{L}_i$, and hence, $\lambda Q_i(S) = \sum_l \lambda_l Q^l_i(S_l) = \sum_{l \in \mathcal{L}_i} \lambda_l Q^l_i(S_l)$. The second inequality holds because $\mathbb{1}_{[i \in \mathcal{I}_{S_l}]} \le \mathbb{1}_{[i \in \mathcal{I}_{S}]}$. Letting $\tilde{\beta} = \{\tilde{\beta}_t = \sum_l \hat{\beta}_{l,t} \mid \forall t\}$, it follows that $(\tilde{\beta}, \hat{\gamma})$ satisfies constraints (28). Therefore, $V^{AF} \le \sum_t \tilde{\beta}_t + \sum_t \sum_i \hat{\gamma}_{i,t} = \underline{V}$. Meissner, Strauss, and Talluri (2013) prove the following that we include for completeness.

(iii) $\bar{V} = V^{CDLP}$ (Meissner, Strauss, and Talluri 2013). Constraints (6) in *dCDLP* are equivalent to

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

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

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

which is exactly constraint (32). □

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

$$\begin{aligned}
V^{dswAR} = \max_{x,\rho} \quad & \sum_t \sum_l \sum_{j \in \mathcal{C}_l} \lambda_l f_j x^l_{j,t} \\
(dswAR)\ \text{s.t.} \quad & \sum_{l \in \mathcal{L}_i} \Bigg[\lambda^l_i x^l_{0,t} + \sum_{s=1}^{t-1} \sum_{j \in \mathcal{J}_i \cap \mathcal{C}_l} \lambda_l x^l_{j,s} \qquad\qquad (33) \\
& \quad + \sum_{j \in \mathcal{J}_i \cap \mathcal{C}_l} \lambda^l_i x^l_{j,t} - \lambda^l_i \rho^l_{it} \Bigg] \le r^1_i \quad \forall i, t \\
& x^l_{0,t} + \sum_{j \in \mathcal{C}_l} x^l_{j,t} = 1 \quad \forall l, t \\
& \frac{x^l_{j,t}}{w^l_j} - x^l_{0,t} + \rho^l_{i,t} \le 0 \quad \forall l, i, j \in \mathcal{J}_i \cap \mathcal{C}_l, t \\
& x^l_{0,t}, x^l_{j,t}, \rho^l_{i,t} \ge 0 \quad \forall l, i, j \in \mathcal{J}_i \cap \mathcal{C}_l, t. \qquad\qquad (34)
\end{aligned}$$

5.2. 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, Strauss, and Talluri (2013), is to formulate the problem by segments and then add a set of consistency conditions called *product-cut* (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 (2017) subsequently show that when the consideration set structure has a certain tree structure, the cuts, in fact, achieve the *CDLP* value. Talluri (2014) 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 approximation.

We begin with a brief description of the PC equalities: Meissner, Strauss, and Talluri (2013) 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 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 (2014) uses choice probabilities (1) to specialize the PC equalities to the MNL model as

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

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

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

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 (2014). If the overlap in the consideration sets of the different segments is not too large, then the number of PC equalities is manageable.

Talluri (2014) shows that adding PC equalities (35)–(37) to the *SBLP* of Gallego, Ratliff, and Shebalov (2015) 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 the following:

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.

6. Computational Experiments

In this section, we compare the upper bounds and the revenues obtained by *CDLP*, *wAR*, and *AF*. We begin by describing the experimental setup.

