Dynamic Capacity Management with Substitution

Robert A. Shumsky
Simon School of Business
University of Rochester
Rochester, NY 14627
shumsky@simon.rochester.edu

Fuqiang Zhang
UC Irvine Graduate School of Management
University of California
Irvine, CA 92697-3125
fzhang@uci.edu

April, 2003. Lastest revision: October, 2004

Abstract

We examine a multiperiod capacity allocation model with upgrading. There are multiple product types, corresponding to multiple classes of demand, and the firm purchases inventory of each product before the first period and thereafter replenishment is not possible. 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 solution is a simple two-step algorithm: first use any available inventory to satisfy same-class demand and then upgrade customers until inventory reaches a protection limit. We describe how to find optimal protection limits by backward induction and show that the protection limits are monotonic in current inventory and in time. The monotonicity results lead to simple bounds for the optimal protection limits, and we demonstrate that heuristic protection limits based on the bounds are effective in solving large problems.

For a model with two time-periods and two products (dedicated and flexible), we compare the optimal initial capacity under our dynamic model with the optimal capacity under a single-period 'static' model. For a given capacity level, the marginal value of the dedicated capacity is always greater under the dynamic model than under the static model. Numerical experiments show that the optimal level of dedicated capacity is greater under the dynamic model than under that static model, although the optimal flexible capacity can be smaller, or greater, under the dynamic model.

Keywords: inventory management, dynamic programming, rationing policy, revenue management.

1 Introduction

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

rental car agency. If the customer's requested car is unavailable, the agency may choose to upgrade the customer to a more expensive car. Customers arrive throughout the day, and this allocation decision must be made when the day's total demand for the higher-value car is still uncertain. Similar decisions appear in other services (e.g., hotels allocating rooms, airlines allocating economy and business-class seats) as well as in manufacturing (e.g., a firm choosing to use a high-value part when a lower-valued part is sufficient but unavailable).

Here we analyze a dynamic multi-product inventory model in which demand arrives in discrete intervals. Throughout this paper we use the terms 'inventory' and 'capacity' interchangeably, for the products may be interpreted as either service capacities or perishable inventories. We describe a stylized model of the problem faced by the rental car agency each day: how much inventory should be acquired before the day begins, and how should that inventory be distributed among customers as the day evolves? The model has the following attributes:

1. There is a single opportunity to invest in inventory before any demand is realized.

2. The period after the initial investment is broken into a finite number of intervals, and the decision-maker allocates inventory to customers after observing all demand within each interval. In practice, demand may arrive continuously, but as Topkis (1968) points out, the assumption that demand arrives in discrete intervals "might be expected to be a good approximation to reality if the intervals are made 'small enough.' "

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 rental agency's demand for a compact car may be met with a full-size car).

5. Inventory may be rationed, so that the firm may choose not to allocate high-class inventory to a lower-class customer.

We are most interested in the impact of characteristic (2), the assumption that customers arrive over a sequence of time intervals. As we will see in the following literature review, there exist numerous articles that examine the benefits of inventory flexibility. However, most authors assume

that all demand appears simultaneously, or that if demand appears over time then inventory is allocated only after the last customer arrives. Given this perfect information about demand, inventory can be allocated optimally. However, for many applications this 'single-period' assumption overestimates profits and, as we shall see, may also overestimate the value of inventory flexibility.

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

After reviewing the literature, we describe the model in Section 3 and show that the single-period formulation provides an upper bound on the expected profit of our dynamic model. In Section 4 we prove that a 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 Section 5 we show that the protection limit of each inventory class is decreasing as time increases and is decreasing in the inventory level of any of the available products.[1] We also derive bounds on both the size of the problem and the optimal protection limits. The bounds on the size of the problem are expressed in terms of the number of products and number of remaining time periods. The bounds on the optimal protection limits are complementary pairs of lower and upper bounds that grow progressively tighter as the computational effort to calculate the bounds increases. Section 5 ends with numerical experiments demonstrating that over a wide range of parameters, the bounds are extremely tight. In fact, bounds based only upon the inventory level of one adjacent product allow us to estimate protection levels that are extremely close to optimal, and these bounds can be calculated quite quickly.

In Section 6 we focus on the optimal capacity decision. In particular, we examine the problem with two products, dedicated inventory and flexible inventory, and two time-periods (we will call

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

this the '2x2 model'). We compare the 2x2 model with a single-period model; this comparison is important, for single-period, static models have often been used to evaluate the benefits of flexible inventory. We prove that, for any given capacity, the marginal value of the dedicated inventory is larger in the dynamic model than in the static model. In numerical experiments we find that the optimal initial level of dedicated inventory is always higher in the dynamic model than in the single-period model. We also find that the optimal initial level of flexible inventory may be lower, or higher, in the dynamic model than in the static model. This finding contradicts the intuition that because flexibility may be used to greater advantage in the static setting than in the dynamic setting, flexible inventory should always be more valuable - and therefore should be more plentiful - in the static setting.

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, address all five. See Van Mieghem, 2003, for a survey of models to study capacity investment and management. Some researchers have focused on single-period multidimensional newsvendor models. 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.

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 period. For the service 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 before the first period. The firm's inability to replenish inventory between periods also distinguishes our work

from the literature on multiperiod inventory models with transshipment, such as Karmarkar (1981), Robinson (1990), and Archibald et al. (1997).

The literature on yield management does focus on environments in which inventory-sizing (or capacity) decisions are made and then inventory must be allocated as demand arrives over time. See McGill and van Ryzin, 1999, for a survey of this literature. The analysis by Brumelle and McGill (1993) characterizes 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). In all of these papers there is a single type of resource, a coach seat on a single-leg flight, 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 (2000) 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. While substitution is allowed during the second, allocation stage, there is no substitution as demand arrives during the booking stage. In our model, substitution may occur during each demand period.

Savin et al. (2004) propose a model that is tailor-made for studying the renting or leasing of capital equipment to multiple customer classes. As in the traditional yield management literature, their model focuses on a single type of capacity, but they formulate the problem as a queueing control problem and allow the rental period to be stochastic rather than uniformly fixed. In contrast, our model focuses on a single day, during which the inventory is depleted. We do not explicitly consider the differential value of customers who will rent for multiple days. Instead, we focus on the value of inventory substitution and the impact of sequential arrivals on the optimal initial capacity.

Researchers have also 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 Section 5, below). Articles by Ha (1997a, 1997b, 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 monotone in customer class. Research by de Véricourt et al. (2001, 2002) describes the benefits of optimal stock allocation for make-to-stock systems and present techniques to calculate optimal parameters for the allocation decision. Frank et al. (2004) 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. All of these papers consider single-item production systems, while we examine a system with multiple products.

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, while 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, while in the second subsection we present two related formulations and bounds on the objective function values 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 $j = 1, 2, ..., n$. Product quality *decreases* as index j increases, so that product j can be used to satisfy a customer of class i as long as $j \leq i$. This is often called 'one-way substitution' and is a common practice in many 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 $t = 1...T$ periods, where T is finite. Demand is independent between periods, although product demands within a period need not be independent. The realized demand in period t is denoted by d_i^t for product $i = 1, 2, ..., n$. The joint cumulative distribution function of a given set of random variables in period t is denoted $F^t(S)$, where $S \subseteq \{1, 2, ...n\}$, so the complete joint distribution function in period t is $F^t(1, 2, ...n)$. The marginal distribution function for product i is F_i^t. Let $\mathbf{D}^t = (d_1^t, d_2^t, ..., d_n^t)$ denote all realized demand in period t (in this paper, capitalized, bold-face characters represent vectors). We assume that each period's demand for a particular product is a non-negative integer, so that $\mathbf{D}^t \in \mathbb{Z}_n^+$. However, all of the following results hold when demand follows any distribution with non-negative support.

Before the first demand period the firm pays c_j for each unit of product j. When a customer arrives, she pays p_j for a product of type j (we assume that all price and cost parameters do not change over time). The firm may also pay a usage cost u_j when a unit is sold. That is, the firm pays c_j up-front, whether the product is sold or not, while u_j is only paid if the product is sold to a customer. 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, and inventory not sold after period T has no salvage value. We also assume that the time horizon is sufficiently short so that there is no discounting of costs or revenues across time periods.

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

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

(A2) $p_1 + v_1 > p_2 + v_2 > ... > p_n + v_n$;

(A3) $u_1 > u_2 > ... > u_n$.

Assumption (A1) states that only one-step upgrading is profitable. In practice, the profit margin accrued from multi-step 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 inventory, 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 have several implications. First, they imply that $\alpha_{jj} > \alpha_{kj}$ for all j and k, so that the maximum margin for product j is achieved by selling to customers of class j. Second, a class j customer should be upgraded when product j is in stock because the margin from a 'horizontal' sale is larger than the profit margin from any present or future upgrade.

Now we describe the firm's decision variables. Let $\mathbf{X}^t = (x_1^t, x_2^t, ..., x_n^t)$, $\mathbf{X}^t \in \mathbb{Z}_n^+$,be the vector of inventories at the beginning of period t, $t = 1, 2, ..., T$. After demand $\mathbf{D}^t$ appears, the firm must make inventory allocation decisions. Let $y_{ij}^t \in \mathbb{Z}^+$ be the quantity of product j allocated to class i demand after demand arrives in period t, and let $\overline{\mathbf{Y}}^t = (y_{ij}^t)$ be the allocation matrix for period t. Let $\Pi^{DYN}(\mathbf{X}^1)$ 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 inventory $\mathbf{X}^1$, while in periods 1 through T the firm allocates its inventory to maximize its revenue.

Dynamic Substitution Model (DYN)

Period 0:

$$\underset{\mathbf{X}^1 \in \mathbb{Z}_n^+}{\text{Max}} \Pi^{DYN}(\mathbf{X}^1) = \underset{\mathbf{X}^1 \in \mathbb{Z}_n^+}{\text{Max}} \{\Theta^1(\mathbf{X}^1) - \sum_j c_j x_j^1\} \tag{1}$$

Period t $(1 \leq t \leq T)$*:*

$$\Theta^t(\mathbf{X}^t) = \underset{\mathbf{D}^t}{\mathrm{E}}\{\underset{\overline{\mathbf{Y}}^t}{\mathrm{Max}}[G^t(\overline{\mathbf{Y}}^t, \mathbf{D}^t) + \Theta^{t+1}(\mathbf{X}^{t+1})]\}$$

where

$$\begin{aligned}
G^t(\overline{\mathbf{Y}}^t, \mathbf{D}^t) &= \sum_{i,j} \alpha_{ij} y_{ij}^t - \sum_i v_i d_i^t \\
\sum_j y_{ij}^t &\leq d_i^t & i = 1, 2, ..., n \\
\sum_i y_{ij}^t &\leq x_j^t & j = 1, 2, ..., n \\
x_j^{t+1} &= x_j^t - \sum_i y_{ij}^t & j = 1, 2, ..., n \\
y_{ij}^t &\in \mathbb{Z}^+ & i, j = 1, 2, ..., n
\end{aligned}$$

We also define $\Theta^{T+1} \triangleq 0$ under the assumption that the inventory has no salvage value. The value of $G^t(\overline{\mathbf{Y}}^t, \mathbf{D}^t)$ is the profit from the single-period capacity problem with substitution. The first inequality is period t's demand constraint, the second inequality is period t's supply constraint, and the last equality calculates the inventory available for period $t+1$ (we will refer to this equality as the 'linking constraint').

