Simple Policies for Managing Flexible Capacity

Ganesh Janakiraman[*] Mahesh Nagarajan[†] Senthil Veeraraghavan[‡]

April 2017

Forthcoming in M&SOM

Abstract

In many scenarios, a fixed capacity is shared flexibly between multiple products. To manage such multi-product systems, firms need to make two sets of decisions. The first one requires setting an inventory target for each product and the second decision requires dynamically allocating the scarce capacity among the products. It is not known how to make these decisions optimally. In this paper, we propose easily implementable policies that have both theoretical and practical appeal. We first suggest simple and intuitive allocation rules that determine how such scarce capacity is shared. Given such a rule, we calculate the optimal inventory target for each product. We demonstrate analytically that our policies are optimal under two asymptotic regimes represented by high service levels (i.e. high shortage costs) and heavy traffic (i.e. tight capacity). We also demonstrate that our policies outperform current known policies over a wide range of problem parameters. In particular, the cost savings from our policies become more significant as the capacity gets more restrictive.

Keywords: Flexible Capacity, Multiple Products, Allocation Rules, Asymptotic Optimality.

1 Introduction

In many industries such as auto-manufacturing, semiconductors, consumer electronics and pharmaceuticals, a firm's ability to carefully manage its flexible capacity is often a significant factor for its long-term success. The focus of this paper is to provide simple decision rules for managing flexibility efficiently - more specifically, rules for allocating limited capacity dynamically across several products.

[*] School of Management, University of Texas at Dallas, Dallas, TX. ganesh@utdallas.edu

[†] Sauder School of Business, University of British Columbia, Vancouver B.C, Canada V6T 1Z2, mahesh.nagarajan@sauder.ubc.ca

[‡] The Wharton School, University of Pennsylvania, 3730 Walnut St, Philadelphia, PA 19104.senthilv@wharton.upenn.edu

To achieve our goal, we study a firm that produces multiple products in every period, using a shared resource with limited capacity. We represent the firm's decisions using a periodically reviewed stochastic inventory model. Production occurs at the beginning of each period. A random demand (for each product) occurs during the period. For all products, the unsatisfied demand at the end of any period is backordered. Linear holding and shortage costs are assessed for all products at the end of every period.

We explore the objective of minimizing the long-run average cost per period. This optimization problem comprises of two sets of related decisions. The first one involves setting the target level for each product, and the second requires an allocation rule that determines how the *scarce* capacity is shared among the products. It is well known that performing these two tasks optimally is difficult (more details on this difficulty in the next section). Therefore, in this work, we propose implementable policies that have both theoretical and practical appeal.

Given the mathematical difficulty of analyzing our problem, we take an approach similar in spirit to papers that study limiting regimes of such stochastic control problems. We first suggest an intuitive *class* of allocation rules called *weighted balancing rules.* These rules are parametrized by a *weight* for each product, and they determine how the scarce capacity in any period is shared amongst multiple products. For every rule in this class, the optimal target level for each product is obtained directly from an application of the newsvendor formula – we refer to the combination of weighted balancing rules with these target levels as *weighted balancing policies.*

To provide theoretical validity to this class of policies, we study two different asymptotic regimes represented by (*i.*) service levels approaching one (i.e., when shortages are prohibitively expensive), and (*ii.*) utilization approaching one (i.e., when there is little slack between capacity and expected aggregate demand). For each of these two regimes, we identify a vector of weights within our policy class which is asymptotically optimal. To investigate how our class of policies performs, in general, we study a set of problems that span a wide range of costs, demand variabilities and capacity utilizations, which are not in the asymptotic regime.

1.1 Our Approach

We focus on a class of policies called *stationary base-stock policies*, which we define below. There is a target or base-stock level corresponding to each product. This target is stationary across periods.

At the beginning of a period, let us assume that the inventory level of each product will be at most equal to that product's base-stock level. (It is easy to see that our definition of a stationary base-stock policy is such that this assumption is satisfied in every period if it is satisfied in the first period). The difference between the inventory level and the base-stock level is called the "opening shortfall". If the aggregate shortfall of all products is smaller than the capacity limit, we produce enough of each product to raise its inventory to its base-stock level. The resulting shortfall ("ending shortfall") is thus zero for every product. If the aggregate shortfall exceeds the capacity, then the entire production capacity is used in such a way that the inventory level of each product does not exceed its base-stock level.

An *allocation rule* describes *how* the capacity is allocated to the different products in any period in which the capacity is insufficient (scarce) for all products to reach their base-stock levels. Thus, *even* within the class of stationary base stock policies, the optimal policy involves two interdependent sets of decisions, namely, base-stock levels (production policy) and an allocation rule. The lack of knowledge of the structure of the optimal allocation rule is thus the main stumbling block. To resolve this difficulty, we first fix the allocation policy, and find the optimal base-stock level under the fixed allocation policy.