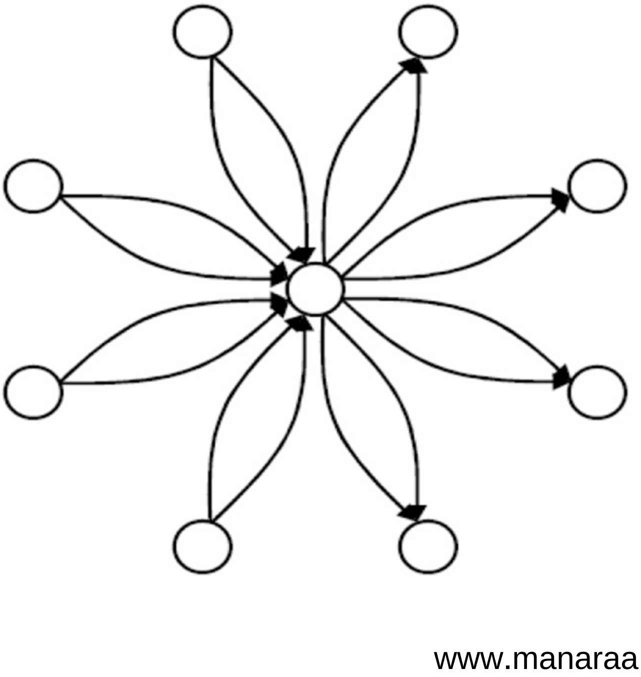
6.1. Test Network

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, and the remaining half have two flights from the hub so that the total number of flights is $2N$. All the flights have identical capacities. Figure 1 shows the structure of the network with $N = 8$.

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 $2N^2$ fare-products connect spoke-to-spoke origin–destination pairs. Half of the $2N^2$ 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 two customer segments. The first segment considers only the low fare products connecting its origin–destination

Figure 1. Structure of the Airline Network with a Single Hub and Eight Spokes



pair, and the second segment considers the high fare-products as well. Therefore, the consideration sets of the different customer segments overlap. Moreover, within each segment choice is governed by the MNL model, and we sample the preference weights of the fare-products in its consideration set from a Poisson distribution with a mean of 100 and set the no-purchase preference weight to be $0.5 \sum_{j\in\mathscr{C}_l} w_j^l$. So the probability that a customer does not purchase anything when all the products in the consideration set are offered is one in three.

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

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

where λ denotes the total arrival rate in a time period. We set $\lambda = 0.9$ and have $\tau = 200$ time periods in all of our test problems. We label our test problems by (N, α), where $N \in \{4, 6, 8\}$ and $\alpha \in \{0.8, 1.0, 1.2, 1.6\}$.

6.2. Results

As we mentioned earlier, it is known that the gap between *CDLP* and affine relaxation diminishes to zero with increasing capacities (Kunnumkal and Talluri 2016). So it is not possible to get large problems in which the gap between weak affine relaxation and *CDLP* values are significant. Most of the benefits of *wAR*, therefore, are likely to happen when the capacities are small, near the end of the booking horizon. We validate this intuition by performing numerical experiments on (i) the differences in the values of the various methods at small capacities, (ii) revenue simulations with small capacities, and (iii) revenue simulations on larger real-world networks in which we turn on *wAR*-recommended controls at the halfway point with *CDLP* recommendations controlling the initial half.

6.2.1. Upper Bounds. Table 1 gives the upper bounds obtained by the benchmark solution methods. The first two columns in the table give the problem characteristics. The third, fourth, and fifth columns, respectively, give the upper bounds obtained by *CDLP*, *wAR*, and *AF*. The last two columns give the percentage gap between the upper bounds obtained by *CDLP* and *wAR* and *CDLP* and *AF*, respectively. We note that by *wAR*, we mean the segment-based weak affine relaxation augmented with product-cut equalities described in Section 5.2, and by *AF*, we mean the reduced formulation *RAF*. We solve *CDLP* and *AF* using column generation and stop when they are within 1% of optimality. Based on initial setup runs, this seemed to provide a good balance between the quality of the solution and the computational effort involved. Figure 2 illustrates how the quality of the solution and the computational time required to solve *CDLP* varies with the stopping criterion for a representative test problem.