Note how this formulation corresponds to the daily problem faced by a rental car agency. First, the agency must decide on $\mathbf{X}^1$, the number of cars that should be available at the beginning of each day. Each time-period t represents an interval within the day (early morning, mid-morning, etc.), and there are T intervals. The unit cost c_j is the cost of having one automobile available for one day. This may include depreciation, as well as any costs associated with transporting the car from another location. The variable cost u_j, on the other hand, is the cost of an actual rental (e.g., wear and tear). The solution to the problem maximizes the expected profits over the entire day. We make the simplifying assumption that the initial inventory vector $\mathbf{X}^1$ is the same on each day, as would be the case if all cars were returned after one day's rental. The queueing control model developed by Savin et al. (2001) relaxes this assumption, although their model focuses on a single product and does not consider the effects of substitution among inventory levels.

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 inventories and $\Pi^{STC}(\mathbf{X})$ the profit function.

Single-period Substitution Model (STC):

$$\underset{\mathbf{X}\in\mathbb{Z}_n^+}{\text{Max}}\ \Pi^{STC}(\mathbf{X}) = \underset{\mathbf{X}\in\mathbb{Z}_n^+}{\text{Max}}\ \underset{\{D^1,D^2...,D^T\}}{\text{E}}[\Theta(\mathbf{X}) - \sum_j c_j x_j] \tag{2}$$

where

$$\Theta(\mathbf{X}) = \underset{\mathbf{Y}}{\text{Max}}[\sum_{i,j} \alpha_{ij} y_{ij} - \sum_i \upsilon_i \sum_t d_i^t]$$

$$\begin{aligned}
\text{s.t.} \quad \sum_j y_{ij} &\leq \sum_t d_i^t & i &= 1,2,...,n \\
\sum_i y_{ij} &\leq x_j & j &= 1,2,...,n \\
y_{ij} &\in \mathbb{Z}^+ & i,j &= 1,2,...,n
\end{aligned}$$

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 inventory occurs as the demand arrives (as in DYN) or after the Tth period (as in STC). In either case, the firm determines the optimal inventory x_j according to the newsvendor fractile and then sells the maximum amount of inventory possible.

Independent Newsvendor Model (NV):

$$\begin{aligned}
&\underset{\mathbf{X}\in\mathbb{Z}_n^+}{\text{Max}}\ \Pi^{NV}(\mathbf{X}) \\
= \ &\underset{\mathbf{X}\in\mathbb{Z}_n^+}{\text{Max}} \sum_j \{ \underset{\{d_j^1,d_j^2...,d_j^T\}}{\text{E}} \left[\alpha_{jj} \min(x_j, \sum_t d_j^t) - \upsilon_j (\sum_t d_j^t) \right] - c_j x_j \}
\end{aligned} \tag{3}$$

Next we compare the profits of the three models from the previous section.

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

Proof. First we show that $\Pi^{NV}(\mathbf{X}) \leq \Pi^{DYN}(\mathbf{X})$. NV allocates capacity to demand without substitution. Therefore, any allocation of inventory that is feasible in NV is also feasible in DYN, while DYN has the additional freedom to substitute products. Therefore, for any demand realization, DYN's profit is greater than, or equal to, the profit of NV, and $\Pi^{NV}(\mathbf{X}) \leq \Pi^{DYN}(\mathbf{X})$.

Next we show that $\Pi^{DYN}(\mathbf{X}) \leq \Pi^{STC}(\mathbf{X})$. In STC, inventory is allocated to customers after the firm observes all demand. Therefore, for a given demand realization $\mathbf{D}^1, \mathbf{D}^2, ..., \mathbf{D}^T$, any allocation decision available in DYN is also a feasible allocation in STC. In addition, there are allocation opportunities in STC that are not feasible in DYN. Therefore, for any demand realization STC's profit is greater than, or equal to, the profit of DYN, and $\Pi^{DYN}(\mathbf{X}) \leq \Pi^{STC}(\mathbf{X})$. ■

This proposition provides us with upper and lower bounds for the dynamic profit function. Because $\Pi^{DYN}(\mathbf{X}) \leq \Pi^{STC}(\mathbf{X})$ for any initial capacity $\mathbf{X}$, $\Pi^{DYN}(\mathbf{X}^{DYN}) \leq \Pi^{STC}(\mathbf{X}^{STC})$, where $\mathbf{X}^{DYN}$ and $\mathbf{X}^{STC}$ are the optimal initial capacity vectors. In qualitative terms, STC's profit function is an upper bound because in that setting there is no demand uncertainty when making the allocation decision.

4 The optimal policy: greedy allocation and then rationing

In this section we will show 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 inventory $\mathbf{X}^t = (x_1^t, x_2^t, ..., x_n^t)$ is available for sale in period t. Proposition 2 will show that we maximize $\Theta^t(\mathbf{X}^t)$ in DYN by following the following algorithm (henceforth we will refer to this procedure as the 'GRA,' for the 'Greedy-then-Rationing Algorithm').

Step 1: Let $y_{ii}^t = \min(d_i^t, x_i^t)$, $i = 1, 2, ..., n$. Satisfy as much class i demand with product i as possible.

Step 2: Let $\mathbf{N}^t$ be the net inventory after full 'parallel' allocation:

$$\mathbf{N}^t = \left(N_1^t, N_2^t, ..., N_n^t\right) = \left(x_1^t - d_1^t, x_2^t - d_2^t, ..., x_n^t - d_n^t\right).$$

Note that N_i^t can be positive if there is excess capacity, negative if demand exceeds capacity, or zero. For $k = 1, ..., n-1$, if $N_k^t > 0$ and $N_{k+1}^t < 0$, then let $y_{k+1,k}^t = N_k^t - \widetilde{p}$. The quantity $\widetilde{p} \in \left[\max(0, N_{k+1}^t + N_k^t), N_k^t\right]$ is the *protection limit*.

The rationale behind the GRA is straightforward. The profit margin from a 'horizontal' sale is larger than the profit margin from any present or future upgrade, so that in Step 1 any available capacity should be used to satisfy demand. To understand Step 2, note that a unit of product k should be upgraded if the current value of the upgrade, $\alpha_{k+1,k}$ is greater than the value of a unit of that product in the next period. Because the marginal value of product k declines as the quantity of product k rises (see Lemma 4, below), a threshold rule is optimal when choosing the number of units to upgrade. The threshold is the protection limit, and if the inventory of a product falls at or below the protection limit, the product will not be used to satisfy demand from a lower class.

To demonstrate rigorously that the GRA is an optimal policy, we must first derive a series of intermediate results about the GRA. First, Lemma 1 states that, after Step 1, the optimization problem breaks into smaller independent 'subproblems':

Lemma 1 *Suppose that at time t after completing Step 1 of the GRA, $\mathbf{N}_i^t \leq 0, i = k+1, \cdots k+j$, so that the inventories of these products have been depleted. Then the optimization problem can be separated into two independent problems: an upper part consisting of products 1 to $k+1$, and a lower part consisting of products $k+j+1$ to n.*

Proof. Given that only single-step upgrading is profitable, products with indices $1, 2, ...k$ will not be used to satisfy demand by classes $k+j+1, ...n$. Therefore, the assignment of products in one group does not affect the capacity or profits of the other group, and the global optimization problem is separable into the two subproblems. ■

In general, after Step 1, the global optimization problem may have been divided into numerous smaller subproblems, each defined by a series of positive net inventories (e.g., $\mathbf{N}_i^t > 0, i = j...k$) and

a single depleted inventory level for the lowest product ($\mathbf{N}^t_{k+1} \leq 0$). Therefore, for each subproblem created after step 1 of the algorithm, *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 step 1 of the GRA. The global optimization at the beginning of time t may be broken into smaller independent subproblems, with boundaries defined by depleted inventories, $x^t_i = 0$. To be explicit, define $B = \{(h^t_1, l^t_1), \cdots, (h^t_m, l^t_m)\}$ as the set of upper and lower limits for the subproblems at time t, i.e., (h^t_i, l^t_i) are the indices of the highest (smallest indexed) and lowest (largest indexed) products in the ith subproblem, so that $h^t_i \leq l^t_i$. Then the profit of the remaining optimization problem at time t, $\Theta^t(\mathbf{X}^t)$ in Equation (1) can be written as the sum of the profits from the subproblems:

$$\Theta^t(\mathbf{X}^t) = \sum\nolimits_{i=1}^{m} \Theta^t_i(\mathbf{X}^t_i) \tag{4}$$

where each subproblem $\Theta^t_i(\mathbf{X}^t_i)$ has the same formulation as $\Theta^t(\mathbf{X}^t)$, although the demand and capacity indices of each subproblem vary from h^t_i to l^t_i, rather than from 1 to n.

To keep the notation simple, for the remainder of this section we will derive the optimal assignment policy for an optimization problem $\Theta^t(\mathbf{X}^t)$ with product indices $i = 1...n$. 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.

Now consider an alternate formulation of $\Theta^t(\mathbf{X}^t)$ that is equivalent to the formulation in Equation (1). Let $\mathbf{Y}^t = (\sum_i y^t_{i1}, ..., \sum_i y^t_{in})$ be the vector of inventory offered to customers in period t. The capacity $\mathbf{Y}^t$ may be less than, or equal to, the available capacity $\mathbf{X}^t$. The jth element of $\mathbf{Y}^t$, denoted by $y^t_j = \sum_i y^t_{ij}$, is the quantity of product j offered for sale during period t. As defined above, $\overline{\mathbf{Y}}^t = (y^t_{ij})$ is the entire allocation matrix for period t.

$$\Theta^t(\mathbf{X}^t) = \mathop{\mathrm{E}}_{\mathbf{D}^t}\{\max_{\substack{\mathbf{Y}^t \in \mathbb{Z}^+_n, \mathbf{X}^{t+1} \in \mathbb{Z}^+_n \\ \mathbf{Y}^t + \mathbf{X}^{t+1} = \mathbf{X}^t}} [H^t(\mathbf{Y}^t, \mathbf{D}^t) + \Theta^{t+1}(\mathbf{X}^{t+1})]\} \tag{5}$$

$$
\begin{aligned}
\textit{where} \qquad H^t(\mathbf{Y}^t, \mathbf{D}^t) \;=\; & \underset{\overline{\mathbf{Y}}^t}{\text{Max}}[\sum_{i,j} \alpha_{ij} y_{ij}^t - \sum_i \upsilon_i d_i^t] \\
& \sum_j y_{ij}^t \le d_i^t \qquad\qquad i = 1, 2, ..., n \\
& \sum_i y_{ij}^t \le y_j^t \qquad j = 1, 2, ..., n \\
& y_{ij}^t \in \mathbb{Z}^+ \qquad\qquad i, j = 1, 2, ..., n.
\end{aligned}
$$

In this formulation, the conditions $\mathbf{Y}^t \ge 0, \mathbf{X}^{t+1} \ge 0$, and $\mathbf{Y}^t + \mathbf{X}^{t+1} = \mathbf{X}^t$ ensure that the supply and linking constraints are satisfied.

