OPERATIONS RESEARCH
Vol. 57, No. 3, May–June 2009, pp. 671–684
ISSN 0030-364X | EISSN 1526-5463 | 09 | 5703 | 0671

DOI 10.1287/opre.1080.0610

Dynamic Capacity Management with Substitution

Robert A. Shumsky
Tuck School of Business, Dartmouth College, Hanover, New Hampshire 03755, robert.shumsky@dartmouth.edu

Fuqiang Zhang
Olin Business School, Washington University in St. Louis, St. Louis, Missouri 63130, fzhang22@wustl.edu

We examine a multiperiod capacity allocation model with upgrading. There are multiple product types, corresponding to multiple classes of demand, and the firm purchases capacity of each product before the first period. Within each period, after demand arrives, products are allocated to customers. Customers who arrive to find that their product has been depleted can be upgraded by at most one level. We show that the optimal allocation policy is a simple two-step algorithm: First, use any available capacity to satisfy same-class demand, and then upgrade customers until capacity reaches a protection limit, so that in the second step the higher-level capacity is rationed. We show that these results hold both when all capacity is salvaged at the end of the last demand period as well as when capacity can be replenished (in the latter case, an order-up-to policy is optimal for replenishment). Although finding the optimal protection limits is computationally intensive, we describe bounds for the optimal protection limits that take little effort to compute and can be used to effectively solve large problems. Using these heuristics, we examine numerically the relative value of strictly optimal capacity and dynamic rationing, the value of perfect demand information, and the impact of demand and economic parameters on the value of optimal substitution.

Subject classifications: inventory/production: uncertainty, stochastic; multi-item; approximations/heuristics.
Area of review: Manufacturing, Service, and Supply Chain Operations.
History: Received February 2007; revisions received June 2007, November 2007; accepted January 2008. Published online in *Articles in Advance* March 11, 2009.

1. Introduction

Many manufacturing and service firms use capacity or inventory flexibility to meet uncertain demand from multiple classes of customers. When capacity for a particular product has been exhausted, demand for that product may be met by a substitute product. For many applications, the assignment of capacity to customers is complicated by the fact that demand arrives over time and capacity must be allocated before demand is fully known.

Consider a manufacturer of personal computers that operates an assemble-to-order system. The firm maintains stocks of critical components such as hard disk drives, which are available in many sizes. If a particular customer's requested size is unavailable, the firm may choose to upgrade the customer to a more expensive size (the customer may or may not be aware of the upgrade, depending upon how the disk drive is formatted before shipping). Customers arrive over time, and therefore the disk drive allocation decision must be made when the demand for each type of drive is still uncertain. Reorder cycles may be lengthy, so that some disk drives must be allocated to customers before stocks can be replenished. Bassok et al. (1999) describe and provide references for similar upgrade problems from the semiconductor and steel industries, and parallel problems are found in the service industries. Car rental agencies upgrade customers to more expensive cars, hotels allocate various grades of rooms among customers, and airlines upgrade customers from economy to business class or first-class seats.

Here we analyze dynamic multiproduct capacity models in which demand arrives in discrete periods. Throughout this paper, we use the term "capacity," although the products may be interpreted as either service capacities or tangible inventory. For this problem, we ask how much capacity should be acquired and how that capacity should be distributed among customers as demand is realized? Our models have the following attributes:

1. Initially, we assume that there is a single opportunity to invest in capacity before any demand is realized. We then consider a more general model with repeated capacity replenishment.

2. The time after the initial purchase (or between replenishments) is broken into a finite number of demand periods (T), and the decision maker allocates capacity to customers after observing demand within each period.

3. Demand that is not satisfied in each period is lost (there is no backlogging).

4. Demand for a product can be met by a product from the next-higher class (for example, a computer manufacturer's demand for a hard disk drive may be met with a larger drive).

5. Capacity may be rationed so that the firm may choose not to allocate high-class capacity to a lower-class customer.

Our models can be seen as an extension of the single-period multiproduct newsvendor models of Bassok et al. (1999), Netessine et al. (2002), and others, to an environment with multiperiod demand. Another model with this flavor is the "newsvendor network" of Van Mieghem and Rudi (2002), but their model allows the firm to replenish capacity between every demand period, whereas in ours replenishment occurs every T periods. Our models are also similar to yield management models in which a firm must find optimal rules for rationing capacity among customer classes. Therefore, this paper can also be seen as a generalization of the yield management problem to include multiple types of capacity as well as the ability to upgrade customers to a higher-capacity class.

After reviewing the literature, in §3 we describe our basic model with a single opportunity for capacity investment and show that a single-period static formulation provides an upper bound on the expected profit of the dynamic model. In §4, we prove that a threshold-rationing scheme is the optimal policy among all possible policies and describe a necessary and sufficient condition for the optimal level of rationing (the number of units to ration is sometimes called the *protection limit*). In §5, we extend the results to the setting in which capacity can be replenished after each sequence of T demand periods. We show that if demands between replenishment opportunities are independent and identically distributed, then the threshold-rationing policy is optimal between replenishments and a stationary order-up-to policy is optimal for each replenishment.

The remainder of this paper focuses on the characteristics, calculation, and relative impact of the threshold-rationing policy. In §6, we show that the protection limit of each capacity class is decreasing as time increases and is decreasing in the capacity level of any of the available products.[1] We also derive complementary pairs of lower and upper bounds on the optimal protection limits that grow progressively tighter as the computational effort needed to calculate each pair of bounds increases. Section 6 then describes numerical experiments demonstrating that over a wide range of parameters, the bounds are extremely tight. In fact, bounds based only upon the capacity level of one adjacent product allow us to estimate protection levels that are extremely close to optimal, and these bounds can be calculated quickly, even for large models with many products.

In §7, we employ these bounds to generate numerical examples using reasonably large capacity quantities and time periods. Using these examples, we first compare the optimal capacities for the single-period static model and the dynamic model described in this paper. We find that the differences between the optimal static and dynamic capacities are usually small, and when they are not, the difference in profits due to using capacity that is optimal for the static model (rather than the dynamic model) for the dynamic case is negligible. In that section, we also numerically examine the value of using optimal rationing, rather than two simple heuristics: (i) upgrading with no rationing, or (ii) no upgrading. We explore how the value of optimal rationing versus these heuristics changes with the availability of advance demand information, economic parameters (e.g., contribution margins and initial costs), and demand parameters (e.g., the variance and within-period correlations of the demand). Finally, in §8 we describe future research.

2. Related Literature

There are many models in the literature that capture a subset of the five characteristics described above, but none, to our knowledge, addresses all five (see the review article by Van Mieghem 2003 for a more complete characterization of the literature on capacity investment and management). Some researchers have focused on single-period "multidimensional newsvendor models," a term used in Van Mieghem (1998). For example, Bassok et al. (1999) propose a general multiproduct inventory model to study the benefits of substitution. Pasternack and Drezner (1991) find the optimal stocking policy for goods with stochastic demand and substitution in both the "up" and "down" directions. Fine and Freund (1990) and Van Mieghem (1998) study optimal levels of flexible and dedicated production capacities. Netessine et al. (2002) study the value of single-level upgrades with an emphasis on the impact of demand correlation on the optimal investment levels. In all of these papers, the firm purchases inventory before demand is realized and distributes the inventory to customers after observing all demand.

Tomlin and Wang (2008) consider a firm that sells two vertically differentiated products to two classes of customers. Both supply and demand are uncertain. The utility-maximizing customers may choose to purchase a second-choice product if the first-choice product is not available. They examine the firm's optimal pricing and inventory allocation policies. Again, theirs is a single-period model in which all allocation decisions are made either before demand is realized or after all demand is realized.

As in our paper, Van Mieghem and Rudi (2002) present a multidimensional newsvendor model that also incorporates multiperiod demand. However, their model allows the firm to replenish inventory between each and every demand period. For the applications we have in mind, adjustments in inventory occur over a longer time scale than the within-period rationing and allocation decisions, so that the firm must find the optimal allocation, given only the inventory it purchases every T demand periods. The firm's inability to replenish inventory between demand periods also distinguishes our work from the literature on multiperiod inventory models with transshipment, such as Karmarkar (1981), Robinson (1990), Archibald et al. (1997), and Axsäter (2003).

The literature on yield management does focus on environments in which capacity-sizing decisions are made and then capacity must be allocated as demand arrives over

time. See McGill and van Ryzin (1999) and Talluri and van Ryzin (2004) for surveys of this literature. Papers by Curry (1990), Wollmer (1992), and Brumelle and McGill (1993) characterize the optimal rationing policy for an airline seat allocation problem in which a fixed seat capacity must satisfy demand for multiple fare classes. The following papers generalize these results by incorporating cancellations and/or overbooking: Bitran and Gilbert (1996), Subramanian et al. (1999), and Zhao and Zheng (2001). Savin et al. (2005) describe a model that is tailor-made for studying the renting or leasing of capital equipment to multiple customer classes. They formulate the problem as a queueing control problem and allow the rental period to be stochastic rather than uniformly fixed. In all of these papers, there is a single type of resource—a coach seat on a single-leg flight or a single type of rental car—so that there is no discussion of "upgrades."

There are a few papers in the yield management area that do address the issue of inventory substitution. Alstrup et al. (1986) study a dynamic overbooking problem with two inventory classes and two-way substitution. Karaesmen and van Ryzin (2004) examine a more general overbooking problem with multiple substitutable inventory classes. Both papers formulate a two-stage model: first a booking stage, and then an allocation stage after all demand is realized. Although substitution is allowed during the second, the allocation stage, there is no substitution as demand arrives during the booking stage. In our model, substitution may occur during each demand period.