Table 1. Comparison of the Upper Bounds for the Hub-and-Spoke Test Problems with Overlapping Consideration Sets

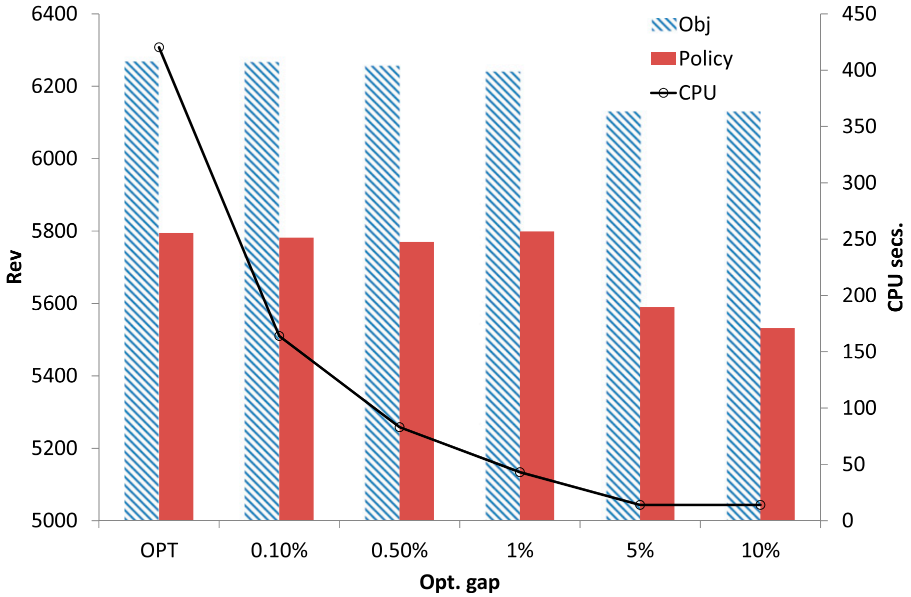
Problem	Average	Upper bound			% gap with *CDLP*	
(N, α)	capacity	*CDLP*	*wAR*	*AF/RAF*	*wAR*	*AF/RAF*
(4, 0.8)	21	7,069	7,094	7,060	–0.35	0.13
(4, 1.0)	16	6,309	6,266	6,241	0.69	1.08
(4, 1.2)	14	5,975	5,907	5,879	1.14	1.60
(4, 1.6)	11	5,207	5,140	5,098	1.30	2.10
(6, 0.8)	13	6,783	6,807	6,773	–0.35	0.14
(6, 1.0)	11	6,240	6,149	6,109	1.46	2.10
(6, 1.2)	9	5,789	5,683	5,645	1.84	2.48
(6, 1.6)	7	4,770	4,704	4,675	1.38	2.01
(8, 0.8)	10	5,921	5,916	5,883	0.08	0.63
(8, 1.0)	8	5,342	5,233	5,193	2.04	2.79
(8, 1.2)	7	4,848	4,719	4,684	2.67	3.37
(8, 1.6)	5	4,170	4,044	3,998	3.03	4.14
				Average	1.24	1.88

AF generates the tightest upper bound, followed by *wAR* and then by *CDLP*. The average percentage gap between *wAR* and *CDLP* is 1.24% although we observe instances in which the gap is as high as 3%. In our test problems, there is overlap in the consideration sets of the different segments, and therefore, *wAR* is not guaranteed to be tighter than *CDLP*. However, we observe that, overall, *wAR* tends to obtain tighter bounds than *CDLP*. The percentage gap between *wAR* and *CDLP* seems to increase with the nominal load factor and the number of spokes in the network. *AF* obtains bounds that are, on average, 1.88% tighter than *CDLP*. *wAR* closes about 70% of the gap between the *AF* and *CDLP* bounds.

6.2.2. Revenue Results. Table 2 gives the expected revenues obtained by the different benchmark methods. We evaluate the revenue performance by simulation and use common random numbers in our simulations. In our revenue simulations, we divide the booking period into five equal intervals. At the beginning of each interval, we resolve the benchmark solution methods to get fresh estimates for the marginal value of capacity on the resources. Recall that 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 in which we resolve the benchmark solution methods.