The class of policies we advocate is the following. For any given vector of base-stock levels, we raise all inventory levels to the base-stock levels in periods in which the aggregate shortfall does not exceed the capacity. In periods in which the capacity is insufficient, all the capacity is used. Only in such periods, the allocation rule becomes relevant. An important aspect of the class of allocation rules proposed by us is that, in any period, the *only* information these decisions require is the opening shortfall of each product. In other words, for a given vector of opening shortfalls, the allocation decisions remain the same for any choice of base-stock levels. For any such allocation rule, the stationary distribution of the vector of ending shortfalls in a period is independent of the base stock levels. This finding has an important implication – the optimal base-stock level for each product can be computed using the newsvendor formula applied to the convolution of that product's demand and its ending shortfall.

Our approach is the following: We restrict attention to simple choices of allocation rules within the aforementioned class, and we choose the base-stock vector corresponding to any particular allocation rule optimally. We use a family of allocation rules which we refer to as *weighted balancing* *rules*. These rules work as follows. Each product is assigned a strictly positive weight which is constant through time. Next, at the beginning of each period, we rank order the products based on their weighted shortfalls (i.e. the shortfall divided by the weight). We then take the highest ranked product (i.e., the one with the largest weighted shortfall), and use the capacity to bring its weighted shortfall to be equal to the weighted shortfall of the second highest product. Next, we allocate capacity to both these products simultaneously until their weighted shortfalls coincide with the third highest product. We continue this procedure with subsequent products until the entire capacity is exhausted. As mentioned earlier, for any vector of weights, the base-stock level for each product is chosen optimally. This completes the description of a *weighted balancing policy*, given the vector of weights.

We now discuss the issue of choosing the weight vector. One special choice is that all weights are equal to 1 - we call the resulting allocation rule as the *symmetric rule*. At the other extreme are choices of the following type: There is some permutation $\{(1), (2), \ldots, (N)\}$ of the N products such that the weight for (1) << the weight for (2) << ... << the weight for (N) (here, we use << to mean "much smaller than"). Intuitively, such a choice mimics the *priority rule*, i.e. the rule which devotes all the available capacity to (1) until its shortfall is zero and then devotes all the remaining capacity to (2) until its shortfall is zero and so on. Later, we will prove that this priority rule, can be approximated by a suitable weighted balancing policy, for every beginning shortfall. We will show under certain assumptions that the symmetric rule is asymptotically optimal in high service level regimes while the priority rule is asymptotically optimal in heavy traffic. But, in general, the heuristic we propose searches over the space (more precisely, a grid) of weight vectors and picks the best vector. We will refer to the policy of using this weight vector along with the corresponding optimal base-stock levels as the *search policy*. Thus, for every problem instance, the search policy is at least as good as the two asymptotically optimal rules mentioned above; therefore, this policy also has the desired optimality property in both the asymptotic regimes.

We conclude this section by summarizing the benefits of the class of weighted balancing policies: (i) In the single product case, our policy (when there is only one product, there is only one policy in this class) is optimal. (ii) When all products are symmetric (i.e. they have identical costs and the distribution of the demand vector is exchangeable), we show (in §4) formally that our policy with symmetric weights is optimal. (iii) In high service level regimes, our policy with symmetric weights is asymptotically optimal. (iv) Finally, in heavy traffic (i.e. when utilization approaches one), the policy with weights chosen to mimic a priority policy is asymptotically optimal.

2 Related Literature

The single product capacitated inventory problem is a special case of our problem. The optimal policy for the single product problem is a modified base stock policy (Federgruen and Zipkin (1986)). However, not much is known about the problem with multiple products and limited capacity due to two sets of difficulties – computational and theoretical. From a computational perspective, the dynamic programming approach to solve this problem becomes intractable due to the curse of dimensionality. Providing simple and cost-effective heuristics which scale well to problems with many products is valuable - we will see that the policies we propose have these desirable attributes.