Given that inventory $\mathbf{Y}^t$ is available for sale in period t, and given demand realization $\mathbf{D}^t$, $H^t(\mathbf{Y}^t, \mathbf{D}^t)$ is a simple transportation problem with a cost structure defined by assumptions (A1) - (A3). Because the data $\mathbf{Y}^t$ and $\mathbf{D}^t$ are integer, the solution to H^t is integer. Bassok et al. (1999) point out that the cost structure of H^t corresponds to a Monge sequence (Hoffman, 1963), so that the following algorithm solves the problem.

Lemma 2 *Given* $\mathbf{Y}^t$ *and* $\mathbf{D}^t$*, the following algorithm solves* $H^t(\mathbf{Y}^t, \mathbf{D}^t)$*:,*
(i) $y_{ii}^t = \min(d_i^t, y_i^t), i = 1...n$
(ii) $y_{i+1,i}^t = \min\left(\left(d_{i+1}^t - y_{i+1}^t\right)^+, \left(y_i^t - d_i^t\right)^+\right), i = 1...n-1.$

This appears to be identical to the GRA: greedy assignment, followed by upgrading. However, we have not yet determined the optimal offered capacity $\mathbf{Y}^t$.

Now let $\Upsilon^t(\mathbf{X}^t)$ be the relaxation of $\Theta^t(\mathbf{X}^t)$ on the real numbers: $\Upsilon^t(\mathbf{X}^t)$ is identical to formulation (5) but with , $\mathbf{Y}^t \in \mathbb{R}_n^+, \mathbf{X}^{t+1} \in \mathbb{R}_n^+$, and $y_{ij}^t \in \mathbb{R}^+$.

Lemma 3 *The function* $\Upsilon^t(\mathbf{X}^t)$ *is concave on* $\mathbf{X}^t \in \mathbb{R}_n^+$.

Proof. The function $\Upsilon^T(\mathbf{X}^T)$ is concave in $\mathbf{X}^T$ because (i) a linear program is jointly concave in variables that determine the right-hand-side of its constraints and (ii) if $f(\mathbf{x}, \boldsymbol{\xi})$ is a concave function for all possible $\boldsymbol{\xi}$, then $E[f(\mathbf{x}, \boldsymbol{\xi})]$ is also concave (see van Slyke and Wets [1966], Proposition 7).

Now assume that $\Upsilon^{t+1}(\ \mathbf{X}^{t+1})$ is concave in $\mathbf{X}^{t+1}$. Because of fact (i) in the previous paragraph, $H^t(\mathbf{Y}^t, \mathbf{D}^t)$ is concave in $\mathbf{Y}^t$ for a given $\mathbf{D}^t$. Therefore, given demand $\mathbf{D}^t$, we maximize the sum

of two concave functions, $H^t(\mathbf{Y}^t, \mathbf{D}^t) + \Theta^{t+1}(\mathbf{X}^{t+1})$, with the constraint $\mathbf{Y}^t + \mathbf{X}^{t+1} = \mathbf{X}^t$. By theorems 5.3 and 5.4 in Rockafeller (1970) this maximal value is concave in $\mathbf{X}^t$. By fact (ii), above, the expected value taken over demand D^t, $\Upsilon^t(\mathbf{X}^t)$, is also concave in $\mathbf{X}^t$. ■

Lemma 3 also implies that the relaxation of $\Pi^{DYN}(\mathbf{X}^1)$ on the real numbers is concave.

We are now ready to show that the GRA is an optimal policy. Using the terminology of Porteus (1975), the set of admissible policies is defined by the constraints of $H^t(\mathbf{Y}^t, \mathbf{D}^t)$, $t = 1...T$, and the GRA defines an admissible structured policy. Because of the capacity constraints, all value functions $\Theta^t(\mathbf{X}^t)$ are finite. In the following Lemma we define a structured value function and show that the GRA attains the optimal value within each period and that the structured value function is preserved under optimization. For this Lemma, 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; $\Delta_k^t(\mathbf{X}^t)$ is the marginal value of one unit of product k.

Lemma 4 *Suppose that Θ^{t+1} has the following properties:*

1. *The GRA solves $\Theta^{t+1}(\mathbf{X})$.*
2. *$\Delta_k^{t+1}(\mathbf{X}) \leq \alpha_{kk}$.*
3. *$\Delta_k^{t+1}(\mathbf{X})$ is decreasing in x_k^{t+1}.*

Then properties (1)-(3) hold for Θ^t.

Proof. We will first show that the GRA attains the optimal value for Θ^t and then we show that properties (2) and (3) are preserved under optimization.

1. From Lemma 2, we know that a greedy/upgrade algorithm is optimal for any given $\mathbf{Y}^t$ and $\mathbf{D}^t$. To show that step 1 of the GRA is optimal, we must show that all available inventory in $\mathbf{X}^t$ is available for step 1's parallel assignment: $y_{ii}^t = \min(d_i^t, x_i^t), i = 1...n$. Suppose that $y_{ii}^t < \min(d_i^t, x_i^t)$. By Lemma 2, $\min(d_i^t, y_i^t) < \min(d_i^t, x_i^t)$, so that $y_i^t < x_i^t$. Therefore, $x_i^t - y_i^t$ units of product i are not used for parallel assignment but are instead held over to the next period. If any of these units had been used for a parallel assignment in period t, we would gain α_{kk} in immediate revenue and lose $\Delta_k^{t+1} \leq \alpha_{kk}$. Therefore, $y_{ii}^t = \min(d_i^t, y_i^t)$ is also an optimal solution.

To show that step 2 is optimal, we must prove the optimality of the rationing scheme for product k, the product at the 'bottom' of the subproblem. Specifically, after Step 1, Lemmas 1 and 2 imply that $N_1^t > 0, N_2^t > 0, ...N_k^t > 0, ... , N_{k+1}^t < 0$. We must determine how much product k capacity should be used to satisfy the unmet demand for product $k+1$. If p is the protection level for product k, we maximize over p,

$$A(p) = \alpha_{k+1,k}(N_k^t - p) + \Theta^{t+1}((N_1^t, N_2^t, ..., p, 0))$$

with the constraint, $\max(0, N_{k+1}^t + N_k^t) \leq p \leq N_k^t$. Because $\Delta_k^{t+1}(\mathbf{X}^{t+1})$ is decreasing in x_k^{t+1}, $A(p+1) - A(p)$ is decreasing in p. Let $\widetilde{p}$ be the unconstrained maximizer of $A(p)$. If $\widetilde{p} \leq Q$, then because the marginal value of $A(p)$ is decreasing, we should save as little inventory as possible and upgrade as much as possible: $x_k^{t+1} = Q$. If $0 \leq \widetilde{p} \leq Q$ then it is optimal to upgrade $N_k^t - \widetilde{p}$ and ration $\widetilde{p}$ units of product k: $x_k^{t+1} = \widetilde{p}$. If $\widetilde{p} > N_k^t$ it is not optimal to use any units of product k for upgrades: $x_k^{t+1} = N_k^t$. This shows that a rationing scheme and the GRA are optimal.

2. To show that property 2 is conserved under optimization, suppose that we begin time period t with capacity vector $\mathbf{X}$. Consider the following cases: (i) $N_k^t < 0$. If the initial capacity vector had been $\mathbf{X} + e_k$, the GRA algorithm would use the additional unit of product k for parallel assignment, and $\Delta_k^t(\mathbf{X}) = \alpha_{kk}$. (ii) $N_k^t \geq 0, N_{k+1}^t \geq 0$. Any extra unit of product k is passed along to the next period, and $\Delta_k^t(\mathbf{X}) = \Delta_k^{t+1}(\mathbf{X}) \leq \alpha_{kk}$. (iii) $N_k^t \geq 0, N_{k+1}^t < 0$. By the GRA algorithm, if $N_k^t + 1 \geq \widetilde{p}$, the extra unit of product k is used for an upgrade, and $\Delta_k^t(\mathbf{X}) = \alpha_{k+1,k} < \alpha_{kk}$. If $N_k^t + 1 < p$, the extra unit is passed along to the next period and $\Delta_k^t(\mathbf{X}) = \Delta_k^{t+1}(\mathbf{X}) \leq \alpha_{kk}$.

3. The optimization problem $\Theta^t(\mathbf{X}^t)$ is equivalent to $\Upsilon^t(\mathbf{X}^t)$ when evaluated with integer data. Lemma 3 indicates that $\Upsilon^t(\mathbf{X}^t)$ is concave in $\mathbf{X}^t$, and therefore $\Delta_k^t(\mathbf{X})$ is decreasing in x_k^t for any $\mathbf{X}$.

■

Proposition 2 *The GRA is an optimal policy from among all admissible policies.*

Proof. Consider the last-period problem, $\Theta^T(\mathbf{X})$. Given that $\Theta^{T+1} = 0$, arguments identical to those in the proof of Lemma 4 show that $\Delta_k^T(\mathbf{X}) \leq \alpha_{kk}$ and $\Delta_k^T(\mathbf{X})$ is decreasing in x_k^T. In addition, the greedy algorithm defined by Hoffman (1963) solves $\Theta^T(\mathbf{X})$, and is a special case of the GRA, with protection limits $\widetilde{p} = 0$. Therefore, the argument of Lemma 4 iterates backwards through $T, T-1, ...1$. ■

The proof for Lemma 4 implies that the optimal protection limit, $\widetilde{p}$, is the smallest value of p such that

$$\Delta_k^{t+1}((N_1^t, ..., N_{k-1}^t, p, 0)) < \alpha_{k+1,k} \tag{6}$$

The marginal value Δ_k^{t+1} depends upon the time period, the current inventories of all products, and the distribution of future demand, and can be difficult to calculate. In the next section we will consider methods for efficiently approximating $\widetilde{p}$.

5 Properties of the protection limits: monotonicity and bounds

In this Section we first find a bound on the amount of data needed to calculate each protection limit. Then we show that the protection limits are monotonic in both the amount of inventory and time, and we use these properties to derive a series of bounds on the protection limits. This Section ends with a description of numerical experiments that demonstrate the quality of the bounds.

5.1 Bounds on the problem size

In the previous section we saw that each rationing problem is associated with a single subproblem. Here we ask: how much data are needed to find the optimal answer for the rationing problem? Two sets of data are certainly relevant: the current capacity in the subproblem and the distribution of future demand. But is the rationing decision dependent upon the capacities and future demands of all products? To answer this question, we define the *size* of a rationing problem as the number of products with data that are relevant to the rationing problem. In the following definition, we use the subproblem indexing scheme introduced in the paragraph before equation 4.

Definition: Let S_i^t be a subset of the product indices for subproblem i at time t: $S_i^t \subseteq \{h_i, ...l_i\}$. Define the set of capacity and demand information

$$\Phi_i^t = \{ \bigcup_{j \in S_i^t} N_j^t, F^{t+1}(S_i^t), F^{t+2}(S_i^t), ..., F^T(S_i^t) \}$$

(recall that $F^t(S)$ is the joint distribution function for products in the set S at time t). Then the *size* of the rationing problem is the cardinality of the smallest set S_i^t such that Φ_i^t is always sufficient to solve optimally the rationing problem.