The columns in Table 2 have a similar interpretation as in Table 1 except that they give the expected total

Figure 2. (Color online) *CDLP* Objective Function Values, Revenues, and CPU Times as a Function of the Stopping Optimality Gap for the Hub-and-Spoke Test Problem (6, 1.0)



Note. OPT means that *CDLP* is solved to optimality.

revenues. In the last two columns, we use a ✓ to indicate that the corresponding benchmark method generates higher revenues than *CDLP* at the 95% level, a ⊙ if the difference in the revenue performance of the benchmark method and *CDLP* is not significant at the 95% level, and an × if the benchmark method generates lower revenues than *CDLP* at the 95% level. *wAR*, on average, generates revenues that are 2.17% higher than *CDLP* although we observe instances in which the gap is as high as 6%. As with the upper bounds, the revenue boosts are more noticeable at the higher load factors. It is interesting to note that the magnitude of the revenue gaps is larger than that of the upper bounds. *AF* generates revenues that are, on average, 1.75% higher than *CDLP*, and its revenue performance is comparable with that of *wAR*.

6.2.3. Robustness Checks.

A natural question that arises concerns the sensitivity of the upper bounds and the revenues to the column-generation stopping criterion. To address this, we compare the performance of the benchmark methods on an additional set of test problems in which we solve *CDLP* to optimality. We continue to work with the hub-and-spoke network structure except that we now associate each origin–destination pair with a single customer segment. Moreover, each segment is only interested in the fare-products connecting the particular origin–destination pair. Therefore, the consideration sets of the different customer segments do not overlap now, and we can solve *CDLP* to optimality using the compact sales-based formulation (*SBLP*). Table 3 gives the upper bounds obtained by the benchmark solution methods, and Table 4 gives the expected revenues. We observe that the nature of the results do not change significantly even when we solve *CDLP* to optimality. *wAR* generates tighter bounds than *CDLP* and closes nearly 75% of the gap between the *AF* and *CDLP* bounds. The revenue performance of *wAR* continues to be superior to that of *CDLP* and is comparable with that of *AF*.

Figure 3 shows a representative plot of how the marginal values of capacity obtained by the benchmark methods change over the course of the booking horizon. Recall that *wAR*, *CDLP*, and *AF* all yield a solution of the form $(\hat{\beta}, \hat{\gamma})$, where $\sum_{k=t}^{\tau} \hat{\gamma}_{i,k}$ can be interpreted as being an estimate of the marginal value

Table 2. Comparison of the Expected Revenues for the Hub-and-Spoke Test Problems with Overlapping Consideration Sets

Problem	Average	Expected revenue			% gap with *CDLP*	
(N, α)	capacity	*CDLP*	*wAR*	*AF*/*RAF*	*wAR*	*AF*/*RAF*
(4, 0.8)	21	6,862	6,828	6,835	−0.49 ×	−0.39 ⊙
(4, 1.0)	16	5,827	5,887	5,913	1.04 ✓	1.48 ✓
(4, 1.2)	14	5,515	5,584	5,650	1.24 ✓	2.45 ✓
(4, 1.6)	11	4,592	4,774	4,750	3.98 ✓	3.44 ✓
(6, 0.8)	13	6,337	6,439	6,291	1.61 ✓	−0.73 ×
(6, 1.0)	11	5,799	5,738	5,730	−1.04 ×	−1.19 ×
(6, 1.2)	9	5,147	5,367	5,236	4.26 ✓	1.71 ✓
(6, 1.6)	7	4,109	4,357	4,390	6.06 ✓	6.85 ✓
(8, 0.8)	10	5,554	5,591	5,557	0.67 ✓	0.05 ⊙
(8, 1.0)	8	4,803	4,894	4,887	1.90 ✓	1.74 ✓
(8, 1.2)	7	4,267	4,384	4,370	2.73 ✓	2.41 ✓
(8, 1.6)	5	3,528	3,674	3,641	4.13 ✓	3.19 ✓
				Average	2.17	1.75

Table 3. Comparison of the Upper Bounds for the Hub-and-Spoke Test Problems with Disjoint Consideration Sets