The theoretical difficulty is as follows. In the finite horizon dynamic program for the single product problem, the cost-to-go function is convex, and that guarantees the optimality of base-stock policies. The cost-to-go function can be shown to be convex even for the multi-product problem; however, this only guarantees the existence of a minimizer (interpreted as the vector of optimal *after-order* inventory levels) but it does not guarantee the optimality of *base-stock policies.*

Moreover, even temptingly simple and intuitive statements do not follow from convexity. For instance, one would imagine that the optimal policy possesses the following property which base-stock policies satisfy: If the inventory levels of all products at the beginning of a period are smaller than their optimal after-order inventory levels, then the inventory level of every product after ordering will be no larger than the optimal after-order inventory level. The careful reader will note that this property does not follow from convexity. In fact, a description of the optimal policy has so far been provided only for the two product case (that too, only for the finite/infinite horizon discounted cost problems, not the average cost problem) by Shaoxiang (2004) who expands on the early work by Evans (1967). For this case, Shaoxiang shows that the optimal policy is a base-stock policy. For the two product case, our weighted balancing rules can be viewed as linear approximations of the monotone switching curve in Shaoxiang (2004).

DeCroix and Arreola-Risa (1998) study the periodic review multi-product problem under both the finite horizon and the infinite horizon discounted cost criteria. They prove the optimality (for the finite horizon) of base-stock policies for the special case where all products are identical both in costs and demand distributions, and when the inventory level of each product in the first period is below its target level. For the general case, they provide a heuristic, but there are no theoretical results on the performance of those policies.

Aviv and Federgruen (2001) study a heuristic for a multi-product inventory system in which "blanks" are produced, and then allocated into multiple finished products. While their heuristic policy is optimal for the single product problem, it can be verified that it is not optimal even for the multi-product problem with symmetric products – this is because the base-stock levels used are obtained by solving a relaxed problem. As with DeCroix and Arreola-Risa's heuristic, there are no theoretical results on the performance of this policy.

While our focus is on periodically reviewed systems, there are counterparts in make-to-stock queues. Ha (1997) studies the two-product case and derives several structural properties of the optimal policy – results are similar to Shaoxiang's discrete time results. The other papers in the area (Zheng and Zipkin, 1990; Zipkin, 1995; Veatch and Wein, 1996; Rubio and Wein, 1996; Pena-Perez and Zipkin, 1997) study multi-product systems in the framework of multi-class make-to-stock queues – that is, the entire attention is on the class of base-stock policies and on finding good policies within this class. This body of work uses a combination of heavy traffic analysis and computational tests to motivate and evaluate various choices of base-stock levels and allocation rules. Most works in this literature stream assume *Poisson* demand processes. Among these papers, Pena-Perez and Zipkin (1997) and Veatch and Wein (1996) are closely related to our paper.

Veatch and Wein (1996) propose and evaluate *index rules*, which suggest any production should be devoted to the product with the lowest index at that time. Our weighted balancing rules are analogous to "linear" index rules. Pena-Perez and Zipkin (1997) argue that a specific *priority rule* is asymptotically optimal under certain assumptions for systems in "heavy traffic", i.e. systems where the aggregate demand rate is *close* to the production capacity rate. Their asymptotic analysis is based on the results of Wein (1992) and uses diffusion approximations. In this paper, we show a parallel result in periodic inventory models with two main strengths: (a) our notion of asymptotic optimality is *strong* (i.e. the difference between the cost of the priority policy and the optimal cost is bounded, while the optimal cost itself approaches infinity in heavy traffic) whereas their notion is *weak* (i.e. the ratio between the cost of the priority policy and the optimal cost approaches one) and (b) our proof is from first principles and does not rely on diffusion approximations.

3 Model Description

We index the products by n, $1 \leq n \leq N$. The holding and backorder cost associated with product n in \$/unit/period are h^n and b^n, respectively. Periods are indexed by $t \geq 1$. In period t, the net-inventory, x_t^n (inventory on hand minus backorders) for each product n is observed and the production quantity, q_t^n, for each product is decided. The total production quantity $q_t = \sum_{n=1}^N q_t^n$ is constrained from above by a capacity limit κ. Next, the demand, D_t^n for each product n is observed. Finally, the cost C_t incurred for this period is computed based on the inventory levels and backorder levels at the end of the period as follows:

$$C_t \;=\; \sum_{n=1}^{N} \left(h^n \cdot (x_t^n + q_t^n - D_t^n)^+ \;+\; b^n \cdot (D_t^n - x_t^n - q_t^n)^+ \right) \;.$$

The optimization problem that we are interested in solving is that of minimizing the long run average cost when the set of admissible (or feasible) policies is the set of all non-anticipatory policies. A formal definition of this problem follows. A non-anticipatory policy π is described by a set of vector-valued functions $\{\pi_t : t = 1, 2, \ldots, \}$ where $q_t^n = \pi_t^n(\mathbf{x}_t)$; here, $\mathbf{x}_t$ is the state vector $(x_t^1, \ldots, x_t^N)$ in period t and π_t is a function from $\mathbb{R}^N \to \mathbb{R}^{N,+}$. Let Π denote the set of all non-anticipatory policies π such that the capacity constraint $\sum_{n=1}^N \pi_t^n(\mathbf{x}) \leq \kappa$ for all $\mathbf{x} \in \mathbb{R}^N$ and for all $t \in \{1, 2, \ldots\}$ is satisfied. If C_t^π denotes the cost incurred by the system in period t when the system follows the policy π, our performance measure is

$$C^\pi = \limsup_{T \to \infty} E[\sum_{t=1}^{T} C_t^\pi]/T \;.$$

The optimal long run average cost is defined as $C^* = \inf_{\pi \in \Pi} C^\pi$.