Researchers have addressed the topics of substitution and rationing in the context of production and inventory control. The model of Topkis (1968) is similar to the problem described in this paper. Topkis also assumes a given initial level of inventory and characterizes the optimal rationing policy as a set of "critical rationing levels," although his model assumes a single type of inventory and multiple demand classes. Topkis shows that, under certain conditions, the critical rationing levels decline over time (analogous results for our model are derived in §6). Articles by Ha (1997a, b, and 2000) consider make-to-stock production systems with several demand classes. These papers show that the optimal stock-rationing policy can be characterized by a sequence of production limits and storage levels that are also monotone in customer class. Research by de Véricourt et al. (2001, 2002) describes the benefits of optimal stock allocation for these make-to-stock systems and characterizes techniques to calculate optimal parameters for the allocation decision. Motivated by a study of the military logistics systems, Deshpande et al. (2003) analyze a service parts inventory system with two demand classes characterized by different arrival rates and shortage costs. It is shown that a static rationing policy is close to optimal in situations typical of the military and high-technology industries. Frank et al. (2003) consider an inventory system in which replenishment is possible and stock may be protected from stochastic demand while it is used to fill higher-priority deterministic demand. Ding et al. (2006) study an inventory system with multiple customer classes and partial backlogging. The likelihood of backlogging is a function of the discount offered to customers. They determine the optimal discounts to offer and characterize the optimal allocation policy for such an inventory system. All of these papers consider single-item production systems, whereas we examine a system with multiple products and substitution.

Kapuscinski and Tayur (2000) study a dynamic capacity reservation problem in a make-to-order environment, in which demands are classified by their waiting-time sensitivities. Eynan (1999) examines the benefits of inventory pooling and shows that these benefits are not significantly reduced by the "cannibalization" of inventory by low-margin customers, but he does not consider the benefits of a rationing policy. Again, these papers focus on problems involving a single product and multiple demand classes, whereas we consider multiple products and demand classes.

3. The Model

In this section, we describe the products offered by the firm, the customer demand classes, the cost and demand parameters (along with a few assumptions about these parameters), and the firm's decision variables. At the end of the first subsection, we present the problem formulation, whereas in the second subsection, we present two related formulations and bounds on the objective function value based on the related formulations.

3.1. Problem Description

Consider a firm that serves N classes of demand by providing N types of products indexed by $j = 1, 2, \ldots, N$. Product quality *decreases* as index j increases, so that product j can be used to satisfy a customer of class i as long as $j \leqslant i$. This is often called "one-way substitution" and is a common practice in many manufacturing and service applications. Products with superior quality are acceptable to customers who request an inferior product, but not vice versa.

Time periods are indexed by t, and demand arrives in each of the $t = 1, \ldots, T$ periods, where T is finite. Demand is independent between periods, although product demands within a period need not be independent. Let $\mathfrak{D}^t = (\mathfrak{d}_1^t, \mathfrak{d}_2^t, \ldots, \mathfrak{d}_N^t)$ denote the demand in period t, an N-variate random variable. Let $\mathbf{D}^t = (d_1^t, d_2^t, \ldots, d_N^t)$ denote all realized demand in period t (we will use boldface characters to represent vectors). Initially, we assume that each period's demand for a particular product is a nonnegative real number, so that $\mathbf{D}^t \in \mathbb{R}_N^+$. We will assume, however, that demand is integer valued when deriving bounds and heuristics in §6 and when conducting numerical experiments in §§6 and 7. Here we also assume that any capacity left over after time T is salvaged. In §5, we consider a model in which capacity is held over for use during another set of demand periods.

Let c_j^a be the purchase cost for each unit of product j and let u_j^a be the usage cost when a unit is sold. That is, the firm pays c_j^a upfront, whether the capacity is sold or not, and only pays u_j^a when a unit of capacity is sold to a customer. Let l_j be the salvage value of product j after period T. One method for assessing total salvage value is to explicitly multiply l_j by the leftover capacity of j after the last demand period. For most of the analysis below, however, we will work with an equivalent formulation in which the salvage value is assessed indirectly by incorporating it into an effective unit purchase cost $c_j = c_j^a - l_j$ and an effective usage cost $u_j = u_j^a + l_j$.

When a customer arrives, she pays p_j for a product of type j. The firm may also pay a penalty cost v_i if it cannot provide a product to a customer of type i. We assume that demand is not backlogged, revenues and costs remain constant over time, and that the time horizon is sufficiently short so that there is no discounting of costs or revenues across demand periods (in §5, we will allow discounting across replenishment intervals).

Let α_{ij} be the unit contribution margin for satisfying a class i customer with product j. We make the following assumptions:

Assumption 1 (A1). $\alpha_{ij} = p_i + v_i - u_j > 0$ *if* $j \leqslant i \leqslant j+1$; $\alpha_{ij} < 0$ *otherwise.*

Assumption 2 (A2). $p_1 + v_1 > p_2 + v_2 > \cdots > p_N + v_N$.

Assumption 3 (A3). $u_1 > u_2 > \cdots > u_N$.

Assumption (A1) states that only one-step upgrading is profitable. In practice, the contribution margin accrued from multistep upgrades is often small, or negative. From a network design perspective, single-step upgrading can often deliver most of the benefits of more complex substitution schemes. For example, when quantifying the value of flexible production capacity, Jordan and Graves (1995) find that a chain of factories, each with a single link to its neighbor (each plant i can produce products i and $i+1$), yields nearly the same sales as a chain of factories with full flexibility (each plant i can produce all products). Here we analyze a similar chain of flexible capacity, although in our model product N cannot be used to upgrade a customer who desires product 1, so that we are missing the last "link" in the chain. Assumptions (A2) and (A3) state that both the revenue $(p_j + v_j)$ and the usage cost u_j decrease in index j. That is, products with higher quality have higher revenues and usage costs. These assumptions imply that $\alpha_{jj} > \alpha_{kj}$ for all $j \neq k$, so that the maximum margin for product j is achieved by selling to customers of class j.

Now we describe the state space of the optimization problem and the firm's decision variables. Let $\mathbf{X}^t = (x_1^t, x_2^t, \ldots, x_N^t)$, $\mathbf{X}^t \in \mathbb{R}_N^+$, be the vector of capacities at the beginning of period t, $t = 1, 2, \ldots, T$. After demand $\mathbf{D}^t$ appears, the firm must make capacity allocation decisions. Let $\Pi^{\mathrm{DYN}}(\mathbf{X}^1)$ be the profit function for our model. We formulate this problem as a dynamic program with $T+1$ steps. In period 0 the firm determines the initial capacity $\mathbf{X}^1$, whereas in periods 1 through T the firm allocates its capacity to maximize its revenue.

Dynamic Substitution Model (DYN)
Period 0:

$$\max_{\mathbf{X}^1 \in \mathbb{R}_N^+} \Pi^{\mathrm{DYN}}(\mathbf{X}^1) = \max_{\mathbf{X}^1 \in \mathbb{R}_N^+} \left\{ \Theta^1(\mathbf{X}^1) - \sum_j c_j x_j^1 \right\}. \tag{1}$$

Period t $(1 \leqslant t \leqslant T)$:

$$\Theta^t(\mathbf{X}^t) = \mathop{\mathrm{E}}_{\mathfrak{D}^t} \left\{ \max_{\substack{\mathbf{Y}^t + \mathbf{X}^{t+1} = \mathbf{X}^t \\ \mathbf{Y}^t \in \mathbb{R}_N^+,\, \mathbf{X}^{t+1} \in \mathbb{R}_N^+}} [H^t(\mathbf{Y}^t \mid \mathbf{D}^t) + \Theta^{t+1}(\mathbf{X}^{t+1})] \right\}, \tag{2}$$

where

$$H^t(\mathbf{Y}^t \mid \mathbf{D}^t) = \max_{\bar{\mathbf{Y}}^t} \left[\sum_{i,j} \alpha_{ij} y_{ij}^t - \sum_i v_i d_i^t \right], \tag{3}$$

$$\sum_j y_{ij}^t \leqslant d_i^t, \quad i = 1, 2, \ldots, N, \tag{4}$$

$$\sum_i y_{ij}^t \leqslant y_j^t, \quad j = 1, 2, \ldots, N, \tag{5}$$

$$y_{ij}^t \in \mathbb{R}^+, \quad i, j = 1, 2, \ldots, N, \tag{6}$$

and $\Theta^{T+1} \equiv 0$. In this formulation, $\mathbf{Y}^t$ is a vector of capacity offered for sale in period t and $y_j^t \equiv (\mathbf{Y}^t)_j$ is the capacity of product j offered for sale. The vector $\mathbf{X}^{t+1}$ is the capacity held over to the next period and the constraints $\mathbf{Y}^t + \mathbf{X}^{t+1} = \mathbf{X}^t$, $\mathbf{Y}^t \in \mathbb{R}_N^+$, and $\mathbf{X}^{t+1} \in \mathbb{R}_N^+$ ensure that the sum of the capacity offered for sale in period t and the capacity held over to the next period do not exceed $\mathbf{X}^t$. The value of $H^t(\mathbf{Y}^t \mid \mathbf{D}^t)$ is the revenue from the single-period capacity problem with substitution, given realized demand $\mathbf{D}^t$. Within problem H^t, $y_{ij}^t \in \mathbb{R}^+$ is the quantity of product j sold to class i demand and $\bar{\mathbf{Y}}^t = (y_{ij}^t)$ is an allocation matrix for period t. Inequality (4) is period t's demand constraint and (5) is period t's supply constraint, i.e., the firm cannot sell more capacity than the capacity offered in period t.

There are two details of the formulation that require further discussion. First, in this formulation there is a distinction between *offered* capacity $(\mathbf{Y}^t)$ and *sold* capacity (y_{ij}^t). Therefore, it is possible that in the optimal allocation $\bar{\mathbf{Y}}^*$, $\sum_i y_{ij}^* < y_j^t$. This implies that some offered capacity is thrown away—it does not generate revenue and is not held over to the next period. We will see below, however, that there is at least one optimal solution to Θ^t in which all offered capacity is sold. (It is true that adding a constraint $\sum_i y_{ij}^t = y_j^t$ to the formulation would eliminate this complication, but the equality constraint would make it more difficult to apply useful results from concave analysis.)