While the notation may be convoluted, the meaning is simple: the size of the rationing problem indicates how many products influence our upgrading decision in subproblem i at time t. First, we know that the size of the rationing problem is limited by the size of the subproblem. Given $l_i^t - h_i^t + 1$ products in the subproblem, the size of the rationing problem is at most $l_i^t - h_i^t + 1$. The following proposition indicates that the size can also be limited by the number of time periods left.

Proposition 3 *Given a subproblem with $\gamma = l_i^t - h_i^t + 1$ products and $\tau = T - t + 1$ time-periods left, then the size of the rationing problem is at most* $\min[\gamma, \tau + 1]$.

Proof. This result follows from the assumption that only single-step upgrades are profitable, so that each time-period in the future connects the current upgrade decision with one additional product. See Appendix A for details ■

5.2 Bounds on the protection limits

Let $\widetilde{p}^t$ be the protection limit for a subproblem at time t. Next, we show that the optimal protection level $\widetilde{p}^t$ is monotone in the inventory state and over time.

Proposition 4 *The optimal protection limit $\widetilde{p}^t$ is decreasing in the state vector $\mathbf{X}^t$.*

Proposition 5 *The optimal protection limit $\widetilde{p}^t$ is decreasing at t rises.*

Proof. For proofs of both propositions, see Appendix B. ■

The proofs of Propositions 4 and 5 show that the marginal value of a product declines if more higher-level inventory is available or if time has elapsed. These results lead directly to sets of upper and lower bounds on the protection limits. Suppose we have a subproblem involving

products $1...k+1$. Let $\widetilde{p}(\mathbf{X})$ be the optimal protection limit of product k, given initial capacity vector $\mathbf{X} = (x_1, ..., x_k, 0)$ (for clarity, we suppress the superscript t). Define a new, truncated capacity vector $\mathbf{X}(i, C) = (C, x_{k-i}, ., x_k, 0)$, $i = 0...k-1$ (if $i = 0$,, then the capacity vector is just (C, x_k)). Setting $C = 0$ indicates that there is no inventory 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). Note that $\mathbf{X}(i, 0)$ and $\mathbf{X}(i, \infty)$ define two smaller subproblems that involve $i+2$ products. Specifically, product $k+1$ has been completely depleted, product k may be rationed, and there are i products with nonzero capacities, products $k-i...k-1$, that may affect the optimal protection level of product k. The capacities $(x_1, ..., x_{k-i-2})$ have no impact on the rationing problem because products $k-i...k$ are 'cut off' by the 0 or infinite inventory of product $k-i-1$. This observation motivates the following set of bounds.

Proposition 6 *For a subproblem with k products,*

$\widetilde{p}(\mathbf{X}(0, \infty)) \leq \widetilde{p}(\mathbf{X}(1, \infty)) \leq ... \leq \widetilde{p}(\mathbf{X}(k-1, \infty))$

$\leq \widetilde{p}(\mathbf{X})$

$\leq \widetilde{p}(\mathbf{X}(k-1, 0)) \leq \widetilde{p}(\mathbf{X}(k-2, 0)) \leq .. \leq \widetilde{p}(\mathbf{X}(0, 0))$.

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

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 1, an inexhaustible supply of inventory 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 $1...k-i-2$. This fact and Proposition 4 imply that for $0 < i \leq k-1$, $\widetilde{p}(\mathbf{X}(i, \infty)) = \widetilde{p}(\infty, x_{k-i}, ., x_k, 0) \geq \widetilde{p}(\infty, \infty, x_{k-i+1}, ., x_k, 0) = \widetilde{p}(\infty, x_{k-i+1}, ., x_k, 0) = \widetilde{p}(\mathbf{X}(i-1, \infty))$. ■

These bounds are useful because the dimensionality of the dynamic program rises with the number of products in the subproblem. These propositions allow us to restrict our attention to a

small subset of products and then produce a range of possible protection limits. The tightness of the bounds rises with the number of products included in the calculations.

5.3 Protection limit bounds: numerical experiments

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 $\widehat{x}$ for the inventory of each product, there are $O(T\widehat{x}^{k-2})$ distinct protection limits to calculate (with $T = 10$, $\widehat{x} = 100$, and $k = 5$, there are over 10 million protection limits). However, Proposition 6 provides us with a series of bounds that allow for a trade-off between accuracy and computational burden. Here we describe numerical experiments designed to test the quality of the two 'loosest' bounds: those bounds determined by the quantity of inventory for a single adjacent product ($\widetilde{p}(\mathbf{X}(1,0))$ and $\widetilde{p}(\mathbf{X}(1,\infty))$) and those bounds determined by two adjacent products ($\widetilde{p}(\mathbf{X}(2,0))$ and $\widetilde{p}(\mathbf{X}(2,\infty))$). There are $O(Tk\widehat{x})$ protection limits to find in the first set of bounds, and $O(Tk\widehat{x}^2)$ in the second set, so that both can be found quickly for reasonably large problems.

In the numerical experiments that follow, we calculate the gaps $\nabla_1(\mathbf{X}) \equiv \widetilde{p}(\mathbf{X}(1,0)) - \widetilde{p}(\mathbf{X}(1,\infty))$ and $\nabla_2(\mathbf{X}) \equiv \widetilde{p}(\mathbf{X}(2,0)) - \widetilde{p}(\mathbf{X}(2,\infty))$ 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 inventory of at most 2 products). Specifically, we ran 144 experiments over a wide variety of parameter configurations. For each experiment we assumed that $k = 5$, $T = 10$, and $\widehat{x} = 30$. We also assume that demands arrive according to Poisson distributions that are independent between demand periods and between products. Here we describe briefly the ranges of the other parameter values. A full description of each of the 144 experiments is available from the authors.

1. *Distribution of capacity across products.* We varied the number of initial units of each product. For example, the initial capacities of products 1, 2 and 3 followed four scenarios: (i) $[30, 20, 10]$, (ii) $[20, 20, 20]$, (iii) $[10, 20, 30]$ and (iv) $[10, 30, 20]$.

2. *Distribution of demand across products.* We varied the total demand for each product, summed over all 10 time periods. For example, the total demand of products 1 varied from 10 to 50 while the total demand for product 5 varied from 50 down to 10.

3. *Distribution of demand over time.* We defined four general scenarios: (i) constant demand for all products, (ii) demand increases over time for all products, (iii) demand decreases for all products, and (iv) demand increases for high-value products (1, 2 and 3) and decreases for low-value products (4 and 5). Scenario (iv) corresponds to the demand pattern that is often seen by airlines and rental-car firms.

4. *Revenue pattern and the value of upgrades.* We defined three scenarios: (i) Large upgrade value for all products (e.g., $\alpha_{11} = 15, \alpha_{22} = 14$, and $\alpha_{21} = 7$), (ii) Small upgrade value for all products (e.g., $\alpha_{11} = 15, \alpha_{22} = 14$, and $\alpha_{21} = 3$), and (iii) large upgrade value for products $1, 2$ and 3, but small upgrade value for product 4.

Combining all of these parameter values produced 144 numerical experiments. These experiments yielded 36,936 one-product protection levels and 1,076,976 two-product protection levels for product 4. In both cases, over 99.5% of the bounds had no gap, and for both sets of protection levels, the maximum gap was just 1 unit. In fact, for $\nabla_1(\mathbf{X})$ only 78 out of approximately 37,000 gaps were 1 rather than 0, while for $\nabla_2(\mathbf{X})$ only 4 gaps out of over a million were 1. Table 1 contains statistics on the distribution of the sizes of the observed gaps.

	# observations	% gap=0	% gap=1	maximum gap
$\nabla_1(\mathbf{X})$	36,936	99.79	0.21	1
$\nabla_2(\mathbf{X})$	1,076,976	99.9996	0.0004	1

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

Therefore, for these experiments, either of the two-product bounds is equivalent to the optimal solution, and the one-product bounds are quite close. In fact, using $\widetilde{p}(\mathbf{X}(1,0)$ as the protection limits produced a negligible decline in expected revenue. We found that the average decline in profit when using the one-product upper bound rather than either two-product bound (essentially, the optimal solution) is $8.62 \times 10^{-7}\%$ and the maximum difference is 0.0022%.

The accuracy of the heuristic protection limits based on these bounds, and the relative ease with which these 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. We will discuss this opportunity in Section 7.

6 Optimal capacity for the static and dynamic 2x2 model

The single-period (static) model has been a popular framework for exploring the impact of flexibility on the optimal level of capacity investment. Using a single-period model, Bassok et al. (1999) and Netessine et al. (2002) show that the optimal level of flexible, class-1 capacity is higher than the optimal level if that product were not available for upgrades (i.e., higher than the newsvendor quantity). Likewise, they show that the optimal level of the lowest-class capacity is lower than the equivalent newsvendor quantity, because customers for the lowest-class product can be upgraded. This section compares optimal capacities for the static model, STC, the dynamic model, DYN, and the newsvendor quantities. In this Section, we assume that each period's demand and capacities are non-negative real numbers: $\mathbf{D}^t \in \mathbb{R}_n^+$ and $\mathbf{X}^t \in \mathbb{R}_n^+$.

6.1 Protection Limits in the 2x2 model

Because it is prohibitively unwieldy to derive and analyze expressions for rationing policies and optimal capacities of the n-product, T-period model, here we examine the simplest possible model that retains both the product flexibility and the dynamic nature of the general model: a model with two products and two time-periods (the '2x2 model'). Given that capacity is continuous and that the profit function is differentiable, the optimal protection p^* limit must satisfy,

$$\frac{P(d_1^2 + d_2^2 \leq p^*)}{P(d_1^2 > p^*)} = \frac{\alpha_{11} - \alpha_{21}}{\alpha_{21}}. \tag{7}$$

One might think of the ratio $\beta = \alpha_{11}/\alpha_{21}$ as a measure of the cost of supply cannibalization. Because the left-hand side of (7) is increasing in p^*, and because the right-hand side of (7) is equal to $\beta - 1$, p^* increases with β. This makes sense: as the cost of supply cannibalization increases, the protection limit should increase, as well.

One can also show that the optimal protection limit p^*, (i) increases as revenue $p_1 + v_1$ increases, (ii) decreases as revenue $p_2 + v_2$ increases, (iii) increases as usage cost u_1 increases (as the usage

cost u_1 rises, the firm is less willing to release expensive capacity to less-lucrative customers), and (iv) increases as second-period demand for either product rises (according to the usual stochastic order). See Shumsky and Zhang (2003) for details.

6.2 Optimal Capacities in the 2x2 model

One might think of the static model is a best case, for the firm is able to gather all demand information and then allocate capacity optimally. Because in the dynamic model the firm is forced to make allocation decisions before all customers have arrived, flexibility may not be used optimally. Therefore, a reasonable prediction is that the solution to the dynamic model should have smaller investments in the highest-class capacity and larger investments in the lowest-class capacity, as compared to the static model. In general, our analysis and numerical experiments confirm this prediction, although there can be exceptions. In fact, given certain parameters, it may be optimal to have more class-1 capacity in the dynamic case than in the static case.