Throughout the paper, we assume that the sequence of random vectors $\{\mathbf{D}_t\}$ is independent and identically distributed across time periods, where $\mathbf{D}_t = (D_t^1, \ldots, D_t^N)$. Note that we allow for the demands of the products to be correlated. We use D^n to denote a random variable with the same distribution as the single period demand for product n and D^{agg} to denote a random variable with the same distribution as the aggregate single period demand. Let $\mu^n = E[D^n]$. We also assume that capacity exceeds aggregate expected demand, i.e., $\mu^{agg} := \sum_{n=1}^{N} \mu^n < \kappa$, which is a necessary condition for the existence of a policy with a finite long-run average cost. Finally, the aggregate demand in a period can exceed capacity with positive probability, i.e., $P\left(D^{agg} > \kappa\right) > 0$. When the above condition does not hold, we can decompose our problem into a set of N newsvendor problems. We also assume that the expected demand for every product is strictly positive, that is, $E[D^n] > 0$ for every n. (If this condition is not met for some product n, then that product has zero demand with probability 1 and can therefore be disregarded.)

Let Π_{BS} denote the subset of stationary base-stock policies described at the beginning of Section 1.1. We now introduce some notation specific to Π_{BS}.

Let S^n denote the target or base-stock level for product n, and $\mathbf{S}$ denote the vector of base-stock levels. In our analysis of stationary base-stock policies, we assume that $x_1^n \leq S^n$ for all n. Let $W_t^n = S^n - x_t^n$; we refer to W_t^n as the opening shortfall for n in period t. Let V_t^n denote the ending shortfall, i.e. shortfall after ordering. So, $V_t^n = W_t^n - q_t^n$. By definition of a base-stock policy, the following condition holds:

$$\text{if } \sum_{n=1}^{N} W_t^n \leq \kappa \ , \quad \text{then } q_t^n \ = \ W_t^n \ \text{ for all } n \ .$$

That is, all inventory levels are raised to their respective targets, if that is feasible. Otherwise, the entire capacity is used for production without the inventory level of any product exceeding its target, i.e.,

$$\text{if } \sum_{n=1}^{N} W_t^n > \kappa \ , \quad \text{then } \sum_{n=1}^{N} q_t^n \ = \ \kappa \ \text{ and } q_t^n \leq W_t^n \text{ for all } n \ .$$

Notice that the exact manner in which the capacity is allocated among products in such periods has not been completely specified yet. We will specify these allocation rules shortly.

Let Π_{BS-B} denote the set of stationary base-stock policies in which the weighted balancing allocation rule is followed. We will refer to these as weighted balancing policies. A verbal description of these allocation rules was given in Section 1.1. Clearly, Π_{BS-B} is a subset of Π_{BS}. A mathematical description of a policy in this class follows.

Weighted Balancing Allocation: Rank the products according to the 'weighted' shortfalls $\{W_t^j/\alpha^j\}$, where α^j is the weight corresponding to product j. Let $\boldsymbol{\alpha} = (\alpha^1, \ldots, \alpha^N)$. The symmetric rule chooses $\boldsymbol{\alpha} = \mathbf{1}$, where $\mathbf{1} = (1, 1, \ldots, 1)$. Let $\tilde{n}$ denote the product with the n^{th} largest value of the weighted shortfall, W_t^j/α^j, breaking ties arbitrarily. Allocate production to product $\tilde{1}$ until its weighted shortfall equals that of $\tilde{2}$, or until the capacity is exhausted. Using the remaining capacity, allocate production to products $\tilde{1}$ and $\tilde{2}$ (proportionally based on their weights so that their weighted shortfalls are always equal) until their weighted shortfalls equal that of $\tilde{3}$, or until the capacity is exhausted. This process is continued until the entire capacity available in the period is exhausted. It is easy to see from this process that any product's shortfall at the end of this process is an increasing function of the shortfall vector at the beginning of the process; that is, for every n and t, V_t^n is an increasing function of $\mathbf{W}_t$. While the description above applies when inventory and production quantities are real-valued, a simple uniform randomization scheme (for breaking ties) can be used to define the policy when these quantities are integer-valued. Note that any policy $\pi \in \Pi_{BS-B}$ is completely specified by a pair $(\mathbf{S}, \boldsymbol{\alpha})$ where $\mathbf{S}$ is a vector of base-stock levels $\mathbf{S}$ and $\boldsymbol{\alpha}$ is a vector of weights.