The second detail for discussion is that we have chosen to use positive real numbers to model capacity. In practice, capacity is often discrete, and demand follows a discrete distribution. In the related literature, capacity has been

modeled as discrete (e.g., Wollmer 1992) or continuous (e.g., Curry 1990). Following the approach of Brumelle and McGill (1993), we begin with a continuous formulation of the problem, and by using subdifferential optimization we show that a rationing algorithm is optimal for either discrete or continuous demand distributions. Then, in Proposition 3, we show that if capacity and demand are discrete (integer valued), then the optimal capacity allocation procedure preserves integrality. In §6, we derive bounds and heuristics for solving large discrete problems.

3.2. Related Models

If we let $T = 1$, model DYN collapses into the single-period (or *static*) model studied by Bassok et al. (1999), Netessine et al. (2002), and others (we will use the acronym STC to refer to this model). For the sake of comparison, we transform the single-period model into an equivalent model with T periods, and we assume that demand arrives in each period as it does in the dynamic model. However, in STC, resources are allocated after all demand is observed. This transformation will help us to compare the performance of STC and DYN, given the same demand. In the following formulation, let $\mathbf{X}$ denote the vector of initial capacities and $\Pi^{\mathrm{STC}}(\mathbf{X})$ the profit function.

Single-Period Substitution Model (STC)

$$\max_{\mathbf{X}\in\mathbb{R}_N^+} \Pi^{\mathrm{STC}}(\mathbf{X}) = \max_{\mathbf{X}\in\mathbb{R}_N^+} \mathop{\mathrm{E}}_{\{\mathfrak{D}^1,\mathfrak{D}^2,\ldots,\mathfrak{D}^T\}}\left[\Theta(\mathbf{X}) - \sum_j c_j x_j\right], \tag{7}$$

where

$$\Theta(\mathbf{X}) = \max_{\mathbf{Y}} \quad \left[\sum_{i,j} \alpha_{ij} y_{ij} - \sum_i v_i \sum_t d_i^t\right] \tag{8}$$

$$\text{s.t.} \quad \sum_j y_{ij} \leqslant \sum_t d_i^t, \quad i = 1, 2, \ldots, N, \tag{9}$$

$$\sum_i y_{ij} \leqslant x_j, \quad j = 1, 2, \ldots, N, \tag{10}$$

$$y_{ij} \in \mathbb{R}^+, \quad i, j = 1, 2, \ldots, N. \tag{11}$$

We also consider the simplest benchmark model, a model without product substitution. This is equivalent to N independent newsvendors (NV). As in DYN and STC, we consider demand that arrives sequentially over T periods. Given independent newsvendors, however, it does not matter whether the allocation of capacity occurs as the demand arrives (as in DYN) or after the Tth period (as in STC). In either case, the firm determines the optimal capacity x_j according to the newsvendor fractile and then sells the maximum amount of capacity possible.

Independent Newsvendor Model (NV)

$$\max_{\mathbf{X}\in\mathbb{R}_N^+} \Pi^{\mathrm{NV}}(\mathbf{X}) = \max_{\mathbf{X}\in\mathbb{R}_N^+} \sum_j \left\{ \mathop{\mathrm{E}}_{\{\mathfrak{D}^1,\mathfrak{D}^2,\ldots,\mathfrak{D}^T\}} \left[\alpha_{jj} \min\left(x_j, \sum_t d_j^t\right) - v_j\left(\sum_t d_j^t\right)\right] - c_j x_j \right\}. \tag{12}$$

In the following proposition, we compare the profits of these three models. When omitted, proofs of propositions and lemmas can be found in the electronic companion that is part of the online version at http://or.journal.informs.org/.

Proposition 1. $\Pi^{\mathrm{NV}}(\mathbf{X}) \leqslant \Pi^{\mathrm{DYN}}(\mathbf{X}) \leqslant \Pi^{\mathrm{STC}}(\mathbf{X})$.

It follows that $\Pi^{\mathrm{DYN}}(\mathbf{X}^{\mathrm{DYN}}) \leqslant \Pi^{\mathrm{STC}}(\mathbf{X}^{\mathrm{STC}})$, where $\mathbf{X}^{\mathrm{DYN}}$ and $\mathbf{X}^{\mathrm{STC}}$ are the optimal initial capacity vectors.

4. The Optimal Policy: Parallel Allocation and Then Rationing

Before explicitly describing the optimal rationing policy for DYN, we first establish two properties for $\Theta^t(\mathbf{X})$, monotonicity (Lemma 1) and concavity (Lemma 2). Note that the proofs of both lemmas do not require the single-step upgrading assumption, so that these monotonicity and concavity results hold under a general upgrading structure.

Lemma 1. $\Theta^t(\mathbf{X})$ *is monotonically increasing in* $\mathbf{X}$.

An immediate implication of Lemma 1 is that there exists an optimal allocation where the offered capacity in each period is fully utilized because otherwise one may improve the profit by passing the unused capacity to the next period. This implication allows us to restrict our attention to the subset of solutions to Θ^t in which $(\mathbf{Y}^*)_j = \sum_i (\overline{\mathbf{Y}}^*)_{ij}$, where $\mathbf{Y}^*$ is the optimal offered vector and $\overline{\mathbf{Y}}^*$ is the optimal allocation matrix (the quantity actually sold). All of the following results will also hold if we admit optimal solutions in which $(\mathbf{Y}^*)_j > \sum_i (\overline{\mathbf{Y}}^*)_{ij}$, but the notation is more complex and the results are not any more informative.

Lemma 2. $\Theta^t(\mathbf{X})$ *is concave in* $\mathbf{X}$.

Lemma 2 implies that $\Pi^{\mathrm{DYN}}(\mathbf{X}^1)$ is also concave.

The analysis in the remainder of this section proves that at any time period t it is optimal to first satisfy demand from class i with capacity from class i and then to consider upgrades, where upgrading is limited by some threshold value. More formally, suppose that capacity $\mathbf{X}^t = (x_1^t, x_2^t, \ldots, x_N^t)$ is available at the beginning of period t. Define $\delta_k\Theta^t = [\delta_k^+\Theta^t, \delta_k^-\Theta^t]$ as the subdifferential of Θ^t with respect to the capacity of product k, where $\delta_k^+\Theta^t$ and $\delta_k^-\Theta^t$ are the right and left derivatives, respectively. Also define $x \wedge y = \min(x, y)$. Proposition 2 will show that the following algorithm maximizes $\Theta^t(\mathbf{X}^t)$ (henceforth, we will refer to this procedure as the PRA, for the Parallel assignment then Rationing Algorithm).

Step 1 (parallel assignment). Let $y_{ii}^t = d_i^t \wedge x_i^t$, $i = 1, 2, \ldots, N$. Satisfy as much class i demand with capacity of product i as possible.

Step 2 (upgrading and rationing). Let $\mathbf{N}^t$ be the difference between parallel demand and capacity:

$$\mathbf{N}^t = (n_1^t, n_2^t, \ldots, n_N^t) = (x_1^t - d_1^t, x_2^t - d_2^t, \ldots, x_N^t - d_N^t). \tag{13}$$

Note that n_i^t can be positive if there is excess capacity after Step 1, negative if demand exceeds capacity, or zero.

For $k = 1, \ldots, N-1$, if $n_k^t > 0$ and $n_{k+1}^t < 0$, then let $(n_k^t - \tilde{p}_k)^+$ be the maximum capacity k offered for upgrading, so that the actual amount of capacity upgraded $y_{k+1,k} = (n_k^t - \tilde{p}_k)^+ \wedge |n_{k+1}^t|$. The quantity $\tilde{p}_k$ is the *protection limit* for product k, and an optimal protection limit satisfies

$$\alpha_{k+1,k} \in \delta_k \Theta^{t+1}(n_1^t, n_2^t, \ldots, n_{k-1}^t, \tilde{p}_k). \tag{14}$$

The rationale behind the PRA is straightforward. From (A1)–(A3), we see that the contribution margin from a parallel allocation is larger than the margin from any present or future upgrade, so that in Step 1 any available capacity should be used to satisfy parallel demand. To understand Step 2, note that a unit of capacity k should be used in period t for an upgrade if the value of the upgrade, $\alpha_{k+1,k}$, is greater than the expected value of that unit in periods $t+1$ through T. Because the marginal value of capacity k in future periods declines as the quantity of capacity k rises (see Lemma 2 above), a threshold rule is optimal when choosing the number of units to upgrade.

To demonstrate rigorously that the PRA is an optimal policy, we must first derive a series of intermediate results. The following lemma establishes the general structure of the optimal policy.

LEMMA 3. *The following algorithm solves* $H^t(\mathbf{Y} \mid \mathbf{D})$:

(i) $y_{ii} = d_i \wedge y_i$, $i = 1, \ldots, N$;

(ii) $y_{i+1,i} = (d_{i+1} - y_{i+1})^+ \wedge (y_i - d_i)^+$, $i = 1, \ldots, N-1$.

Lemma 3 allows us to rewrite $H^t(\mathbf{Y} \mid \mathbf{D})$ as

$$H^t(\mathbf{Y} \mid \mathbf{D}) = \sum_{i=1}^{N} \alpha_{ii}(d_i \wedge y_i) + \sum_{i=1}^{N-1} \alpha_{i+1,i}((d_{i+1} - y_{i+1})^+ \wedge (y_i - d_i)^+). \tag{15}$$

This appears to be identical to the PRA: parallel assignment, followed by upgrading. However, we have not yet determined the optimal offered capacity $\mathbf{Y}^t$, and therefore have not demonstrated that in Step 1 of PRA all available capacity should be used to satisfy parallel demand and that in Step 2 a threshold policy is optimal. Lemma 3, however, does split the optimal policy into two simple decisions: how much capacity to offer for parallel assignment, and then how much capacity to upgrade.

Before answering these questions, the following lemma shows that, after Step 1 of the PRA, the optimization problem breaks into smaller independent "subproblems":

LEMMA 4. *Suppose that at time t after completing Step* 1 *of PRA, net capacity* $n_i^t \leqslant 0$, $i = k+1, \ldots, k+j$, *so that the capacities of these products have been depleted. Then, the optimization problem can be separated into two independent subproblems*: *an upper part consisting of products* 1 *to* $k+1$, *and a lower part consisting of products* $k+j+1$ *to* N.