Recall that the objective function of the static model, $\Pi^{STC}(x_1, x_2)$, is defined in equation (2) and that the dynamic model, $\Pi^{DYN}(x_1, x_2)$, is defined in equation (1). Let (x_1^{STC}, x_2^{STC}) and (x_1^{DYN}, x_2^{DYN}) be the optimal capacities for each of these models. Appendix C contains explicit formulations for these models as well as first-order conditions for the optimal capacities. These first-order conditions lead to the following result, and the proof of this Proposition is also included in Appendix C.

Proposition 7 *In the 2x2 case,* $\partial\Pi^{DYN}(x_1, x_2)/\partial x_2 \geq \partial\Pi^{STC}(x_1, x_2)/\partial x_2$ *for any capacities* x_1 *and* x_2*.*

Proposition 7 indicates that the marginal value of an additional unit of type-2 inventory is more valuable in the dynamic environment than in the static environment. The terms of the partial derivative $\partial\Pi^{DYN}(x_1^*, x_2^*)/\partial x_2$ in Appendix C suggest why: extra type-2 capacity can be useful for protecting 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. While Proposition 7 is not sufficient to show that $x_2^{DYN} \geq x_2^{STC}$, 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, $x_2^{DYN} \geq x_2^{STC}$. We describe examples of these experiments below.

There is no analogue of Proposition 7 for type-1 capacity: $\partial \Pi^{DYN}/\partial x_1 \lesseqgtr \partial \Pi/\partial x_1$. In addition, we will see examples below in which $x_1^{DYN} \leq x_1^{STC}$ and $x_1^{DYN} > x_1^{STC}$.

In the following numerical experiments we assume that all demands are normally distributed and truncated at 0, although the coefficient of variation will be sufficiently small so that truncation does not significantly affect the results. For the STC model, we assume that the total type-1 and type-2 demands are distributed with mean $\mu_i^1 + \mu_i^2 = 100$ and standard deviations $\sigma(D_i^1 + D_i^2) = 30, i = 1, 2$. For DYN, when we split demand between the first and second periods, we will hold these total-demand parameters constant. Specifically, if a proportion r of type-i demand occurs in the first period, then $D_i^1 \sim N(100r, 30\sqrt{r})$ and $D_i^2 \sim N(100(1-r), 30\sqrt{(1-r)})$, so that the standard deviation of the *total* demand is 30. In the first set of experiments described here, the revenue and cost parameters are $\alpha_{11} = 40, \alpha_{21} = 15, \alpha_{22} = 20$, $c_1 = 12$, and $c_2 = 10$. These parameters imply that the newsvendor critical ratios for type-1 and type-2 are 0.7 and 0.5, respectively.

The numerical experiments examine four models: NV, STC, DYN, and the dynamic model with no rationing (protection level $p = 0$). The first-order conditions for STC and the dynamic models are described in Appendix C, and the solution to the newsvendor problem is well known. The optimal capacities of each model were found numerically, using Monte Carlo Integration and a simple search procedure.

We found that optimal capacities for the static and dynamic models diverged significantly when (i) a majority of type-2 demand occurs in the first period and (ii) a majority of type-1 demand occurs in the second period. Therefore, in the dynamic model we 'unbalance' the demand to emphasize this point. Given that r is the proportion of type-2 demand in the first period and $1-r$ is the proportion of type-1 demand in the first period, we varied r from 0.4 to 1.

For example, when $r = 0.5$, demands for both products are distributed equally between periods. In this case there is almost always insufficient demand in the first period of the dynamic model to require any upgrading, so that there is little risk of supply cannibalization, type-1 capacity is rarely

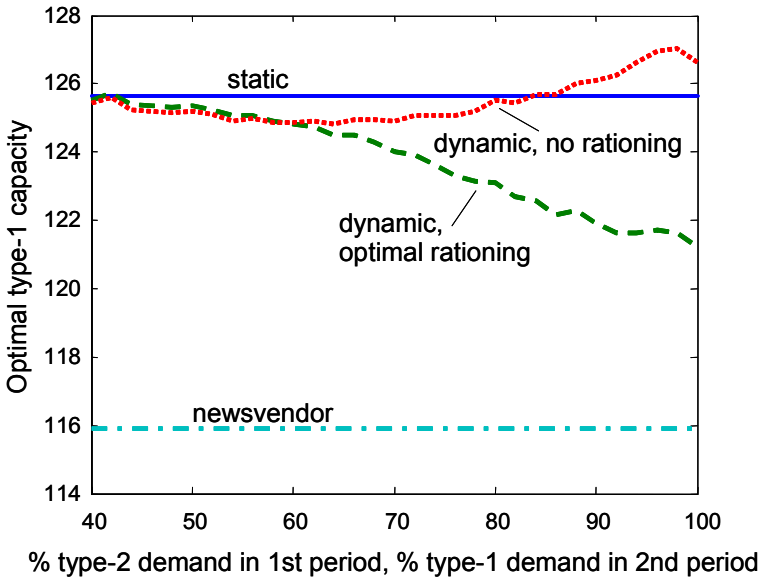

Figure 1: Optimal type-1 capacity

rationed, the particular rationing policy does not matter, and there is little difference between the static and dynamic models. However, as r rises, the early appearance of type-2 demand and the late appearance of type-1 demand forces the firm to either upgrade type-2 demand or ration type-1 products. The model with $r = 1$ is analogous to the standard yield management problem, in which low-fare passengers arrive first, followed by high-fare passengers.

Figures 1 and 2 show the optimal type-1 and type-2 capacity values, respectively, for each model. In Figure 1 the dynamic model's optimal type-1 capacity, x_1^{DYN} is consistently below the optimal capacity from the static model, x_1^{STC}, although we have found that the opposite can be true (see below). A more pronounced pattern is shown in Figure 2, where we see that the optimal type-2 capacities can be significantly higher in the dynamic model ($x_2^{DYN} \geq x_2^{STC}$). The extra type-2 capacity acts as a buffer to prevent cannibalization of more lucrative type-1 capacity. This role for type-2 capacity is particularly important when there is no rationing, thus inflating the optimal type-2 capacity.

To see that it is possible to have $x_1^{DYN} > x_1^{STC}$, consider an experiment with the following revenue and cost parameters: $\alpha_{21} = 4, \alpha_{22} = 5$, and $c_2 = 1$ (we will try a variety of values for both

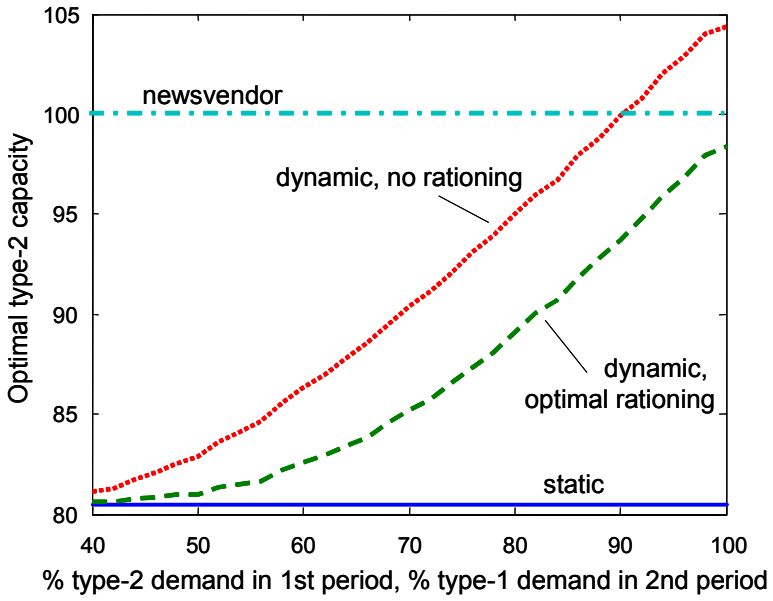

Figure 2: Optimal type-2 capacity

α_{11} and c_1). The total demands are still $N(100, 30)$, and we assume that $r = 1$, so that in the dynamic model there is no type-1 demand in the first period and no type-2 demand in the second period. The parameters α_{22} and c_2 imply that the newsvendor problem's critical ratio is 0.8 for product 2. This ratio will be substantially higher for product 1 in the following examples, for we will vary α_{11} from 5 to 80 and will use two low values of c_1: 1.5 and 0.5. The second value indicates that the initial purchase cost of product 1 is less than the cost of product 2, although the usage cost may be significantly greater for product 1 than product 2.

Figure 3 shows the optimal type-1 capacities from the dynamic and static procedures, x_1^{DYN} and x_1^{STC}. Here the optimal dynamic type-1 capacities are higher than the optimal static capacities. This difference is again caused by the problem of supply cannibalization in the dynamic case. For demand realizations in which cannibalization occurs, an additional unit of type-1 product always has the marginal value $\alpha_{11} - \alpha_{21}$ in the dynamic case, but may have no value in the static case. This effect is largest when the profitability of a type-1 sale is greatest, i.e., when α_{11} is large and when c_1 is low. In addition, this risk of supply cannibalization is even greater when protection limits are lowered. If there is no rationing, the differences between the optimal dynamic and static

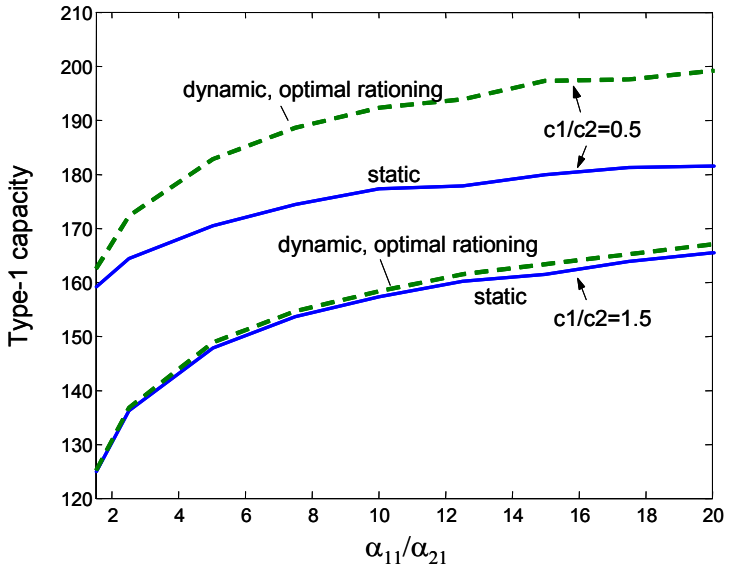

Figure 3: Optimal type-1 capacity can be larger in the dynamic model

capacities are consistently larger than the differences seen in Figure 3.