Priority Allocation: Let $\{(1), (2), \ldots, (N)\}$ denote any permutation of $\{1, 2, \ldots, N\}$. Then, a priority rule defined by this permutation works as follows: In every period, allocate production to (1) until its shortfall is zero or the entire capacity is consumed; then, proceed to (2) and do the same until all product shortfalls are zero or the capacity is consumed. Note that the priority policy is lexicographic, and does not strictly belong to the class of weighted balancing rules. Nevertheless, it can be arbitrarily approximated by weighted balancing policies. We show this formally in Section 4.

4 Weighted Balancing Policies: Preliminaries

We start by explaining the connection between weighted balancing policies and the known structural properties of the optimal policy for the special case of two products. Shaoxiang (2004) shows for the infinite horizon, discounted cost version of this problem that an optimal policy satisfies the following: There exists a base-stock vector (S^1, S^2) such that once the inventory vector reaches a point which is componentwise smaller than (S^1, S^2), then the inventory vector in every subsequent period is also smaller than (S^1, S^2). Thus, the effective state space (i.e. possible inventory vectors) is $(-\infty, S^1] \times (-\infty, S^2]$; so, it is sufficient (for our purposes since we consider the average cost version of the problem) to study the optimal policy within this "rectangle". Within this region, the optimal policy is completely described by a monotone switching curve or function $\overline{x}^2(x^1)$.

Our weighted balancing policies work exactly like the optimal policy except that they replace $\overline{x}^2(x^1)$ with the function $(\frac{\alpha_2}{\alpha_1}) \cdot x^1$. In other words, computing the optimal policy involves finding or searching for the function $\overline{x}^2(x^1)$, i.e. the optimal "switching curve" within the space of all increasing functions, whereas, the best weighted balancing policy is found by searching for the best ratio α_2/α_1.

Consider any policy in Π_{BS-B} defined by a base-stock vector $\mathbf{S}$ and a weight vector $\boldsymbol{\alpha}$. Let $V_t^{agg} = \sum_{n=1}^N (S^n - x_t^n - q_t^n)^+$ denote the aggregate ending shortfall in period t. Let $D_t^{agg} = \sum_{n=1}^N D_t^n$ similarly denote the aggregate demand the system faces in period t. We now make a few observations about the vector of individual shortfalls of the products and the aggregate shortfall under any policy in Π_{BS-B}. All proofs are relegated to the appendix.

Lemma 1. *Consider any policy in Π_{BS-B} defined by a base-stock vector $\mathbf{S}$ and a weight vector $\boldsymbol{\alpha}$. Assume $x_1^n = S^n$ for all n. Then, (i) the distribution of the vector of ending shortfalls (its aggregate) is independent of $\mathbf{S}$ for all t, (ii) the sequence of distributions of this vector (its aggregate) converges to a limiting distribution as $t \to \infty$, (iii) these limiting distributions are also independent of $\mathbf{S}$, and (iv) the evolution of the aggregate shortfall process $\{V_t^{agg}\}$, is described by the recursive equation $V_{t+1}^{agg} = (V_t^{agg} + D_t^{agg} - \kappa)^+$.*

For any policy $\pi \in \Pi_{BS-B}$ defined by the pair $(\mathbf{S}, \boldsymbol{\alpha})$, we use $\mathbf{V}_t^{\boldsymbol{\alpha}}$ to denote the vector of shortfalls in period t. (Note that it is not necessary to include the argument $\mathbf{S}$ in the notation for the shortfall vector since its distribution does not depend on $\mathbf{S}$.) Let $\mathbf{V}_\infty^{\boldsymbol{\alpha}}$ denote the limiting distribution of $\mathbf{V}_t^{\boldsymbol{\alpha}}$; thus, $V_\infty^{\boldsymbol{\alpha},n}$ is the *steady-state* shortfall of product n. Let $\Phi^{\boldsymbol{\alpha},n}$ denote the distribution of the convolution of $V_\infty^{\boldsymbol{\alpha},n}$ and D^n. (Note: The steady state shortfall distributions can be computed by simulation. An interesting question for future research would be the identification of a more efficient and exact method for computing these distributions. Roundy and Muckstadt

(2000) present such a method for single-product, capacitated inventory systems.) Let

$$\mathbf{S}^{\boldsymbol{\alpha}*} = (S^{\boldsymbol{\alpha}*,1}, \ldots, S^{\boldsymbol{\alpha}*,N}), \text{ where } S^{\boldsymbol{\alpha}*,n} = (\Phi^{\boldsymbol{\alpha},n})^{-1}\left(\frac{b^n}{b^n + h^n}\right).$$

We will now show that the base-stock vector $\mathbf{S}^{\boldsymbol{\alpha}*}$ is the optimal choice of $\mathbf{S}$ within the subset of those policies in Π_{BS-B} that use the weight vector $\boldsymbol{\alpha}$.