In general, after parallel assignment, the global optimization problem may have been divided into numerous smaller subproblems, each defined by a series of positive net capacities (e.g., $n_i^t > 0$, $i = j, \ldots, k$) and a single depleted capacity level for the lowest product ($n_{k+1}^t \leqslant 0$). Therefore, for each subproblem created after parallel allocation, *there is only one upgrading and rationing decision to be made*: How much capacity of class k do we use for upgrades of unfilled demand from class $k+1$?

The same observation applies at the beginning of time t, before parallel assignment. The global optimization at the beginning of time t may be broken into smaller independent subproblems, with boundaries defined by depleted capacities, $x_i^t = 0$. By convention, for these subproblems we do not include the "0" capacities of the boundary products. To be explicit, define $B = \{(h_1^t, l_1^t), \ldots, (h_m^t, l_m^t)\}$ as the set of upper and lower limits for the subproblems at time t, i.e., (h_i^t, l_i^t) are the indices of the highest (smallest indexed) and lowest (largest indexed) products in the ith subproblem, so that $h_i^t \leqslant l_i^t$ and $x_j^t > 0$, $h_i^t \leqslant j \leqslant l_i^t$. Then, the profit of the remaining optimization problem at time t, $\Theta^t(\mathbf{X}^t)$ in Equation (2), can be written as the sum of the profits from the subproblems:

$$\Theta^t(\mathbf{X}^t) = \sum_{i=1}^{m} \Theta_i^t(\mathbf{X}_i^t), \tag{16}$$

where each subproblem $\Theta_i^t(\mathbf{X}_i^t)$ has the same formulation as $\Theta^t(\mathbf{X}^t)$, although the demand and capacity indices of each subproblem vary from h_i^t to l_i^t, rather than from 1 to N.

In the remainder of this section, we will derive the optimal policy for an optimization problem $\Theta^t(\mathbf{X}^t)$ with product indices $i = 1, \ldots, N$. Because the subproblems are independent, and because the objective function of the global problem is the sum of the values of the subproblems, the following results apply to any subproblem, as well as to the global optimization problem.

We now show that the PRA is an optimal policy, given all possible policies. Because demands are independent between time periods, we consider only fixed policies that depend upon the time period and capacity state, but do not depend upon observed demand realizations. Using the terminology in Porteus (1975), the set of admissible policies is defined by the constraints of Θ^t and H^t, $t = 1, \ldots, T$, and the PRA defines an admissible structured policy. Because of the capacity constraints, all value functions $\Theta^t(\mathbf{X})$ are finite for finite $\mathbf{X}$.

The following lemma establishes that the value function $\Theta^t(\mathbf{X})$ has the following three properties: (1) the PRA is an optimal policy; (2) the marginal value of one unit of capacity in the next period is at most α_{kk}, the value from a parallel assignment; (3) the value function for the next period is concave in the capacity passed on to the next period. We show that property (1) is preserved under induction: First, we show that property (2) implies that the full parallel assignment in Step 1 of the PRA is optimal. We then

invoke property (3) to show that in Step 2 of the PRA there is an optimal upgrading threshold, as defined by condition (14). Finally, because the PRA is optimal, it follows that the marginal value of a unit of capacity is bounded by α_{kk} and, from Lemma 2, the value function is concave.

LEMMA 5. *Suppose that* Θ^{t+1} *has the following properties*:
1. *The PRA solves* $\Theta^{t+1}(\mathbf{X})$;
2. $\delta_k^-\Theta^{t+1}(\mathbf{X}) \leqslant \alpha_{kk}$;
3. $\Theta^{t+1}(\mathbf{X})$ *is concave in* $\mathbf{X}$.

Then, properties (1)–(3) *hold for* Θ^t.

PROOF. Here we will sometimes write the vector $\mathbf{Y}$ as $(y_k, \mathbf{Y}_{-k})$ to emphasize the value of the vector's kth component. Define

$$\widehat{\Theta}^t(y_k, \mathbf{Y}_{-k}, \mathbf{X} \mid \mathbf{D}) = H^t(\mathbf{Y} \mid \mathbf{D}) + \Theta^{t+1}(\mathbf{X} - \mathbf{Y})$$

so that

$$\Theta^t(\mathbf{X} \mid \mathbf{D}) = \max_{\mathbf{Y} \leqslant \mathbf{X}} \widehat{\Theta}^t(y_k, \mathbf{Y}_{-k}, \mathbf{X} \mid \mathbf{D}).$$

Let $\beta_k^+\widehat{\Theta}^t$ and $\beta_k^-\widehat{\Theta}^t$ be the right and left derivatives of $\widehat{\Theta}^t$ with respect to y_k and let $\beta_k\widehat{\Theta}^t$ be the subdifferential of $\widehat{\Theta}^t$ with respect to y_k. We first prove property 1 and then show that properties 2 and 3 are preserved under optimization.

1. To show that Step 1 of the PRA is optimal, we see from Lemma 3 that we need only show that $y_k^* \geqslant (d_k \wedge x_k)$, where y_k^* is the optimal offered (and sold) capacity of product k. That is, we will show that all available capacity in $\mathbf{X}$ is available for parallel allocation in Step 1. We consider two cases, $y_k \leqslant x_k \leqslant d_k$ and $y_k \leqslant d_k < x_k$. When $y_k \leqslant x_k \leqslant d_k$,

$$\beta_k^-\widehat{\Theta}^t(y_k, \mathbf{Y}_{-k}, \mathbf{X} \mid \mathbf{D}) = \alpha_{kk} - \delta_k^-\Theta^{t+1}(\mathbf{X} - \mathbf{Y}) \tag{17}$$

$$\geqslant 0, \tag{18}$$

where Equation (17) follows from the derivative of Equation (15), and inequality (18) follows from induction Assumption (2). Inequality (18) and the constraint $\mathbf{Y} \leqslant \mathbf{X}$ imply that $y_k^* = x_k$. When $y_k \leqslant d_k < x_k$, Equation (17) and inequality (18) also apply as long as $y_k \leqslant d_k$. Therefore, $y_k^* \geqslant d_k$. (Note that $y_k^* > d_k$ if some capacity of product k is used for upgrading.) Therefore, in general, $y_k^* \geqslant (d_k \wedge x_k)$.

To show that Step 2 of the PRA is optimal after Step 1 has been completed, we first note that by Lemma 3, any upgrading can only occur when $d_k < y_k \leqslant x_k$ and $d_{k+1} > x_{k+1}$. Given these conditions,

$$\beta_k^+\widehat{\Theta}^t(y_k, \mathbf{Y}_{-k}, \mathbf{X} \mid \mathbf{D}) = \begin{cases} \alpha_{k+1,k} - \delta_k^+\Theta^{t+1}(\mathbf{X} - \mathbf{Y}) \\ \quad \text{for } d_k + d_{k+1} > y_k + x_{k+1}, \\ 0 \quad \text{otherwise;} \end{cases} \tag{19}$$

and

$$\beta_k^-\widehat{\Theta}^t(y_k, \mathbf{Y}_{-k}, \mathbf{X} \mid \mathbf{D}) = \begin{cases} \alpha_{k+1,k} - \delta_k^-\Theta^{t+1}(\mathbf{X} - \mathbf{Y}) \\ \quad \text{for } d_k + d_{k+1} \geqslant y_k + x_{k+1}, \\ 0 \quad \text{otherwise.} \end{cases} \tag{20}$$

Given that $d_{k+1} > x_{k+1}$, from Step 1 of the PRA we know that $x_{k+1} - y_{k+1}^* = 0$. Therefore, given that Step 1 has been completed, the derivatives of Θ^{t+1} in (19) and (20) are equal to the derivatives in the equivalent subproblems bounded by products k and l, where $l = 1$ or $x_{l-1} - y_{l-1}^* = 0$. That is, in (19), $\delta_k^+\Theta^{t+1}(\mathbf{X} - \mathbf{Y}) = \delta_k^+\Theta^{t+1}(n_l, \ldots, n_{k-1}, x_k - y_k)$ and in (20), $\delta_k^-\Theta^{t+1}(\mathbf{X} - \mathbf{Y}) = \delta_k^-\Theta^{t+1}(n_l, \ldots, n_{k-1}, x_k - y_k)$.

Let $p_k = x_k - y_k$. We will identify sufficient conditions on p_k to maximize $\widehat{\Theta}^t$. Given that $d_k < y_k \leqslant x_k$ and $d_{k+1} > x_{k+1}$, from Equation (15), H^t is concave in p_k. Because Θ^{t+1} is also concave, $\widehat{\Theta}^t$ is the sum of concave functions and is therefore also concave.

Now, recall the PRA's threshold condition (14), $\alpha_{k+1,k} \in \delta_k\Theta^{t+1}(n_l, \ldots, n_{k-1}, \tilde{p}_k)$ for some $\tilde{p}_k$. This condition, Equations (19) and (20), and the concavity of Θ^{t+1} imply that

$$0 \in \beta_k\widehat{\Theta}^t(\tilde{p}_k, \mathbf{Y}_{-k}, \mathbf{X} \mid \mathbf{D}). \tag{21}$$

Therefore, the threshold $\tilde{p}_k$ maximizes $\widehat{\Theta}^t$ in Step (ii) of Lemma 3, and it is optimal to sell capacity k to customers $k+1$ as long as more than $\tilde{p}_k$ units remain unsold. Finally, note that $\tilde{p}_k$ is a function of $(n_l, \ldots, n_{k-1})$ and is independent of the available capacity $n_k = x_k - d_k$ and the unmet demand $n_{k+1} = d_{k+1} - x_{k+1}$. Given $\tilde{p}_k$, the optimal amount of capacity k sold is $y_k^* = d_k + (n_k - \tilde{p}_k)^+ \wedge |n_{k+1}|$.