7 Conclusions and Further Research

In this paper we formulate a flexible capacity investment and allocation problem in which demand arrives over a sequence of discrete time intervals. Because total demand from the most lucrative customers is uncertain when inventory allocation decisions must be made, the firm may hold back, or ration, some products before the last time interval. We show that the optimal assignment policy involves two steps: greedy allocation, followed by upgrading that is limited by a protection limit. We show that the protection limits satisfy certain reasonable properties: the protection limits decrease as inventories increase, and the protection limits decrease over time. We use these properties to derive simple bounds on the protection limits.

Previous work on capacity investment decisions and the impact of flexibility have focused on static models. Previous results have shown that the optimal quantity of flexible (dedicated) capacity is higher (lower) than the optimal newsvendor quantities. We have found that for the

2x2 dynamic model, the optimal capacities can be pushed back toward the newsvendor quantities, although there are cases in which the optimal level of flexible capacity can be greater in the dynamic case than in the static case. We have also seen that the ordering of demand arrival is a significant factor, for differences between the static and dynamic model are greatest when low-class demand tends to arrive first and high-class demand arrives later.

Given the promising results showing that the bounds can lead to accurate approximations of the optimal protection levels, we expect to use heuristic protection levels based on the bounds (e.g., the average of the bounds) to examine problems with large numbers of periods and products. Using these heuristics, we will explore the impact on total expected profit of reducing the number of demand periods, which is roughly equivalent to gathering advance demand information. We will also examine the consequences of using sub-optimal policies, such as a greedy policy ('upgrade whenever possible') and a no-upgrade policy that separates the problem into simple newsvendor problems. Finally, the heuristics may allow us to consider optimal capacity levels for problems larger than the 2X2 problem of Section 6.

There are also many possible extensions to the model, such as the inclusion of backlogging or discounting, and incorporating inter-period demand dependence that would allow the firm to update protection levels as demand arrives. 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.

Finally, in many real-world environments customer arrivals cannot be divided into time-periods, and an important extension of the analysis would be to compare our dynamic model with a model that features continuous arrivals (e.g., type-1 and 2 customers arrive according to a Poisson or diffusion process). However, the multi-period model approximates a continuous models as the number of periods increases. In addition, a model with a small number of discrete demand periods may be reasonable approximation when different customer classes tend to arrive in different periods, as is often the case in yield management applications.

Appendix A: Proposition 3

Proof: From Lemma 1 we know that rationing decisions within the subproblem involve only

products within that subproblem, so that the rationing problem is bounded by the number of products, γ. To prove that the size is bounded by $\tau + 1$, we will use induction, which begins with a base case showing that when $\tau = 2$, the size of the rationing problem is at most 3. Denote the two periods by 1 and 2. Without loss of generality, let the indices for the products be $1, 2, ..., k$, where $k = \gamma > \tau + 1$, so $k > 3$. Suppose that after Step 1 products $1, 2, ..., k-1$ have net capacities N_1^1, N_2^1, ..., $N_{k-1}^1 > 0$, while $N_k^1 < 0$ so that there is extra demand for the lowest-class product to be upgraded. The rationing problem is this: what portion of N_{k-1}^1 should be used to satisfy class-k demand, and what portion should be 'protected' for class $k-1$ demand in period 2? Let p ($p \leq N_{k-1}^1$) be the quantity of N_{k-1}^1 that is not used in the first period, and passed to the second period. Then the expected profit, to be maximized over the decision variable p, is

$$\begin{aligned}
\Gamma(p) \;=\; & \alpha_{k,k-1} \min[\left|N_k^1\right|, (N_{k-1}^1 - p)] + \sum\nolimits_{i=1}^{k-2} \alpha_{ii} E[\min(d_i^2, N_i^1)] \\
& + \sum\nolimits_{i=1}^{k-3} \alpha_{i+1,i} E\{\min[(d_{i+1}^2 - N_{i+1}^1)^+, (N_i^1 - d_i^2)^+]\} \\
& + \alpha_{k-1,k-1} E[\min(d_{k-1}^2, p)] + \alpha_{k-1,k-2} E\{\min[(d_{k-1}^2 - p)^+, (N_{k-2}^1 - d_{k-2}^2)^+]\} \\
& + \alpha_{k,k-1} E[\min(d_k^2, (p - d_{k-1}^2)^+].
\end{aligned}$$

The first term on the right-hand side is the profit from upgrading in period 1, and the remaining terms represent profits from period 2. Our decision variable, p, does not appear in the second and third terms. The information in the remaining terms involve only $N_k^1, N_{k-1}^1, N_{k-2}^1$ and the demand realizations of products $k, k-1$, and $k-2$. Therefore, the information needed to solve the rationing problem is $\Phi^1 = \{N_k^1, N_{k-1}^1, N_{k-2}^1, F^2(k, k-1, k-2)\}$ and the size of the rationing problem is at most 3.

Next we show that if the Proposition is true for $\tau = 2, 3, ...k$ then it is also true for $\tau = k + 1$. First we specify the induction assumption, with the current time index = 1 and with $\tau = k$ time periods to go. Let the indices for the products in the subproblem be $1, 2, ..., k+2$. The induction assumption is that the smallest information set that is always sufficient to solve the rationing problem is

$$\Phi^1 = \{N_2^1, N_3^1, ..., N_{k+2}^1, F^2(2, 3, ..., k+2), ..., F^T(2, 3, ..., k+2)\}.$$

If the number of products is greater than $k + 2$, the induction assumption is that when $\tau = k$, the smallest information set that is always sufficient to solve the rationing problem includes only

the 'bottom' $k+1$ products. Therefore, the size of the problem with $\tau = k$ is at most $k+1$ and product 1 is not included in the information set Φ^1.

Now consider the problem with $\tau = k+1$ and product indices $1, 2, ..., k+3$ (again, it is easy to extend this logic when the number of products is greater than $k+3$). In the first period, the remaining capacity after Step 1 is $(N_1^1, N_2^1, ..., N_{k+2}^1, N_{k+3}^1)$, where $N_i^1 > 0$ for $1 \leq i \leq k+2$ but $N_{k+3}^1 < 0$. We want to show that the size of this problem is $k+2$. As in the base case, the objective function for the rationing problem includes the profit in the current period, $\alpha_{k+3,k+2} \min[\left|N_{k+3}^1\right|, (N_{k+2}^1 - p)]$, plus expected profits from the subproblem in future periods. Given that there are $k+3$ products, the proof is complete if we can show that product N_1^1 and its future demands $d_1^2, d_1^3, ...d_1^T$ will have no influence on the rationing decision in the current period. This implies that product 1 is not included in the information set Φ^1, and therefore the size of the rationing problem is limited by the remaining $k+2$ products.

After making the upgrade decision for period 1, we calculate the capacities available for period 2, observe realized demand in period 2, and perform the parallel allocation in Step 1. For this new subproblem three scenarios are possible: (i) At least one class of product of a higher class than $k+2$ runs out in period 2: $N_i^2 < 0$ for some $i = 1, ..., k+1$. Then the problem splits into two subproblems and product 1 will have no impact on the rationing made in period 1; (ii) All classes higher than $k+2$ have extra capacity $(N_i^2 > 0, i = 1, ..., k+1)$, but $N_{k+2}^2 < 0$. From this point on, demand for product $k+3$ cannot be filled and product $k+3$ drops out of the subproblem. Therefore, the new period-2 subproblem is identical to the subproblem in the induction assumption: $\tau = k$,and the subproblem includes only products $1, 2, ...k+2$. Let Φ^2 be the information set for this period-2 rationing problem. From the induction assumption, Φ^2 does not include product 1, and therefore product 1 is not included in Φ^1; (iii) $N_i^2 > 0, i = 1, ..., k+2$. In this case, we still have a rationing problem with demand from products $i = 1, ..., k+3$ and supply from products $i = 1, ..., k+2$, but now with $\tau = k$. The induction assumption states that Φ^2 includes only the 'bottom' $k+1$ products, $i = 3, ..., k+3$. Therefore, Φ^2 does not include product 1, and neither does Φ^1. ■

Appendix B: Monotonicity results

Proposition 4 The optimal protection limit $\widetilde{p}^t$ is decreasing in the inventory vector $\mathbf{X}^t$.

Proof: Consider two subproblems in time period t, and without loss of generality assume that the subproblem's product indices are $1, ..., k+1$. Before step 1, the first subproblem has inventories $\mathbf{X}^t$, where $x_i^t > 0, i = 1, ..., k$, and $x_{k+1}^t = 0$. The second subproblem has inventories $\widehat{\mathbf{X}}^t = \mathbf{X}^t + \mathbf{e}_j$ with $1 \leq j \leq k-1$. Let $\Delta_k^t(\mathbf{X}^t)$ be the marginal value of an additional unit of product k in time-period t, given inventory $\mathbf{X}^t$. To prove that the proposition is true, we proceed by backwards induction, with two induction assumptions: (i) the optimal protection limit $\widetilde{p}^t$ is decreasing in the inventory vector $\mathbf{X}^t$ (this is the Proposition) and (ii) in the *next* time-period, the marginal value of product k is decreasing in the capacity vector. That is, $\Delta_k^{t+1}(\widehat{\mathbf{X}}^{t+1}) \leq \Delta_k^{t+1}(\mathbf{X}^{t+1})$ for $\widehat{\mathbf{X}}^{t+1} = \mathbf{X}^{t+1} + \mathbf{e}_j, 1 \leq j \leq k-1$.

Before showing that the induction assumptions are true for all t, we first prove that assumption (ii) implies (i). Recall that the protection limit $\widetilde{p}^t$ solves a concave optimization problem in one variable, with the solution specified by condition (6). The left-hand-side of (6) is the marginal value of an increase in the quantity of product k made available in the next period. Therefore, the protection limit rises or falls as the marginal value of product k in the next period rises or falls. Furthermore, if $\widehat{\mathbf{X}}^t = \mathbf{X}^t + \mathbf{e}_j$ for some $1 \leq j \leq k-1$, then $\widehat{x}_j^{t+1} \geq x_j^{t+1}$, because the extra capacity of the higher-level product is either passed along or used to satisfy demand in period t. Therefore, given induction assumption (ii), an increase in $\mathbf{X}^t$ may lead to a decrease in the marginal value of product k in the next period, and $\widetilde{p}^t$ is decreasing in $\mathbf{X}^t$.

Now consider the rationing problem at time T. Netessine et al. (2002) show that the profit function of the single-period model with one-step upgrading is submodular in its capacity $\mathbf{X}$. In other words, the marginal value of a product is decreasing in the quantity of any other product. Therefore, $\Delta_k^T(\widehat{\mathbf{X}}^T) \leq \Delta_k^T(\mathbf{X}^T)$ for $\widehat{\mathbf{X}}^T = \mathbf{X}^T + \mathbf{e}_j$, $1 \leq j \leq k-1$. From the discussion in the last paragraph, this also implies that the optimal protection limit $\widetilde{p}^{T-1}$ is decreasing in the inventory vector $\mathbf{X}^{T-1}$.