Lemma 2. *Consider the class of weighted balancing policies, Π_{BS-B}. Within the subclass of policies which use the weight vector $\boldsymbol{\alpha}$, the policy with the base-stock vector $\mathbf{S}^{\boldsymbol{\alpha}*}$ is optimal.*

Next, we discuss the special case in which all products are "symmetric", i.e. the products are identical in terms of cost parameters and the distribution of the demand vector is exchangeable. We are able to make stronger statements about the optimal policy for this special case. We first formally state our assumption.

Assumption 1. *The following conditions hold. (a) $h^n = h$ and $b^n = b$ for all n. (b) The distribution of the vector $(D^1, \ldots, D^N)$ is exchangeable, that is, the joint distribution of $(D^1, \ldots, D^N)$ is identical to the joint distribution of $(D^{\theta(1)}, \ldots, D^{\theta(N)})$ for any permutation $(\theta(1), \ldots, \theta(N))$ of $(1, \ldots, N)$.*

Lemma 3. *Consider the policy in Π_{BS-B} defined by a base-stock vector $\mathbf{S}$ and the weight vector $\mathbf{1}$. Assume that $x_1^n = S^n$. Under Assumption 1 (b), the following statements hold.*
(i) The distribution of $\mathbf{V}_t^{\mathbf{1}}$ is exchangeable for all t.
(ii) The distribution of $\mathbf{V}_\infty^{\mathbf{1}}$, the limiting random vector mentioned in Lemma 1, is exchangeable.

Next, we show that the policy in Π_{BS-B} that uses the symmetric allocation rule and the corresponding optimal base-stock vector (as defined in Lemma 2) is optimal over all policies, not just base-stock policies, when all products are identical. This result is the average cost version of Theorem 3 of DeCroix and Arreola-Risa (1998)[1], which pertains to the finite horizon and infinite horizon discounted cost problems.

Theorem 4. *Consider the policy in Π_{BS-B} with the weight vector $\mathbf{1}$ and the base-stock vector $\mathbf{S}^{\mathbf{1}*}$. Under Assumption 1, this policy minimizes the long run average cost per period $\limsup_{T\to\infty} E[\sum_{t=1}^{T} C_t^\pi]/T$ over Π, the class of all non-anticipatory policies.*

In the next result, we show that shortfalls under the priority policy can be approximated by policies in Π_{BS-B}.

Lemma 5. *Let $((1),(2),\ldots,(N))$ denote any permutation of $\{1,2,\ldots,N\}$. Consider any given shortfall vector $\mathbf{W}$ (before ordering) in any period. Let $\boldsymbol{\alpha}_m$ be defined by $\alpha_m^{(1)} = 1$ and $\alpha_m^{(j)} = m \cdot \alpha_m^{(j-1)}$ for $j \in \{2,\ldots,N\}$. Let $\mathbf{V}^P$ and $\mathbf{V}^{\boldsymbol{\alpha}_m}$ denote the shortfall vectors after ordering under the priority rule (with priorities $(1) > (2) > \ldots > (N)$) and the weighted balancing rule (with weight vector $\boldsymbol{\alpha}_m$), respectively. Then, for every $\epsilon > 0$, there exists a sufficiently large M such that $|\mathbf{V}^{\boldsymbol{\alpha}_m} - \mathbf{V}^P| < \epsilon$ for all $m > M$, where $|(u_1,u_2,\ldots,u_n)| = \max\{u_1,u_2,\ldots,u_n\}$.*

5 High Service Level Asymptotics

We show under some technical assumptions that if the joint distribution of demands for all the products is exchangeable and the holding costs for all products are identical, then the best base-stock policy under the symmetric allocation rule is asymptotically optimal along a sequence of problems in which the backorder costs are scaled by a factor β that approaches ∞. In more practical terms, when the cost parameters are such that service levels for all products are high (in any reasonable policy), the best base-stock policy under the symmetric allocation rule is close to being optimal. We note that we do *not* restrict the backorder cost parameters for the products to be identical in this analysis. We proceed to state our assumptions formally, and then present our analysis.