2. Given that we optimize using the PRA,

$$\begin{aligned} &\delta_k^-\Theta^t(\mathbf{X}) \\ &= \alpha_{kk}\Pr(\mathfrak{d}_k \geqslant x_k) + \alpha_{k,k+1} \\ &\quad \cdot \Pr(\mathfrak{d}_k < x_k, \mathfrak{d}_{k+1} > x_{k+1}, \mathfrak{d}_k + \mathfrak{d}_{k+1} \geqslant x_k + x_{k+1} - \tilde{p}_k) \\ &\quad + \delta_k^-\Theta^{t+1}(\mathbf{X})[1 - \Pr(\mathfrak{d}_k \geqslant x_k) - \Pr(\mathfrak{d}_k < x_k, \mathfrak{d}_{k+1} > x_{k+1}, \\ &\qquad \mathfrak{d}_k + \mathfrak{d}_{k+1} \geqslant x_k + x_{k+1} - \tilde{p}_k)]. \end{aligned} \tag{22}$$

By Assumption 2 of this lemma, and because Assumptions (A2) and (A3) of the model imply that $\alpha_{k,k+1} < \alpha_{kk}$, $\delta_k^-\Theta^{t+1}(\mathbf{X}) \leqslant \alpha_{kk}$.

3. The concavity of Θ^t follows from the optimality of the PRA and Lemma 2. □

In the proof, we have shown that the properties in Lemma 5 are preserved under backwards induction. Therefore, to show that the PRA is an optimal policy, we need only show that the three properties hold for period T. From Hoffman (1963), the PRA is optimal (with no rationing) in period T, α_{kk} remains an upper bound on the unit value of capacity, and Lemma 2 implies the concavity is preserved.

PROPOSITION 2. *The PRA is an optimal policy from among all admissible policies.*

PROOF. Consider the last-period problem, $\Theta^T(\mathbf{X})$. Given that $\Theta^{T+1} \equiv 0$, arguments identical to those in the proof of Lemma 5 show that $\delta_k^-\Theta^T(\mathbf{X}) \leqslant \alpha_{kk}$. From Lemma 2, $\Theta^T(\mathbf{X})$ is concave in $\mathbf{X}$. In addition, the greedy algorithm defined by Hoffman (1963) solves $\Theta^T(\mathbf{X})$, and is a special

case of the PRA with protection limits $\tilde{p}_k = 0$. Therefore, the argument of Lemma 5 iterates backwards through $T, T-1, \ldots, 1$. □

Next, we show that there exists an integer optimal rationing policy given that initial capacity is integer valued and that each period's demands are integer valued. Therefore, problems that are integer in demand and capacity have integer solutions. Let P^t be the set of protection levels for all upgrading problems in period t. Then, we have:

PROPOSITION 3. *If* $\mathbf{X}^1$ *and demand vectors* $\mathfrak{D}^1, \ldots, \mathfrak{D}^T$ *are integer valued, then there exists an optimal integer rationing policy* $(\tilde{P}^1, \ldots, \tilde{P}^T)$.

For any such integer-valued problems in period t, define $\Delta_k^t(\mathbf{X}^t) = \Theta^t(\mathbf{X}^t + e_k) - \Theta^t(\mathbf{X}^t)$, where $\mathbf{e}_k$ is the kth unit vector. The optimal protection limit, $\tilde{p}_k$, is the smallest value of p such that

$$\alpha_{k+1,k} \geqslant \Delta_k^{t+1}(n_1^t, \ldots, n_{k-1}^t, p,). \tag{23}$$

The marginal value Δ_k^{t+1} depends upon the time period, the current capacities of all products, and the distribution of future demand, and therefore can be difficult to calculate. In the next section, we consider methods for efficiently approximating $\tilde{p}_k$ for integer-valued problems.

5. A Model with Capacity Replenishment

Now assume that every T periods the firm can replenish capacity, and then the firm faces another capacity-rationing problem. To avoid confusion, we will continue to use the term *demand period* to describe each of the T relatively short time periods imbedded between each replenishment opportunity. We will use the term *replenishment interval* or just *interval* to describe each of the longer time periods between each replenishment. There is a finite number, R, of replenishment intervals. Assume that demands among the replenishment intervals are independent and identically distributed according to the random vectors $\mathfrak{D} = (\mathfrak{D}^1, \ldots, \mathfrak{D}^T)$, as defined in §3. Before the first interval, the firm invests in capacity $\mathbf{X}$ and is charged the effective unit cost $\mathbf{c} = (c_1, \ldots, c_N)$. Leftover capacity is held between intervals, and at the beginning of each interval capacity can be replenished at cost $\mathbf{c}$. Capacity left over at the end of each interval is assessed holding cost $\mathbf{h} = (h_1, \ldots, h_N)$. Capacity left over after the last interval R has value $\mathbf{c}$ (i.e., it can be sold for the initial effective cost). All other costs and revenues are as described in §3. Cash flows are discounted using discount factor $0 < \gamma \leqslant 1$ for each interval, and all costs and revenues are assumed to be expressed as beginning-of-period monetary units. Finally, note that the following analysis and results are similar to the analysis and results in Van Mieghem and Rudi (2002, §4).

Before analyzing the replenishment problem, we extend the notation for the single-interval DYN problem. Let $\Pi(\mathbf{X}; \mathbf{l})$ represent $\Pi^{\mathrm{DYN}}(\mathbf{X})$, given a vector of effective salvage values $\mathbf{l} = (l_1, \ldots, l_N)$. In particular, for our replenishment model, leftover capacities are not salvaged but do cost $\mathbf{h}$, so the relevant single-interval problem is $\Pi(\mathbf{X}; -\mathbf{h})$. In addition, an effective salvage value vector $\gamma\mathbf{c} - \mathbf{h}$ will be useful for the analysis of the multi-interval replenishment problem. Recall that problem $\Pi(\mathbf{X}; \mathbf{l})$ is concave in $\mathbf{X}$ (see the discussion after Lemma 2). Let $\mathbf{X}^*$ be an unconstrained maximizer of the single-interval rationing problem with salvage value $\mathbf{l} = \gamma\mathbf{c} - \mathbf{h}$:

$$\mathbf{X}^* \in \arg\max_{\mathbf{X} \in \mathbb{R}_N^+} \Pi(\mathbf{X}; \gamma\mathbf{c} - \mathbf{h}).$$

Given that the capacity at the beginning of any period is $\mathbf{Z} \leqslant \mathbf{X}^*$, we have the following stationary optimal policy.

PROPOSITION 4. *Given capacity* $\mathbf{Z} \leqslant \mathbf{X}^*$ *at the beginning of a replenishment interval, an optimal replenishment policy is to order up to* $\mathbf{X}^*$*, and the PRA is an optimal rationing policy within the interval.*

If $\mathbf{Z} > \mathbf{X}^*$, the PRA may not be an optimal rationing policy within the interval, and the replenishment policy may be much more complicated (see Song and Xue 2007 for an example of such complex replenishment policies in a multiproduct setting). If we relax the assumption that stock left over after period R can be sold for $\mathbf{c}$, the replenishment policy may again be much more complicated and the PRA may no longer be optimal within each interval. Likewise, if demand is not stationary, the PRA may not be optimal. Finding optimal policies for these more general multi-interval cases may be an interesting area for additional research.

6. Properties of the Protection Limits: Monotonicity and Bounds

For the remainder of this paper, we focus on the single-interval problem with one initial opportunity to purchase capacity, followed by T demand periods. The results of the previous section imply that all of the following results also apply to the problem with multiple replenishments, where the initial capacity $\mathbf{X}^1$ can be interpreted as the capacity after replenishment.

In this section, we show that the protection limits are monotonically decreasing in both the amount of capacity and time, and we use these properties to derive a series of bounds on the protection limits. We then describe numerical experiments that demonstrate the tightness of the bounds. We end with a discussion of how the bounds can be used to accurately approximate the protection limits for large problems. Throughout this section, we assume that $\mathbf{X}^t \in \mathbb{Z}_N^+$ and $\mathfrak{D}^t \in \mathbb{Z}_N^+$.

6.1. Monotonicity and Bounds on the Protection Limits

Let $\tilde{p}^t$ be the optimal protection limit for a subproblem at time t. We show that $\tilde{p}^t$ is monotonically decreasing in the capacity state and over time.

PROPOSITION 5. *The optimal protection limit $\tilde{p}^t$ is decreasing in the state vector $\mathbf{X}^t$.*

PROPOSITION 6. *The optimal protection limit $\tilde{p}^t$ is decreasing in t.*

These propositions lead directly to sets of upper and lower bounds on the protection limits. We derive the bounds by restricting our attention to a limited number, i, of products above the one that might be rationed. To find the upper (lower) bounds on the protection limit, we set the capacity of the product immediately above those i products to 0 (∞).

Specifically, suppose that we have a subproblem involving products $1, \ldots, k+1$. Let $\tilde{p}(\mathbf{X})$ be the optimal protection limit of product k, given initial capacity vector $\mathbf{X} = (x_1, \ldots, x_k)$ (for clarity, we suppress the superscript t). Define a new, truncated capacity vector $\mathbf{X}(i, C) = (C, x_{k-i}, \ldots, x_k)$, $i = 0, \ldots, k-1$ (if $i = 0$, then the capacity vector is just (C, x_k)). Setting $C = 0$ indicates that there is no capacity of product $k-i-1$, and we use the notation $C = \infty$ to indicate that there is no capacity constraint for product $k-i-1$. That is, with $\mathbf{X}(i, \infty)$, any quantity of demand available to be upgraded from product $k-i$ to product $k-i-1$ provides revenue of $\alpha_{k-i,\,k-i-1}$ per unit (here we assume that demand is finite, so that the objective function is still bounded). Therefore, $\mathbf{X}(i, 0)$ and $\mathbf{X}(i, \infty)$ define two smaller subproblems that involve $i+2$ products. In each of these subproblems, product $k+1$ may be completely depleted, product k may be rationed, and there are i products with nonzero capacities, products $k-i, \ldots, k-1$, that may affect the optimal protection level of product k. The capacities $(x_1, \ldots, x_{k-i-2})$ have no impact on the rationing problem because products $k-i, \ldots, k$ are "cut off" by the zero or infinite capacity of product $k-i-1$.