Problem	Average	Upper bound			% gap with *CDLP*	
(N, α)	capacity	*CDLP*	*wAR*	*AF/RAF*	*wAR*	*AF/RAF*
(4, 0.8)	21	7,180	7,176	7,155	0.06	0.35
(4, 1.0)	16	6,462	6,377	6,352	1.31	1.70
(4, 1.2)	14	6,138	6,053	6,027	1.38	1.81
(4, 1.6)	11	5,389	5,304	5,277	1.57	2.08
(6, 0.8)	13	6,918	6,891	6,860	0.39	0.84
(6, 1.0)	11	6,357	6,241	6,205	1.83	2.39
(6, 1.2)	9	5,799	5,683	5,654	2.00	2.50
(6, 1.6)	7	4,796	4,704	4,672	1.91	2.57
(8, 0.8)	10	6,040	5,992	5,959	0.79	1.33
(8, 1.0)	8	5,460	5,328	5,288	2.43	3.15
(8, 1.2)	7	4,993	4,857	4,817	2.73	3.52
(8, 1.6)	5	4,243	4,129	4,089	2.70	3.63
			Average	1.59	2.16	

of resource i at time t. The marginal values of capacity are used in (8) to obtain a control policy. *CDLP* yields static marginal values, and the *wAR* and *AF* marginal values change with time. The *wAR* and *AF* marginal values start decreasing toward the end of the booking horizon, reflecting the perishability of the resources. Consequently, we expect the controls based on them to have superior revenue performance.

As another robustness check, we compare performance of the dynamic programming decomposition approaches based on *CDLP* and *wAR*. Liu and van Ryzin (2008) describe how the *CDLP* dual solution can be used to decompose the network problem into a number of single resource problems, and Zhang and Adelman (2009) show that this approach obtains a bound that 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); we omit the details. Table 5 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 5 give the upper bounds obtained by *DP* − *CDLP* and *DP* − *wAR*, respectively, and the last column gives the percentage gap in the upper bounds obtained by *DP* − *CDLP* and *DP* − *wAR*. The results are in line with the earlier observations. *DP* − *wAR* generates bounds that are, on average, 1.24% tighter that *DP* − *CDLP*, and we observe gaps as high as 4.4%. It is also worthwhile noting that, in many cases, the *wAR* bound (from Table 3) is itself tighter than *DP* − *CDLP*.

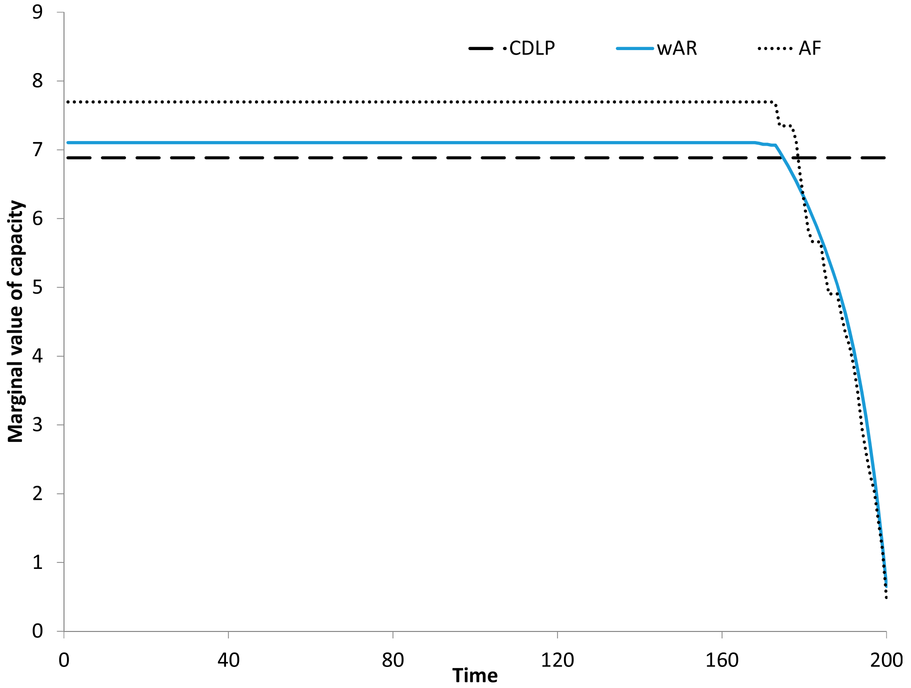
6.2.4. Larger Real-World Networks. Finally, to understand how the performance of the solution methods scales with the size of the problem, we test them on a larger flight network with a longer booking horizon. Our larger network is based on that of a European carrier network and has 30 flight legs that connect around 125 origin–destination pairs. There are, on average, six fare-products that are offered between each origin–destination pair so that the total number of fare-products is 752. There are, on average, two customer segments interested in the fare-products between each origin–destination pair, and their consideration sets overlap so that the total number of customer segments is 402. We set the fares and flight capacities according to the given data and use the observed demand for a fare-product as a proxy for its preference weight. Our base case has a booking horizon of length $\tau = 640$ periods. We vary the length of the booking horizon and proportionally scale the flight-leg capacities to obtain different test problems. In particular, we consider $\tau \in \{160, 320, 480, 640\}$ in our computational experiments.