Assumption 2. *The following conditions hold.*
(a) All products have identical holding costs, that is, $h^n = h$ for all $n \in \{1,\ldots,N\}$.
(b) The distribution of the vector $(D^1,\ldots,D^N)$ is exchangeable.
(c) Consider the steady state shortfall vector, $\mathbf{V}_\infty^{\mathbf{1}} = (V_\infty^{\mathbf{1},1}, V_\infty^{\mathbf{1},2},\ldots,V_\infty^{\mathbf{1},N})$, under the symmetric balancing policy. (Recall that this vector is also exchangeable under (b).) Let $W_\infty^{\mathbf{1},1}$ denote the convolution $V_\infty^{\mathbf{1},1} + D^1$. The limit

$$\lim_{x\to\infty} \frac{E[W_\infty^{\mathbf{1},1} - x \mid W_\infty^{\mathbf{1},1} > x]}{x} \;=\; 0\ .$$

The first two statements in the assumption above are self-explanatory. Statement (c) needs some discussion which we provide at the end of this section after we develop our asymptotic result under Assumption 2.

When the demand vector has an exchangeable distribution, let us employ $C^*(h,b)$ to denote the optimal long run average cost of a system in which *all* the products have the same holding cost parameter h and the same backorder cost parameter b. When the backorder costs are not identical (which might generally be the case), we use $C^*(h,\mathbf{b})$ to denote the same except that $\mathbf{b}$ represents the vector of backorder costs over all the products. We denote the long-run average cost of the policy in Π_{BS-B} with parameters $(\mathbf{S^{1*}},\mathbf{1})$ as $C^{\mathbf{1}*}(h,\mathbf{b})$. Finally, we denote the lowest backorder cost parameter in $\mathbf{b}$ by $\min(\mathbf{b})$ and the average of all the individual itemwise backorder costs by $\text{avg}(\mathbf{b})$.

In what follows, we note that $C^*(h,b)$ can be evaluated using Lemma 3. In our analysis, we use the cost $C^*(h,b)$ as a basis for cost comparisons across various policies because we know how it can be computed. Recall that the distribution of $\mathbf{V}^{\mathbf{1}}_{\infty}$ is exchangeable when the distribution of $(D^1,\ldots,D^N)$ is exchangeable. Then, we know that

$$C^*(h,b) \;=\; N\cdot L(h,b,V^{\mathbf{1},1}_{\infty}+D^1)\;, \tag{1}$$

where $L(h,b,X)$ is the optimal cost of a single product newsvendor problem with holding and penalty cost parameters h and b respectively, and facing a demand distribution of X, i.e.,

$$L(h,b,X) \;=\; \min_y \; h\cdot E[(y-X)^+] \;+\; b\cdot E[(X-y)^+]\;.$$

Before proceeding to the details of the analysis leading to the asymptotic optimality result of Theorem 7, we outline the main steps. In Lemma 6, we show that $C^*(h,\text{avg}(\mathbf{b}))$ and $C^*(h,\min(\mathbf{b}))$ are upper and lower bounds, respectively, on $C^{\mathbf{1}*}(h,\mathbf{b})$ - the long run average cost of the optimal symmetric policy. Notice that both the bounds are optimal costs of systems in which the products are symmetric in costs. (Recall that throughout this section we assume that the vector of product demands has an exchangeable distribution.) Thus, we can express these bounds as the optimal costs of certain newsvendor problems involving the convolution of demands and shortfalls as explained in the previous paragraph. Our goal is to show that the ratio $\frac{C^{\mathbf{1}*}(h,\beta\cdot\mathbf{b})}{C^*(h,\beta\cdot\mathbf{b})}$ approaches 1 as β approaches

∞. Thus, it is sufficient to show that the ratio of the optimal costs of the two newsvendor problems alluded to above converges to 1 since one of these optimal costs is an upper bound on the numerator of the ratio of interest and the other is a lower bound on its denominator. We establish this convergence in the proof of Theorem 7 by making use of a result in Huh et al. (2009) (presented in our appendix as Lemma 13) for the standard newsvendor problem under a mild distributional assumption on demand. Since the newsvendor problems of our interest involve the convolution of a product's demand and its shortfall, we have to demonstrate that this convolution also satisfies their assumption - statement (c) of Assumption 2 guarantees it.

Lemma 6. *Under Assumptions 2 (a)-(b), the following inequalities hold:*

$$C^*(h, \min(\mathbf{b})) \;\leq\; C^*(h, \mathbf{b}) \;\leq\; C^{\mathbf{1}*}(h, \mathbf{b}) \;\leq\; C^*(h, \operatorname{avg}(\mathbf{b})) \;. \tag{2}$$

We are now ready to derive an upper bound on the ratio $\frac{C^{\mathbf{1}*}(h,\mathbf{b})}{C^*(h,\mathbf{b})}$ and show that this ratio approaches 1 as $\mathbf{b}$ is scaled by a factor β which approaches ∞.