Now suppose that product $k-i-1$ has zero capacity. Proposition 5 implies that the protection limit of product k remains the same or declines as the capacity of $k-i-1$ increases from zero. Therefore, the protection level for the subproblem with capacity vector $\mathbf{X}(i, 0)$, $\tilde{p}(\mathbf{X}(i, 0))$, is an upper bound on $\tilde{p}(\mathbf{X})$. This upper bound becomes tighter as i increases and more levels of capacity are added above k. Likewise, if product $k-i-1$ has very large capacity, the protection limit of product k remains the same or increases as the capacity of $k-i-1$ decreases. This implies that $\tilde{p}(\mathbf{X}(i, \infty))$ provides a series of lower bounds, and these lower bounds are also progressively tighter as i increases. We make these statements more precise in the following proposition.

PROPOSITION 7. *For a subproblem with k products,*

$$\begin{aligned}&\tilde{p}(\mathbf{X}(0, \infty))\\ &\quad\leqslant \tilde{p}(\mathbf{X}(1, \infty)) \leqslant \cdots \leqslant \tilde{p}(\mathbf{X}(k-1, \infty))\\ &\quad\leqslant \tilde{p}(\mathbf{X})\\ &\quad\leqslant \tilde{p}(\mathbf{X}(k-1, 0)) \leqslant \tilde{p}(\mathbf{X}(k-2, 0)) \leqslant \cdots \leqslant \tilde{p}(\mathbf{X}(0, 0)).\end{aligned}$$

PROOF. The tightest bounds, $\tilde{p}(\mathbf{X}(k-1, \infty)) \leqslant \tilde{p}(\mathbf{X}) \leqslant \tilde{p}(\mathbf{X}(k-1, 0))$, follow from Proposition 5. Now consider $\tilde{p}(\mathbf{X}(i, 0))$ for $0 < i \leqslant k-1$. From Proposition 5 and Lemma 4, $\tilde{p}(\mathbf{X}(i, 0)) = \tilde{p}(0, x_{k-i}, \ldots, x_k) \leqslant \tilde{p}(0, 0, x_{k-i+1}, \ldots, x_k) = \tilde{p}(0, x_{k-i+1}, \ldots, x_k) = \tilde{p}(\mathbf{X}(i-1, 0))$.

For the lower bounds, note that setting $C = \infty$ has a similar impact on the size of the subproblem as setting $C = 0$. As in Lemma 4, an inexhaustible supply of capacity splits the subproblem into smaller pieces: If product $k-i-1$ can satisfy any quantity of demand, then the protection limit of product $k > k-i-1$ does not depend upon the capacity levels of products $1, \ldots, k-i-2$. This fact and Proposition 5 imply that for $0 < i \leqslant k-1$, $\tilde{p}(\mathbf{X}(i, \infty)) = \tilde{p}(\infty, x_{k-i}, \ldots, x_k) \geqslant \tilde{p}(\infty, \infty, x_{k-i+1}, \ldots, x_k) = \tilde{p}(\infty, x_{k-i+1}, \ldots, x_k) = \tilde{p}(\mathbf{X}(i-1, \infty))$. □

These bounds are useful because the dimensionality of the dynamic program rises with the number of products in the subproblem. Specifically, for many problems of reasonable size, calculation of the optimal protection limits using backwards induction is impossible. For a subproblem with T time periods, k products, and a maximum of $\hat{x}$ for the capacity of each product, there are $O(T\hat{x}^{k-2})$ distinct protection limits to calculate (with $T = 10$, $\hat{x} = 100$, and $k = 5$, there are over 10 million protection limits). However, Proposition 7 provides us with a series of bounds that allow for a trade-off between accuracy and computational burden. In the next section, we will focus on $[\tilde{p}(\mathbf{X}(1, 0)), \tilde{p}(\mathbf{X}(1, \infty))]$, *one-product bounds*, determined by the capacity of a single adjacent product. There are $O(Tk\hat{x})$ protection limits associated with these bounds. We will also examine the accuracy of $[\tilde{p}(\mathbf{X}(2, 0)), \tilde{p}(\mathbf{X}(2, \infty))]$, *two-product bounds*, and there are $O(Tk\hat{x}^2)$ of these. If either of these bounds are sufficiently tight, then protection limits chosen between these bounds will be both nearly optimal and easy to calculate. In the next section, we will find that, indeed, there is rarely any gap between either the one- or two-product bounds, so that they provide us with methods for solving problems with large numbers of products.

6.2. Protection Limit Bounds: Numerical Experiments

We now summarize numerical experiments that test the quality of the one-product and two-product bounds described above. Full details of the parameters are available in §2 of the online appendix. In all of the experiments, we have five products ($k = 5$) and 10 time periods ($T = 10$). Each product has a maximum initial capacity of 30 ($\hat{x} \leqslant 30$) and a maximum total mean demand of 50 across all time periods ($\sum_{t=1}^{10} E[\mathfrak{D}^t] \leqslant 50$). In one subset of experiments, we assume that demands arrive according to Poisson distributions that are independent between demand periods and between products. In another subset, we assume that demands arrive according to multivariate normal distributions, truncated at zero and rounded to the nearest integer.

Table 1. Size of gaps for one-product and two-product bounds.

Gap type	No. of gaps calculated	Percentage of gap > 0 (no. of instances)	Maximum gap	Mean revenue loss	Maximum revenue loss
$\nabla_1(\mathbf{X})$	123,012	0.3%(395)	1	0.00001%	0.0031%
$\nabla_2(\mathbf{X})$	4,442,196	0.0001%(6)	1	≈0	≈0

For this latter subset, we vary the within-period coefficient of correlation among all demands from −0.25 to 0.9.

Given these demand distributions, we run 408 experiments using a wide variety of parameter values. We vary the ratio of demand to capacity for each product, the distribution of demand over time, the pattern of mean demand between products and across time periods, and the marginal contribution of parallel sales and upgrades. We consider both realistic and extreme cases, e.g., for one extreme case we set the initial capacity to be $\mathbf{X}^1 = [1, 45, 1, 45, 1]$. Note that in these experiments, we assume an arbitrary initial capacity, whereas in the numerical experiments of §7 below, we will always use the optimal initial capacity.

For each scenario, we calculate the gaps $\nabla_1(\mathbf{X}) \equiv \tilde{p}(\mathbf{X}(1,0)) - \tilde{p}(\mathbf{X}(1,\infty))$ and $\nabla_2(\mathbf{X}) \equiv \tilde{p}(\mathbf{X}(2,0)) - \tilde{p}(\mathbf{X}(2,\infty))$ for product 4 (note that $\nabla_2(\mathbf{X}) = \mathbf{0}$ for products 1, 2, and 3 because the protection limits of these products depend upon the capacity of at most two products). The 408 experiments yield 123,012 one-product bounds and 4,442,196 two-product bounds for product 4. Table 1 summarizes the results of the experiments. For both the one-product and two-product bounds, the maximum gap is just one unit, and most of the bounds have no gap at all. In fact, for the two-product bounds, just six out of the 4.4 million gaps are one.

Therefore, for these experiments, either of the two-product bounds is equivalent to the optimal solution, and the one-product bounds are quite close. Using the one-product upper bound on the protection limit rather than either two-product bound produces a small loss in expected revenue, just 0.00001% on average and a maximum revenue loss of 0.0031%. Additional experiments described in §7 with another set of five-product problems produce similar results: Out of 27,000 protection levels calculated, over 99% of the gaps $\nabla_1(\mathbf{X})$ are zero, and the maximum gap is again one.

The accuracy of the heuristic protection limits based on these bounds, and the relative ease with which one- and two-product bounds can be calculated, provide us with an opportunity to compare the static and dynamic formulations in a realistic context, with large numbers of products and time periods.

7. The Value of Optimal Capacity and Allocation: Numerical Experiments

This section describes the results from analytical and numerical studies designed to understand how the parameters of the model affect two quantities: (i) the value of optimal upgrading, and (ii) the value of using the capacity that is strictly optimal, given that optimal upgrading will be applied, rather than using capacity that is optimal for the simpler static model. Here we calculate the value of optimal upgrading as the difference between the profit generated from the DYN model and the profit generated from two simpler heuristics, the NV model and a *greedy heuristic* in which $y^t_{k+1,k} = [(d^t_{k+1} - x^t_{k+1})^+ \wedge (x^t_k - d^t_k)^+]$ for $k = 1, \ldots, N-1$, i.e., all possible upgrading is performed in each period. We calculate the value of strictly optimal capacity as the difference between the profits generated by DYN and a *hybrid heuristic* in which the initial capacity is optimal for the STC problem and then optimal upgrading is used, once customers begin arriving.

We assess the impact of model parameters on the quantities (i) and (ii) described above. In particular, we examine the impact of three attributes of the model:

- *Availability of advance demand information*. In the one-period model (STC), all demand information is available when all allocation occurs, so that capacity may be assigned to customers without any possibility of cannibalization. In practice, demand information may become available in small increments over time, and we examine the impact of the incremental release of demand information by varying the number of periods in the DYN model.
- *Economic parameters*, the contribution margins α_{ij} and the initial capacity costs c_j.
- *Demand parameters*, the variance and within-period correlations of the demand.

Our experiments include almost 5,000 parameter scenarios with a two-product model and 20 scenarios with a five-product model. In §4 of the online appendix, we include an expanded version of this section with full details on the parameter values used in the scenarios, the algorithms used to find the optimal capacities and protection levels, and the results of the experiments. Here we will provide an overview of the models, summarize the results, and present illustrative examples.

First, we describe parameters that are common to all of the models. For all two-product scenarios, the total mean demand for each product over all periods is 60 units, whereas for the five-product scenarios the total mean demand for each product over all periods is 20 units. For every scenario, demand for high-value products rises over the time periods, whereas demand for low-value products declines. This is consistent with environments that are amenable to yield management techniques, in which high-value customers tend to arrive after low-value customers. In all experiments, we chose economic parameter ranges that were bounded either by (A1)–(A3) or by limits imposed by real-world

applications, e.g., $c_1 > c_2$, the unit cost of product 1 should be greater than the unit cost of product 2.

In general, our numerical experiments lead to the following observations,

1. Profits from DYN and the hybrid heuristic are nearly identical, so that using capacity that is optimal for STC when paired with optimal upgrading, produces results that are close to results when using the optimal capacity for DYN.