Assume that induction assumptions (i) and (ii) hold for periods t and $t+1$, respectively, and we will show that (ii) is true for t and therefore (i) is true for $t-1$. Given a realization of demand in period t, $\mathbf{D}^t$, after Step 1 we are left with the net capacity vectors $\mathbf{N}^t = \mathbf{X}^t - \mathbf{D}^t$ and $\hat{\mathbf{N}}^t = \widehat{\mathbf{X}}^t - \mathbf{D}^t$ (note $\mathbf{N}^t$ and $\hat{\mathbf{N}}^t$ only differ in the jth element, and by one unit). To find the marginal value of

an extra unit of product k, we must consider a variety of scenarios. In each of these cases, an extra unit of product k may be used for one of three things. The unit may be used for a parallel assignment to a customer of class k (denoted by '$\underrightarrow{k}$' and '$\underrightarrow{\widehat{k}}$' given $\mathbf{N}^t$ and $\hat{\mathbf{N}}^t$, respectively), it may be used to upgrade a customer of class $k+1$ (denoted '$\underrightarrow{k+1}$' and '$\underrightarrow{\widehat{k+1}}$') and it may not be used in period t but passed along to period $t+1$ (denoted '$\underrightarrow{t+1}$' and '$\underrightarrow{\widehat{t+1}}$'). Before cataloguing an exhaustive list of scenarios, we consider the following observation:

Observation: Suppose that in period t, $N_k^t > 0$, and that the extra unit of product k is not allocated in period t but is passed along to the next period ('$\underrightarrow{t+1}$'). Then one of the following must be true:

Case A: We have the event $\underrightarrow{t+1}$ because all excess type-$(k+1)$ demand has been upgraded and the protection limit has not yet been reached. In this case $\Delta_k^t(\mathbf{X}^t) \leq \alpha_{k+1,k}$ because the quantity of available capacity is larger than the protection limit.

Case B: We have the event $\underrightarrow{t+1}$ even though there is still excess type-$(k+1)$ demand to be upgraded. In this case, the protection limit *has* been reached. Here we can also make a somewhat surprising conclusion: there were *no* upgrades in period t. This can be shown by contradiction. Suppose that there were upgrades in period t. Then there was one type-$(k+1)$ customer who hit the protection limit during the period and was not upgraded. But if we add an extra unit of type-k product, then this unit will be used to upgrade that customer, and we have $\underrightarrow{k+1}$, instead of the assumed event, $\underrightarrow{t+1}$. Also, in this case, $\Delta_k^t(\mathbf{X}^t) \geq \alpha_{k+1,k}$ because the protection limit has been reached.

The same reasoning can be applied when we have residual capacity $\hat{\mathbf{N}}^t$ and event $\underrightarrow{\widehat{t+1}}$: only **Case A** and **Case B** are possible.

Now we are ready to list all possible sample paths and examine, for each path, the marginal value of an extra unit of product k given inventories $\mathbf{X}^t$and $\widehat{\mathbf{X}}^t$. We begin by looking at a relatively simple case in which our subproblem 'splits' because we run out of capacity for a high-level product:

(1) $\hat{N}_i^t \leq 0$ for some $j \leq i \leq k-1$, so that the demand for some product in the chain between j and $k-1$ is greater than the corresponding capacity $\widehat{\mathbf{X}}^t$ (thus also $\mathbf{X}^t$). Then, the allocation

problem separates in period $t+1$ and the one extra unit of product j in $\widehat{\mathbf{X}}^t$ has no impact on the marginal value of product k. Therefore, $\Delta_k^t(\widehat{\mathbf{X}}^t) = \Delta_k^t(\mathbf{X}^t)$.

(2) For the remaining scenarios we assume that $\hat{N}_i^t > 0$ for all $j \leq i \leq k-1$. We define subcases according to the value of $\hat{N}_k^t = N_k^t$, the amount of product k available after Step 1. We consider (2.1) $\hat{N}_k^t \geq 0$ and (2.2) $\hat{N}_k^t < 0$. Unfortunately, each of these cases will also have subcases, and subsubcases!

(2.1) $\hat{N}_k^t = N_k^t \geq 0$. Here there are two subcases, $N_{k+1}^t = 0$ and $N_{k+1}^t < 0$ (we cannot have $N_{k+1}^t > 0$, according to the definition of the subproblem).

(2.1.1) If $\hat{N}_{k+1}^t = N_{k+1}^t = 0$ then there will be no upgrading and $\hat{N}_k^t = N_k^t$ will be passed to period $t+1$. Therefore, by the induction assumption, we know $\Delta_k^t(\widehat{\mathbf{X}}^t) \leq \Delta_k^t(\mathbf{X}^t)$.

(2.1.2) If $\hat{N}_{k+1}^t = N_{k+1}^t < 0$, then the extra unit of product k *may* be used to upgrade demand for product $k+1$. This is the most complex case because the extra unit may be used differently, given $\mathbf{X}^t$ and $\widehat{\mathbf{X}}^t$ (recall that the protection limit may be lower under $\widehat{\mathbf{X}}^t$). Because $\hat{N}_k^t = N_k^t \geq 0$ there is no type-k demand remaining, so we cannot have $\underrightarrow{k}$ or $\widehat{\underrightarrow{k}}$. Therefore, we have four cases: $(\underrightarrow{k+1}, \widehat{\underrightarrow{k+1}})$, $(\underrightarrow{t+1}, \widehat{\underrightarrow{t+1}})$, $(\underrightarrow{k+1}, \widehat{\underrightarrow{t+1}})$, and $(\underrightarrow{t+1}, \widehat{\underrightarrow{k+1}})$.

(2.1.2.1) $(\underrightarrow{k+1}, \widehat{\underrightarrow{k+1}})$: In this case, $\Delta_k^t(\widehat{\mathbf{X}}^t) = \Delta_k^t(\mathbf{X}^t) = \alpha_{k+1,k}$.

(2.1.2.2) $(\underrightarrow{t+1}, \widehat{\underrightarrow{t+1}})$: From the Observation above, the same amount of product k is passed to period $t+1$ under $\mathbf{X}^t$ and $\widehat{\mathbf{X}}^t$. For **Case A**, all demand for product $k+1$ is upgraded, and the same quantity $N_k^t - d_{k+1}^t$ is passed to period $t+1$ under both $\mathbf{X}^t$ and $\widehat{\mathbf{X}}^t$. For **Case B**, there is no upgrading, so N_k^t is passed to period $t+1$ under both $\mathbf{X}^t$ and $\widehat{\mathbf{X}}^t$. Then by the induction assumption, we know $\Delta_k^t(\widehat{\mathbf{X}}^t) \leq \Delta_k^t(\mathbf{X}^t)$.

(2.1.2.3) $(\underrightarrow{k+1}, \widehat{\underrightarrow{t+1}})$: Here the additional unit in $\mathbf{X}^t$ is used for upgrading, for a marginal value of $\alpha_{k+1,k}$. Under $\widehat{\mathbf{X}}^t$, we are passing along the extra unit, and for **Case A** we know that $\Delta_k^t(\widehat{\mathbf{X}}^t) \leq \alpha_{k+1,k} = \Delta_k^t(\mathbf{X}^t)$. **Case B** implies an upgrade occurred under $\mathbf{X}^t$ while the same unit of capacity was protected under $\widehat{\mathbf{X}}^t$, implying a larger protection limit under $\widehat{\mathbf{X}}^t$. But the induction assumption indicates that protection limits are decreasing under $\widehat{\mathbf{X}}^t$. Therefore, **Case B** cannot occur.

(2.1.2.4) $(\underrightarrow{t+1}, \widehat{\underrightarrow{k+1}})$: Under $\mathbf{X}^t$ we again consider **Case A** and **Case B**. For **Case A**, we observed that all demand must have been upgraded and that there is more inventory than the

protection limit. However, we also know that under $\widehat{\mathbf{X}}^t$ the protection limit is the same, or smaller, than under $\mathbf{X}^t$ so both $\underrightarrow{t+1}$ and $\widehat{\underrightarrow{k+1}}$ cannot occur simultaneously, and **Case A** is impossible. Given **Case B**, under $\widehat{\mathbf{X}}^t$ the extra unit of product k is used for upgrading, with marginal value $\alpha_{k+1,k}$. Under $\mathbf{X}^t$ we know the marginal value of the additional unit is at least as high as $\alpha_{k+1,k}$ because the unit is passed to the next period even though there is an upgrading opportunity. Again, we have $\Delta_k^t(\widehat{\mathbf{X}}^t) \leq \Delta_k^t(\mathbf{X}^t)$.

(2.2) $\hat{N}_k^t = N_k^t < 0$. Because it is always optimal to complete parallel allocations (Step 1), this case implies events $\underrightarrow{k}$ and $\widehat{\underrightarrow{k}}$: we always assign an extra unit of product k to unmet k demand. However, to calculate the marginal value of this assignment, we have to consider whether this 'marginal' customer had already been satisfied by an upgrade to capacity $k-1$. Therefore, we consider four cases:

(2.2.1) For both $\mathbf{X}^t$ and $\widehat{\mathbf{X}}^t$, the additional unit of product k satisfies a type-k customer who otherwise would have been turned away. In this case, $\Delta_k^t(\widehat{\mathbf{X}}^t) = \Delta_k^t(\mathbf{X}^t)$.

(2.2.2) For both $\mathbf{X}^t$ and $\widehat{\mathbf{X}}^t$, the additional unit of product k satisfies a type-k customer who otherwise would have been upgraded to product $k-1$. In this case, $\Delta_k^t(\widehat{\mathbf{X}}^t) = \Delta_k^t(\mathbf{X}^t)$.

(2.2.3) Under $\mathbf{X}^t$ the customer would not have been upgraded (would have been turned away), but under $\widehat{\mathbf{X}}^t$ the additional unit of product k satisfies a type-k customer who otherwise would have been upgraded to product $k-1$. In this case, $\Delta_k^t(\widehat{\mathbf{X}}^t) = \alpha_{kk} - \alpha_{k,k-1} + \Delta_{k-1}^{t+1}(\widehat{\mathbf{X}}^t)$. Because the last unit of product $k-1$ had been used for upgrading, we know $\Delta_{k-1}^{t+1}(\widehat{\mathbf{X}}^t) \leq \alpha_{k,k-1}$. Therefore, $\Delta_k^t(\widehat{\mathbf{X}}^t) \leq \Delta_k^t(\mathbf{X}^t) = \alpha_{kk}$.

(2.2.4) Under the last scenario, the marginal customer would have been upgraded under $\mathbf{X}^t$ but not upgraded under $\widehat{\mathbf{X}}^t$. However, our induction assumption states that under $\widehat{\mathbf{X}}^t$ the protection limit is the same, or smaller, than under $\mathbf{X}^t$. Therefore, this scenario cannot occur.

We have shown that for all possible scenarios $\Delta_k^t(\widehat{\mathbf{X}}^t) \leq \Delta_k^t(\mathbf{X}^t)$ and, therefore, the protection limit is decreasing in the state vector. ■

Proposition 5 The optimal protection limit $\widetilde{p}^t$ is decreasing over time.

Proof: Consider two rationing problems with the same state vector $\mathbf{N} = (N_1, N_2, ..., N_{k+1})$. Let problem 1 be in period t_1, while problem 2 is in period t_2, and $t_1 < t_2$. Let $\widetilde{p}^1$ and $\widetilde{p}^2$ be the optimal protection limits for product k in the two problems, respectively. To prove $\widetilde{p}^1 \geq \widetilde{p}^2$ we

first show that the marginal value of product k, passed to the next period, is higher in problem 1 than that in problem 2. In particular, we show that this is true for any sample path between t_1 and t_2.