Table 6 shows the upper bounds and the computational times for the three solution methods for the large flight network. The first column gives the length of the booking horizon. The second column gives the minimum, maximum, and average flight leg capacity in the network. The third, fourth, and fifth columns, respectively, give the upper bounds obtained by *CDLP*, *wAR*, and *AF*. The next two columns, respectively, give the percentage gap between the upper bounds obtained by *wAR* and *AF* relative to the *CDLP* bound. The last three columns, respectively, give the CPU seconds required by *CDLP*, *wAR*, and *AF*. All of our computational experiments are carried out on a Xeon E5 desktop, and we use CPLEX 12.6 to solve all LPs. We solve *CDLP* and *AF* by column generation to within 1% of optimality and report the corresponding solution times. We solve *wAR* (with product-cut equalities) to optimality as it has a compact formulation. We see

Table 4. Comparison of the Expected Revenues for the Hub-and-Spoke Test Problems with Disjoint Consideration Sets

Problem	Average	Expected revenue			% gap with *CDLP*	
(N, α)	capacity	*CDLP*	*wAR*	*AF/RAF*	*wAR*	*AF/RAF*
(4, 0.8)	21	5,755	5,748	5,744	−0.13 ⊙	−0.19 ⊙
(4, 1.0)	16	5,263	5,242	5,305	−0.39 ⊙	0.80 ✓
(4, 1.2)	14	5,056	5,080	5,136	0.47 ⊙	1.57 ✓
(4, 1.6)	11	4,413	4,570	4,580	3.56 ✓	3.78 ✓
(6, 0.8)	13	5,487	5,531	5,473	0.81 ✓	−0.25 ⊙
(6, 1.0)	11	5,047	5,127	5,098	1.58 ✓	1.00 ✓
(6, 1.2)	9	4,665	4,764	4,760	2.12 ✓	2.02 ✓
(6, 1.6)	7	3,824	4,101	4,075	7.23 ✓	6.56 ✓
(8, 0.8)	10	4,829	4,888	4,862	1.22 ✓	0.69 ✓
(8, 1.0)	8	4,343	4,434	4,456	2.09 ✓	2.61 ✓
(8, 1.2)	7	3,969	4,091	4,125	3.08 ✓	3.93 ✓
(8, 1.6)	5	3,384	3,579	3,570	5.77 ✓	5.49 ✓
				Average	2.28	2.34

Figure 3. (Color online) Marginal Values of Capacity Obtained by *CDLP*, *wAR*, and *AF* as a Function of Time for the Hub-and-Spoke Test Problem with Disjoint Consideration Sets and Problem Parameters (6, 1.6)



that the gap between the *CDLP* and *AF* upper bounds shrinks as we have longer booking horizons and larger flight capacities. This is in line with the result in Kunnumkal and Talluri (2016), who show that the *AF* bound is within a factor of $1 + 1/min_i\{r_i^1\}$ of *CDLP*, where r_i^1 is the capacity of flight leg i. Because the *wAR* bound is sandwiched between the *CDLP* and *AF* bounds, the improvements from the *wAR* upper bound also tend to be relatively small when the capacities are large. Indeed, we see that the benefits of *wAR* (in terms of both the upper bound and the computation time) are the greatest for the test problems with a relatively smaller number of time periods and flight-leg capacities.