2. Profits from DYN (or the hybrid heuristic) are consistently within 1% of STC, so that perfect demand information has relatively little value as long as optimal upgrading is used.

3. As T grows, profits from DYN and the hybrid heuristic decline relative to STC, but not by much. For reasonably large T (say, $T \geqslant 5$), profits from NV dominate profits from the greedy heuristic. This appears to be due to the fact that when T is large, advance demand information is not available for many of the allocation decisions. Therefore, in this environment the risk of cannibalization is higher under the greedy heuristic, and it is safer to keep products separate, as in the NV model.

In addition, we learned that the optimal upgrading policy generated by DYN is most valuable, as compared to the simple heuristics, when

4. $\alpha_{i+1,i}/\alpha_{ii}$ is close to one for all $i = 1, \ldots, N-1$. If upgrades have a relatively high value, then using optimal upgrading provides significant profit above the newsvendor solution with no upgrading.

5. $c_1 - c_N$ is close to zero and $c_N \leqslant c_i \leqslant c_1$, $i = 2, \ldots, N-1$. If c_1 is close to c_N, product-1 capacity is relatively inexpensive, whereas product-N capacity is relatively expensive. Therefore, it is optimal to invest in large amounts of the high-value product and relatively little lower-value product. This increases the likelihood of low-value shortage and high-value surplus, providing more opportunities to upgrade.

6. α_{11}/α_{NN} is close to one. When α_{11}/α_{NN} is large ($\gg 1$), parallel revenues from high-value products represent a high proportion of the revenue, so that upgrading provides relatively less value. On the other hand, if parallel contribution margins of products 1 and N are similar, then type-1 parallel revenues do not dominate and optimal upgrading is valuable.

7. Demand variance is high. With uncertain demand, mismatches between demand and capacity are more likely to occur, so that optimal upgrading becomes useful.

8. Demand correlation between products is low. Under low correlation, it is more likely that a stockout for a low-type product is paired with a surplus of a higher-type product, thus increasing the value of optimal upgrading.

In the following sections, we present examples to illustrate many of these observations. Again, the online appendix contains more details on all the experiments and the complete set of evidence for points 1–8.

7.1. The Value of Using Optimal Capacity in the Dynamic Model

Using both analysis and numerical experiments, we have found that the optimal initial capacities under STC and DYN may differ substantially in certain stylized environments. Analysis of a model with two products, two time periods, and continuous demand and capacity shows that the marginal value of an additional unit of type-2 capacity is more valuable under DYN than under STC. This is because extra type-2 capacity can be useful as a buffer to protect against supply cannibalization, upgrades of type-2 customers in the first period that lead to a shortage of type-1 capacity for type-1 customers in the second period. Although this result is not sufficient to show that the optimal quantity of type-2 capacity under DYN is always greater than the optimal type-2 capacity under STC ($x_2^{\mathrm{DYN}} \geqslant x_2^{\mathrm{STC}}$), we have conducted thousands of numerical experiments using a wide variety of parameters and two types of distribution functions (truncated normal and uniform), and in every case, $x_2^{\mathrm{DYN}} \geqslant x_2^{\mathrm{STC}}$. A subset of these experiments are described in §3 of the online appendix. There is no analogue of these results for type-1 capacity, and the online appendix contains examples in which $x_1^{\mathrm{DYN}} \leqslant x_1^{\mathrm{STC}}$ and $x_1^{\mathrm{DYN}} > x_1^{\mathrm{STC}}$. The online appendix also describes results that characterize how the protection limit in the two-product, two-period model changes with the product contribution margins, the distribution of product demand, and the correlation between demand distributions.

Again, we do find that large differences between the optimal initial capacities for STC and DYN can occur in extreme cases, e.g., when there are just two periods, all of type-2 demand in period 1, and all of type-1 demand in period 2. In more realistic cases, however, optimal capacities for STC and DYN are often nearly identical, and when they are not, there is a negligible difference in profits due to using STC-optimal capacity for the dynamic case (our hybrid heuristic) rather than using DYN-optimal capacity. This observation is useful because finding the optimal capacity for DYN is significantly more difficult than finding the optimal capacity for STC because the capacity optimization in DYN must take the future dynamic rationing policy into account, and therefore must evaluate the full dynamic program, given any initial capacity level. This can be cumbersome, even when taking advantage of the bounds described in Proposition 6. The value of STC given any capacity level, however, requires few relatively simple calculations (see Netessine et al. 2002).

To examine the impact of using optimal capacity, for each scenario we found the percentage increase in the expected profit due to using the optimal capacity for DYN rather than using the hybrid heuristic with the STC-optimal capacity. That is,

$$\text{value of using optimal capacity} \equiv \frac{\Pi^{\mathrm{DYN}}(\mathbf{X}^{\mathrm{DYN}}) - \Pi^{\mathrm{DYN}}(\mathbf{X}^{\mathrm{STC}})}{\Pi^{\mathrm{DYN}}(\mathbf{X}^{\mathrm{DYN}})}, \tag{24}$$

where $\Pi^X(\mathbf{X}^Y)$ is the profit from model X when starting with capacity that is optimal for model Y.

Of the nearly 5,000 scenarios in the two-product model, for 48% of the scenarios the DYN and STC capacities were identical. Overall, the average value of using optimal capacity—the average value of definition (24)—was 0.0008 (0.08% of the DYN profit). The 90th percentile of (24) among scenarios was 0.25%, and the maximum value was 2%. Results for the five-product experiments were similar: an average value of 0.01% and a maximum value of 0.05%. Therefore, ignoring the dynamic rationing policy when finding the initial capacity by using STC for capacity optimization, and then using optimal rationing, almost always performs as well as the much more complex capacity optimization in DYN.

7.2. The Value of Optimal Upgrading

In this section, we compare the profit from DYN with the profit from the NV model (no upgrading) and the greedy heuristic (myopic upgrading). First, we examine how DYN and the heuristics perform as the number of periods changes. These experiments quantify the value of advance demand information. Then, we examine the effects of changes in the financial and demand parameters.

7.2.1. Advance Demand Information. The model in DYN is equivalent to STC if the firm has a perfect demand forecast: If the firm knows exactly who is coming and when, then it can optimally allocate capacity among customers as if all customers had arrived in the same time period. Here we construct a series of DYN models, each of which has the same total demand over all periods. The models have an increasing number of periods, T, and we release less demand information within each period as T increases. The impact of this change can be seen in Figure 1, which displays profits from the five-product DYN model, hybrid heuristic, NV model, and greedy heuristic, all as a fraction of the upper-bound STC profit (e.g., the top line in the figure shows $\Pi^{\mathrm{DYN}}(\mathbf{X}^{\mathrm{DYN}})/\Pi^{\mathrm{STC}}(\mathbf{X}^{\mathrm{STC}})$ as T varies). In this example, as the number of periods in DYN increases, information availability decreases (more allocations are being made with less demand information), and the allocation is less effective than the allocation in STC. The figure also demonstrates other elements of points 1–3 above: Relative profits from DYN and the hybrid heuristic are virtually identical at the top of the figure (the lines essentially overlap) and both remain within 1% of STC. We also see that the results from NV are superior to the greedy heuristic for $T \geqslant 5$.

Over all two-product experiments, when $T = 20$ the median difference between the DYN formulation and the STC upper bound is also less than 1%. This leads to perhaps the most important point, observation 2, that perfect demand information has relatively little value as long we use the optimal initial capacity and optimal upgrading.

7.2.2. Impact of the Economic and Demand Parameters. In Figure 1, for large T, the difference between DYN and NV is approximately 3% of the STC profit, and the difference between DYN and greedy can be much larger. Now we ask, in general, when is it worth implementing optimal upgrading, the DYN policy, rather than one of the simpler heuristics? When assessing the value of optimal upgrading, we will compare DYN with the NV heuristic because the NV heuristic dominates the greedy heuristic for most problems with large T.

For example, Figure 2 illustrates observation 4, above—that the value of optimal upgrading rises with the relative upgrade value, $\alpha_{i+1,i}/\alpha_{i,i}$. This figure was generated using the five-product model with $T = 10$. As in Figure 1, the

Figure 1. Profits as a fraction of STC profits in the five-product model as the number of periods, T, varies.

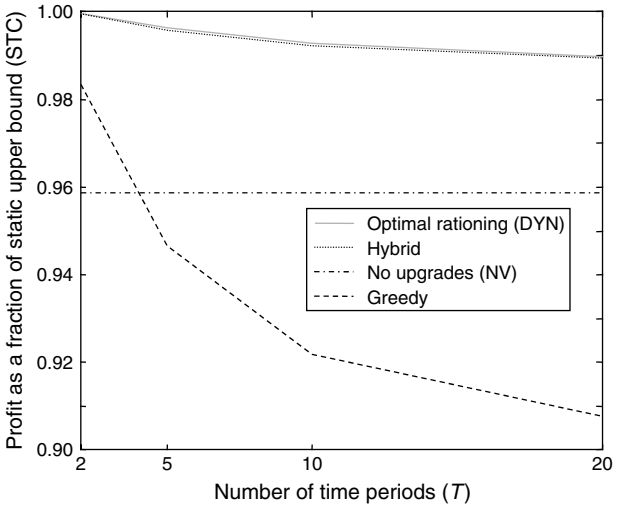

Figure 2. Profits as a fraction of STC profits in the five-product model as the contribution margin of upgrading varies.