Suppose that an extra unit of product k is passed to t_1+1 in problem 1 and to t_2+1 in problem 2, and consider the demand arriving in problem 1 during periods t_1+1 to t_2. There are two possible cases. First, if no demand for any product is satisfied during those periods. In this case, problem 1 is equivalent to problem 2 at period t_2+1. Second, if a positive amount of demand is satisfied during those periods. Then at period t_2+1, the capacity vector of problem 1 is strictly smaller than that of problem 2. By the reasoning in the proof of Proposition 4, the marginal value of a unit of product k passed to the next period is higher for problem 1 than for problem 2. From the rationing optimality condition (6), $\widetilde{p}^1 \geq \widetilde{p}^2$. ■

Appendix C: 2x2 capacity optimization and Proposition 7

Given an arbitrary protection limit p, the objective function of the dynamic 2x2 model is

$$
\begin{aligned}
&\Pi^{DYN}(x_1,x_2) \\
= \underset{\mathbf{D}^1,\mathbf{D}^2}{\mathrm{E}} &\left\{
\begin{array}{l}
\alpha_{11}\min(d_1^1,x_1)+\alpha_{22}\min(d_2^1,x_2)+\alpha_{21}\min\left[(d_2^1-x_2)^+,(x_1-p-d_1^1)^+\right] \\
+\alpha_{11}\min\left[d_1^2,(x_1-d_1^1)^+-\min\left[(d_2^1-x_2)^+,(x_1-p-d_1^1)^+\right]\right] \\
+\alpha_{22}\min\left[d_2^2,(x_2-d_2^1)^+\right] \\
+\alpha_{21}\min\left[\begin{array}{c}\left\{d_2^2-(x_2-d_2^1)^+\right\}^+, \\ \left\{\begin{array}{c}(x_1-d_1^1)^+- \\ \min\left[(d_2^1-x_2)^+,(x_1-p-d_1^1)^+\right]-d_1^2\end{array}\right\}^+\end{array}\right] \\
-c_1x_1-c_2x_2
\end{array}
\right\}.
\end{aligned}
$$

The objective function of STC is

$$
\Pi^{STC}(x_1,x_2)=\underset{\mathbf{D}^1,\mathbf{D}^2}{\mathrm{E}}\left\{
\begin{array}{l}
\alpha_{11}\min(d_1^1+d_1^2,x_1)+\alpha_{22}\min(d_2^1+d_2^2,x_2) \\
+\alpha_{21}\min\left[(d_2^1+d_2^2-x_2)^+,(x_1-d_1^1-d_1^2)^+\right]-c_1x_1-c_2x_2
\end{array}
\right\}.
$$

For convenience let $x_T=x_1+x_2$, $x_{T-p}=x_1-p+x_2$,and $d_T=d_1^1+d_2^1+d_1^2+d_2^2$. Using techniques similar to those described by Netessine and Rudi (2002), we find the following partial derivative:

$$\begin{aligned}\frac{\partial \Pi^{DYN}}{\partial x_2} =& \alpha_{22} P(d_2^1 > x_2) - \alpha_{21} P(d_2^1 > x_2, d_1^1 + d_2^1 \leq x_{T-p}) \\ &+ \alpha_{11} P(d_2^1 > x_2, d_1^1 + d_2^1 \leq x_{T-p}, d_1^1 + d_2^1 + d_1^2 > x_T) \\ &+ \alpha_{22} P(d_2^1 \leq x_2, d_2^1 + d_2^2 > x_2) \\ &+ \alpha_{21} P(d_2^1 > x_2, d_1^1 + d_2^1 \leq x_{T-p}, d_1^1 + d_2^1 + d_1^2 \leq x_T, d_T > x_T) \\ &- \alpha_{21} P(d_2^1 \leq x_2, d_2^1 + d_2^2 > x_2, d_T \leq x_T) - c_2\end{aligned}$$

The third probability term with coefficient α_{11} merits special attention. This is the incremental profit when an additional unit of type-2 capacity leads to fewer upgrades in the first period, and thus more type-1 sales in the second period. This is the marginal benefit due to a reduction in the cannibalization of capacity.

The partial derivative for the single-period problem is

$$\frac{\partial \Pi^{STC}}{\partial x_2} = \alpha_{22} P(d_2^1 + d_2^2 > x_2) - \alpha_{21} P(d_2^1 + d_2^2 > x_2, d_T \leq x_T) - c_2.$$

Proposition 7 In the 2x2 case, $\partial \Pi^{DYN}(x_1, x_2)/\partial x_2 \geq \partial \Pi^{STC}(x_1, x_2)/\partial x_2$ for any capacities x_1 and x_2.

Proof: In the expression for $\partial \Pi^{DYN}/\partial x_2$ there are two probability terms multiplied by the constant α_{22}:

$$\alpha_{22} \left\{ P(d_2^1 > x_2) + P(d_2^1 \leq x_2, d_2^1 + d_2^2 > x_2) \right\} = \alpha_{22} P(d_2^1 + d_2^2 > x_2).$$

Therefore, the terms with the coefficient α_{22} are equal in $\partial \Pi^{DYN}/\partial x_2$ and $\partial \Pi^{STC}/\partial x_2$. In addition, both expressions include the term '$-c_2$'. Let Ψ denote the remaining terms in $\partial \Pi^{DYN}/\partial x_2$ (the second, third, fifth and sixth terms). We now show that Ψ is greater than or equal to the remaining term in $\partial \Pi^{STC}/\partial x_2$:

$$\begin{aligned}\Psi &\geq \alpha_{21} \left\{ \begin{array}{l} -P(d_2^1 + d_2^2 > x_2, d_T \leq x_T) \\ +P(d_2^1 > x_2, d_T \leq x_T) \\ -P(d_2^1 > x_2, d_1^1 + d_2^1 \leq x_{T-p}, d_1^1 + d_2^1 + d_1^2 \leq x_T, d_T \leq x_T) \end{array} \right\} \\ &\geq -\alpha_{21} P(d_2^1 + d_2^2 > x_2, d_T \leq x_T).\end{aligned}$$

where the first inequality follows by replacing α_{11} with α_{21} and rearranging the probability terms. This result applies for any protection level p, including the optimal protection level p^*. Therefore, $\partial\Pi^{DYN}(x_1, x_2)/\partial x_2 \geq \partial\Pi^{STC}(x_1, x_2)/\partial x_2$. ■

Acknowledgements

We are grateful to William Cooper, Marshall Freimer, Serguei Netessine, Dan Zhang, two anonymous referees, and an associate editor for their helpful suggestions. We also thank participants at graduate seminars at the University of Minnesota, the University of Toronto, and the University of California, Irvine, 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", *European Journal of Operational Research*, 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 Science*, 43(2): 173-183.

Bassok, Y., R. Anupindi, R. Akella. 1999. "Single-period multiproduct inventory models with substitution", *Operations Research*, 47(4): 632-642.

Bitran, G.R. and S.M. Gilbert. 1996. "Managing hotel reservations with uncertain arrivals", *Operations Research*, 44(1): 35-49.

Brumelle, S.L. and J.I. McGill. 1993. "Airline seat allocation with multiple nested fare classes", *Operations Research*, 41(1): 127-137.

de Véricourt F., F. Karaesmen and Y. Dallery. 2001."Assessing the Benefits of Rationing and Scheduling Policies for a Make-to-Stock Production System", *Manufacturing and Service Operation Management*, 2 (3): 105-121.

de Véricourt F., F. Karaesmen and Y. Dallery. 2002. "Optimal Stock Allocation for a Capacitated Supply System", *Management Science* 48 (11): 1486-1501.

Eynan, Amit. 1999. "The multi-location inventory centralization problem with first-come, first-served allocation", *European Journal of Operational Research*, 114: 38-49.

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

Frank, K.C., R. Zhang and I. Duenyas.

Ha, A.Y. 1997a. "Inventory rationing in a make-to-stock production system with several demand classes and lost sales". *Management Science*, 43: 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/Ek/1 make-to-stock queue", *Management Science*, 46(1): 77-87.

Hoffman, A.J. 1963. "On simple linear programming problems", V. Klee, ed., *Convexity: Proceedings of Symposia in Pure Math*, Vol. 7.

Jordan, W.C. and S.C. Graves. 1995. "Principles on the benefit of manufacturing process flexibility", *Management Science*, 41(4): 577-594.

Kapuscinski, R. and S. Tayur. 2000. "Dynamic capacity researvation in a make-to-stock environment", Working Paper, Carnegie Mellon University.

Karaesmen and van Ryzin. 2000. "Overbooking with substitutable inventory classes", Working Paper, Columbia University.

Karmarkar, U.S. 1981. "The multiperiod multilocation inventory problem", *Operations Research*, 29: 215-228.

McGill, J. and van Ryzin. 1999. "Revenue management: research overview and prospects", Transportation Science, Vol. 33, No. 2, pgs. 233-256.

Netessine, S., G. Dobson, R.A. Shumsky. 2002. "Flexible service capacity: optimal investment and the impact of demand correlation", *Operations Research*, Vol. 50, No. 2, pgs. 375-388.

Netessine, S., and N. Rudi. 2002. "Centralized and competitive inventory models with substitution", forthcoming in *Operations Research*, available at

http://omg.ssb.rochester.edu/omgHome/Rudi/page/.

Pasternack, B.A. and Z. Drezner. 1991. "Optimal inventory policies for substitutable commodities with stochastic demand", *Naval Research Logistics*, 38: 221-240.

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

Robinson, L.W. 1990. "Optimal and approximate policies in multiperiod, multiproduct inventory models with transshipment", *Operations Research*, 38(2): 278-295.

Robinson, L.W. 1995. "Optimal and approximate control policies for airline booking with sequential nonmonotonic fare classes", *Operations Research*, 43(2): 252-263.

Rockafellar, R. T. 1970. *Convex Analysis.* Princeton University Press, Princeton, New Jersey.

Savin, S.V., M.A. Cohen, N. Gans, and Z. Katalan. 2004. "Capacity management in rental businesses with heterogeneous customer bases", forthcoming in *Operations Research.*

Shumsky, R.A., and F. Zhang. 2003. "Dynamic capacity management with substitution: working paper", Simon School, University of Rochester.

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

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

Van Mieghem, J.A. 1998. "Investment strategies for flexible resources", *Management Science*, 44(8): 1071-1078.

Van Mieghem, J.A. and N. Rudi. 2002. "Newsvendor networks: dynamic inventories and capacities management with discretionary pooling", *Manufacturing & Service Operations Management*, 4(4): 313-335.

Van Mieghem, J. A. 2003. "Capacity Management, Investment, and Hedging: Review and Recent Developments", *Manufacturing & Service Operations.* Vol. 4, No. 4.

Van Slyke R. and R. Wets. 1966. "Programming under uncertainty and stochastic optimal control", *SIAM Journal on Control*, 4(1): 179-193.

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