Inspired by the observations in the preceding paragraph, we test the effect on the revenue performance by switching to *wAR* toward the end of the booking horizon. In particular, we consider our base case test problem with a booking horizon of length $\tau = 640$ periods. For each sample path, we solve *CDLP* at the start of the booking horizon and use the *CDLP* control policy up to time period $t = 320$. At that point, we switch to *wAR* and use the *wAR* control policy for the remaining time periods. Table 7 shows the expected revenue obtained by this hybrid control policy and benchmarks it with other control policies. The first column in Table 7 describes the control policy using a pair in which the first element denotes the solution method used to obtain the controls for time periods 1–320 and the second element denotes the solution method used to obtain the controls for time periods 321–640. So $(CDLP, wAR)$ refers to the control policy described here. On the other hand, $(CDLP, CDLP)$ refers to a control policy that uses *CDLP* controls for time periods 1–320, resolves *CDLP* at time $t = 320$, and uses the refreshed *CDLP* solution to make the decisions for the remaining time periods. We use Φ to indicate that we do not refresh the controls at the halfway point $(t = 320)$. So, for example, $(CDLP, \Phi)$ refers to a control policy in which we solve *CDLP* only at the beginning of the booking horizon. The second column gives the expected revenues obtained by the control policies, and

Table 5. Comparison of the Upper Bounds Obtained by the Dynamic Programming Decomposition Approaches for the Hub-and-Spoke Test Problems with Disjoint Consideration Sets

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
		Average	1.24

Table 6. Comparison of the Upper Bounds and Computation Times for the Large Network with 30 Flight-Legs

Problem	Cap	Upper bound			% gap with *CDLP*		CPU seconds		
τ	(minimum, maximum, average)	*CDLP*	*wAR*	*AF*/*RAF*	*wAR*	*AF*/*RAF*	*CDLP*	*wAR*	*AF*/*RAF*
160	(3, 9, 5)	65,530	64,305	62,536	1.87	4.57	720	129	2,727
320	(5, 17, 10)	127,844	126,979	124,863	0.68	2.33	1,566	670	4,730
480	(8, 25, 14)	188,880	188,153	185,876	0.38	1.59	2,648	1,566	5,208
640	(10, 33, 18)	249,638	249,236	246,483	0.16	1.26	2,961	3,017	7,400

the last column gives the percentage difference with the $(CDLP, \Phi)$ control. We use a ✓ to indicate that the revenue differences are significant at the 95% level. We observe that control policies based on *wAR* generate noticeably higher revenues. As observed previously, the revenue gaps tend to be larger than the corresponding gaps in the upper bounds. $(CDLP, wAR)$ provides about a 2% revenue boost compared with $(CDLP, CDLP)$. Therefore, switching from *CDLP* to *wAR* at the halfway mark ($t = 320$) can lead to significantly higher revenues. An added benefit is that *wAR* also tends to have significantly shorter run times when we solve it at the halfway mark.

7. Conclusions

CDLP and the affine relaxation are two methods in the literature that give upper bounds on the value function for choice network revenue management. Although *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, by analyzing the affine relaxation, we obtain a weaker but tractable approximation. We show that our approximation yields an upper bound that is in between the *CDLP* and the affine bounds. Our relaxation retains the appeal of the formulation discovered in Gallego, Ratliff, and Shebalov (2015) in that it involves solving a compact LP, eliminating the need for constraint or column generation. We extend our approximation to the mixture-of-multinomial-logits model with disjoint as well as with overlapping consideration sets. Our computational study indicates that our approximation typically produces upper bounds that are close to the affine bound (achieving nearly 75% reduction of the gap between it and the *CDLP*), have good revenue performance (obtaining, on average, above 95% of the revenues obtained by policies from the affine relaxation), and can be a tractable alternative to solving the affine relaxation with running times typically a fraction of that of the reduced affine relaxation.