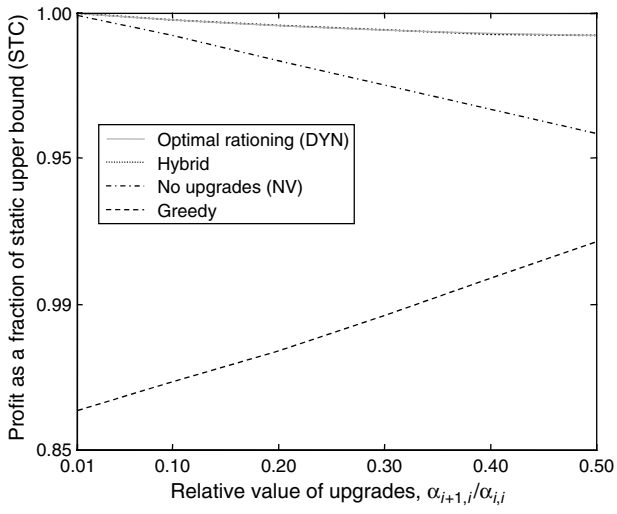

figure displays profits as a fraction of STC, and again the DYN and hybrid results are nearly identical. We also see that as the value of upgrades rises, the relative performance of the NV solution declines so that the value of optimal upgrading increases. Similar plots for observations 5–8, above, are available in §4 of the online appendix.

8. Conclusions and Future Research

In this paper, we formulate a flexible capacity investment and allocation problem in which demand arrives over a sequence of discrete time periods. Because total demand from the most lucrative customers is uncertain when capacity allocation decisions must be made, the firm may hold back, or ration, some products before the last time period. We show that the optimal assignment policy involves two steps: greedy allocation, followed by upgrading that is limited by a protection limit. We extend these results to a model in which the firm may replenish its capacity every T time periods and derive heuristics for generating nearly optimal protection limits for large problems.

We then explore the impact on total expected profit of using dynamic-optimal capacity and optimal upgrading. We also examine the consequences of using suboptimal policies, such as a greedy policy ("upgrade whenever possible") and a no-upgrade policy that separates the problem into simple newsvendor problems. We find that using optimal capacity for the static problem, when paired with optimal upgrading, produces profits that are close to profits when using the optimal capacity for the dynamic problem. We also find that under optimal upgrading, profits are consistently close to the upper-bound profits of the static problem, so that perfect demand information has relatively little value as long as optimal upgrading is used. Finally, we find that using optimal upgrading rather than no upgrading or greedy upgrading (no rationing) is particularly important when the value of upgrading is high, products are close together in terms of cost or marginal revenue, demand variance is high, and demand correlation is low.

There are many possible extensions to the model, such as the inclusion of backlogging and incorporating interperiod demand dependence that would allow the firm to update protection levels as demand arrives. It would also be interesting to relax the single-step upgrading assumption. For this latter extension, determining the actual values of optimal booking limits can be difficult, particularly in problems with large numbers of flexible products and time periods, so that recursive and/or heuristic methods for finding booking limits would be useful.

Most of the literature on upgrading and substitution assumes that customers do not react strategically to the firm's actions, and this assumption also applies to our model. However, customers may intentionally demand a lower-quality product in the hope of getting upgraded to a higher-quality product. This may not be an issue if the higher-quality product can be degraded (for example, a large hard disk drive can be formatted to be a smaller one at very little cost). However, degrading the product quality is not practical in most service industries. The impact of strategic customer behavior on firms' optimal capacity investment and upgrading decisions is an interesting direction for future research.

Finally, in many real-world environments customer arrivals cannot be divided into time periods, and an extension of the analysis would be to compare our dynamic model with a model that features continuous arrivals (e.g., customers arrive according to a Poisson or diffusion process). As Topkis (1968) points out, however, the assumption that demand arrives in discrete periods "might be expected to be a good approximation to reality if the intervals are made 'small enough' " (p. 161). In addition, a model with a small number of discrete demand periods may be a reasonable approximation when different customer classes tend to arrive in different periods, as is often the case in yield management applications.

9. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://or.journal.informs.org/.

Endnote

1. Throughout this paper, we use *decreasing* for *nonincreasing* and *increasing* for *nondecreasing*.

Acknowledgments

The authors are grateful to William Cooper, Marshall Freimer, Serguei Netessine, Jeannette Song, Jan Van Mieghem, Dan Zhang, two anonymous referees, and an associate editor for their helpful suggestions. They also thank participants at research seminars at the Operations Research Center of MIT, the University of Minnesota, the University of Toronto, the University of California, Irvine, and Washington University in St. Louis for their comments.

References

Alstrup, J., S. Boas, O. B. G. Madsen, R. V. V. Vidal. 1986. Booking policy for flights with two types of passengers. *Eur. J. Oper. Res.* **27** 274–288.

Archibald, T. W., S. A. E. Sassen, L. C. Thomas. 1997. An optimal policy for a two depot inventory problem with stock transfer. *Management Sci.* **43**(2) 173–183.

Axsäter, S. 2003. A new decision rule for lateral transshipments in inventory systems. *Management Sci.* **49**(9) 1168–1179.

Bassok, Y., R. Anupindi, R. Akella. 1999. Single-period multiproduct inventory models with substitution. *Oper. Res.* **47**(4) 632–642.

Bitran, G. R., S. M. Gilbert. 1996. Managing hotel reservations with uncertain arrivals. *Oper. Res.* **44**(1) 35–49.

Brumelle, S. L., J. I. McGill. 1993. Airline seat allocation with multiple nested fare classes. *Oper. Res.* **41**(1) 127–137.

Curry, R. E. 1990. Optimal airline seat allocation with fare classes nested by origins and destinations. *Transportation Res.* **24**(3) 193–204.

de Véricourt, F., F. Karaesmen, Y. Dallery. 2001. Assessing the benefits of different stock-allocation policies for a make-to-stock production system. *Manufacturing Service Oper. Management* **3**(2) 105–121.

de Véricourt, F., F. Karaesmen, Y. Dallery. 2002. Optimal stock allocation for a capacitated supply system. *Management Sci.* **48**(11) 1486–1501.

Deshpande, G. V., M. A. Cohen, K. Donohue. 2003. A threshold inventory rationing policy for service-differentiated demand classes. *Management Sci.* **49**(6) 683–703.

Ding, Q., P. Kouvelis, J. Milner. 2006. Dynamic pricing through discounts for optimizing multiple class demand fulfillment. *Oper. Res.* **54**(1) 169–183.

Eynan, A. 1999. The multi-location inventory centralization problem with first-come, first-served allocation. *Eur. J. Oper. Res.* **11** 438–449.

Fine, C. H., R. M. Freund. 1990. Optimal investment in product-flexible manufacturing capacity. *Management Sci.* **36**(4) 449–466.

Frank, K. C., R. Q. Zhang, I. Duenyas. 2003. Optimal policies for inventory systems with priority demand classes. *Oper. Res.* **51**(6) 993–1002.

Ha, A. Y. 1997a. Inventory rationing in a make-to-stock production system with several demand classes and lost sales. *Management Sci.* **43**(8) 1093–1103.

Ha, A. Y. 1997b. Stock rationing policy for a make-to-stock production system with two priority classes and backordering. *Naval Res. Logist.* **44** 457–472.

Ha, A. Y. 2000. Stock rationing in an $M/E_k/1$ make-to-stock queue. *Management Sci.* **46**(1) 77–87.

Hoffman, A. J. 1963. On simple linear programming problems. V. Klee, ed. *Convexity*: *Proceedings of Symposia in Pure Mathematics*, Vol. 7. American Mathematical Society, Providence, RI.

Jordan, W. C., S. C. Graves. 1995. Principles on the benefit of manufacturing process flexibility. *Management Sci.* **41**(4) 577–594.

Kapuscinski, R., S. Tayur. 2000. Dynamic capacity reservation in a make-to-stock environment. Working paper, University of Michigan, Ann Arbor, MI.

Karaesmen, I., G. van Ryzin. 2004. Overbooking with substitutable inventory classes. *Oper. Res.* **52**(1) 83–104.

Karmarkar, U. S. 1981. The multiperiod multilocation inventory problem. *Oper. Res.* **29**(2) 215–228.

McGill, J., G. J. van Ryzin. 1999. Revenue management: Research overview and prospects. *Transportation Sci.* **33**(2) 233–256.

Netessine, S., G. Dobson, R. A. Shumsky. 2002. Flexible service capacity: Optimal investment and the impact of demand correlation. *Oper. Res.* **50**(2) 375–388.

Pasternack, B. A., Z. Drezner. 1991. Optimal inventory policies for substitutable commodities with stochastic demand. *Naval Res. Logist.* **38** 221–240.

Porteus, E. L. 1975. On the optimality of structured policies in countable stage decision processes. *Management Sci.* **22**(2) 148–157.

Robinson, L. W. 1990. Optimal and approximate policies in multiperiod, multiproduct inventory models with transshipment. *Oper. Res.* **38**(2) 278–295.

Savin, S. V., M. A. Cohen, N. Gans, Z. Katalan. 2005. Capacity management in rental businesses with heterogeneous customer bases. *Oper. Res.* **53**(2) 617–631.

Song, J.-S., Z. Xue. 2007. Demand management and inventory control for substitutable products. Working paper, Duke University, Durham, NC

Subramanian, J., S. Stidham, C. J. Lautenbacher. 1999. Airline yield management with overbooking, cancellations, and no-shows. *Transportation Sci.* **33**(2) 147–167.

Talluri, K., G. van Ryzin. 2004. *The Theory and Practice of Revenue Management*. Kluwer Academic Publishers, Boston.

Tomlin, B., Y. Wang. 2008. Pricing and operational recourse in coproduction systems. *Management Sci.* **54**(3) 522–537.

Topkis, D. M. 1968. Optimal ordering and rationing policies in a nonstationary dynamic inventory model with n demand classes. *Management Sci.* **15**(3) 160–176.

Van Mieghem, J. A. 1998. Investment strategies for flexible resources. *Management Sci.* **44**(8) 1071–1078.

Van Mieghem, J. A. 2003. Capacity management, investment, and hedging: Review and recent developments. *Manufacturing Service Oper. Management* **5**(4) 269–302.

Van Mieghem, J. A., N. Rudi. 2002. Newsvendor networks: Inventory management and capacity investment with discretionary activities. *Manufacturing Service Oper. Management* **4**(4) 313–335.

Wollmer, R. D. 1992. An airline seat management model for a single leg route when lower fare classes book first. *Oper. Res.* **40**(1) 26–37.

Zhao, W., Y. S. Zheng. 2001. A dynamic model for airline seat allocation with passenger diversion and no-shows. *Transportation Sci.* **35**(1) 80–98.