Table 7. Comparison of the Expected Revenues Obtained by the Different Control Policies for the Large Network with 30 Flight-Legs and 640 Time Periods

Control policy (1–320, 321–640)	Expected revenue	% gap with $(CDLP, \Phi)$
$(CDLP, \Phi)$	222,648	
(wAR, Φ)	231,410	3.94 ✓
$(CDLP, CDLP)$	226,351	1.66 ✓
$(CDLP, wAR)$	230,892	3.70 ✓
(wAR, wAR)	234,593	5.36 ✓

References

Ahuja RK, Orlin JB, Stein C, Tarjan RE (1994) Improved algorithms for bipartite network flow. *SIAM J. Comput.* 23(5):906–933.

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

Bront JJM, Méndez-Díaz I, Vulcano G (2009) A column generation algorithm for choice-based network revenue management. *Oper. Res.* 57(3):769–784.

Chaneton J, Vulcano G (2011) Computing bid-prices for revenue management under customer choice behavior. *Manufacturing Service Oper. Management* 13(4):452–470.

Gallego G, Ratliff R, Shebalov S (2015) A general attraction model and sales-based linear program for network revenue management under customer choice. *Oper. Res.* 63(1):212–232.

Gallego G, Iyengar G, Phillips R, Dubey A (2004) Managing flexible products on a network. Technical Report TR-2004-01, Columbia University, New York.

Gallego G, Li A, Truong V-A, Wang X (2016) Online personalized resource allocation with customer choice. Technical Report, Columbia University, New York.

Golrezaei N, Nazerzadeh H, Rusmevichientong P (2014) Real-time optimization of personalized assortments. *Management Sci.* 60(6): 1532–1551.

Grötschel M, Lovász L, Schrijver A (1988) *Geometric Algorithms and Combinatorial Optimization* (Springer, Berlin Heidelberg).

Hosseinalifam M, Marcotte P, Savard G (2016) A new bid price approach to dynamic resource allocation in network revenue management. *Eur. J. Oper. Res.* 255(1):142–150.

Kunnumkal S, Talluri K (2016) A note on relaxations of the choice network revenue management dynamic program. *Oper. Res.* 41(1):158–166.

Liu Q, van Ryzin GJ (2008) On the choice-based linear programming model for network revenue management. *Manufacturing Service Oper. Management* 10(2):288–310.

McFadden D, Train K (2000) Mixed MNL models for discrete response. *J. Appl. Econometrics* 15(5):447–470.

Meissner J, Strauss AK (2012) Network revenue management with inventory-sensitive bid prices and customer choice. *Eur. J. Oper. Res.* 216(2):459–468.

Meissner J, Strauss AK, Talluri KT (2013) An enhanced concave programming method for choice network revenue management. *Production Oper. Management* 22(1):71–87.

Peeters R (2003) The maximum edge biclique problem is NP-complete. *Discrete Appl. Math.* 131(3):651–654.

Rusmevichientong P, Shmoys D, Tong C, Topaloglu H (2014) Assortment optimization under the multinomial logit model with random choice parameters. *Production Oper. Management* 23(11): 2023–2039.

Strauss AK, Talluri KT (2017) Tractable consideration set structures for assortment optimization and network revenue management. *Production Oper. Management* 26(7):1359–1368.

Talluri KT (2014) New formulations for choice network revenue management. *INFORMS J. Comput.* 26(2):401–413.

Talluri KT, van Ryzin GJ (2004) *The Theory and Practice of Revenue Management* (Kluwer, New York).

Vossen TWM, Zhang D (2015) Reductions of approximate linear program for network revenue management. *Oper. Res.* 63(6):1352–1371.

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