Theorem 7. *Under Assumptions 2 (a)-(b), the cost of using the symmetric allocation rule (and its corresponding optimal base-stock vector) relative to the optimal cost can be bounded as follows:*

$$\left(\frac{C^{\mathbf{1}*}(h, \mathbf{b})}{C^*(h, \mathbf{b})}\right) \;\leq\; \left(\frac{C^*(h, \operatorname{avg}(\mathbf{b}))}{C^*(h, \min(\mathbf{b}))}\right).$$

Moreover, if Assumption 2 (c) also holds, this ratio converges to 1 *as the backorder cost parameters grow, in the following sense:*

$$\lim_{\beta\to\infty}\left(\frac{C^{\mathbf{1}*}(h, \beta\mathbf{b})}{C^*(h, \beta\mathbf{b})}\right) \;=\; 1.$$

Discussion on Assumption 2 (c)

In this section, we argue that Assumption 2 (c) is closely related to the assumption that the aggregate demand in a period has a light-tailed distribution. A non-negative random variable X is said to have a light-tailed distribution if there exist non-negative constants A_1 and A_2 such that $P(X \geq x) \leq A_1 e^{-A_2 x}$ for all $x \geq 0$. The family of light-tailed distributions contains the popular family of IFR (increasing failure rate) distributions.

We also use the concept of *associated random variables* (Esary et al., 1967) defined as follows. Let $\mathbf{T}$ denote a vector of random variables $(T_1, T_2, \ldots, T_n)$, for some $n \in \mathbb{N}$. These random variables are said to be associated if $Cov(f(\mathbf{T}), g(\mathbf{T})) \geq 0$ for all non-decreasing functions f and g, such that $E[f(\mathbf{T})]$, $E[g(\mathbf{T})]$ and $E[f(\mathbf{T})g(\mathbf{T})]$ exist, where Cov denotes covariance. Esary et al. show that independent random variables are associated and that the set consisting of a single random variable is associated.

Assumption 3. *The distribution of the aggregate single-period demand, D^{agg}, is light-tailed. Moreover, the single period demands of the N products (that is, D^1, D^2, ..., D^N) are associated random variables.*

When $N = 1$, Assumption 3 reduces to the assumption that the single-period demand distribution is light tailed, which implies Assumption 2 (c), as follows: The steady state shortfall distribution is the same as the steady state waiting time distribution in a $GI/G/1$ queue. A famous result in Queuing Theory called the *Cramér* $-$ *Lundberg* approximation can then be used to show that the distribution of $W_\infty^{\mathbf{1},1}$ (recall that $N = 1$) has an asymptotically exponential tail, which immediately implies Assumption 2 (c). For details, see Glasserman (1997) and Huh et al. (2016).

We now discuss the connection between Assumption 3 and Assumption 2 (c) when $N > 1$. We first state a useful result.

Lemma 8. *Assumption 3 implies that the random variable $W_\infty^{\mathbf{1},1}$ is light-tailed.*

If $W_\infty^{\mathbf{1},1}$ is light-tailed, then, it is "typically true" (see Section 4(a) of Chapter VI in Asmussen and Glynn, 2007) that $\limsup_{x\to\infty} E[W_\infty^{\mathbf{1},1} - x \mid W_\infty^{\mathbf{1},1} > x] < \infty$, which implies Assumption 2 (c). In fact, it appears from the literature that, except in pathological cases, a light-tailed random variable X will have a bounded mean excess function (that is, $\limsup_{x\to\infty} E[X - x \mid X > x] < \infty$). The following assumption is one such "non-pathological" condition on $W_\infty^{\mathbf{1},1}$.

Assumption 4. *The random variable $W_\infty^{\mathbf{1},1}$ has a strictly positive density in $(0, \infty)$ and the limit of the reciprocal of the failure rate of $W_\infty^{\mathbf{1},1}$ exists; that is, $\lim_{x\to\infty} \frac{P(W_\infty^{\mathbf{1},1} \geq x)}{\frac{d}{dx}\left(P(W_\infty^{\mathbf{1},1} \leq x)\right)}$ exists.*

Lemma 9. *Assumption 2 (c) is satisfied if Assumptions 3 and 4 are satisfied.*

6 Heavy Traffic Asymptotics

In this section, we assume without loss of generality that the products are numbered in such a way that $h^1 \geq h^2 \geq \ldots \geq h^N$. We show that when $b^N = \min\{b^j\}$, the priority rule which assigns priorities based on the order $(1, 2, \ldots, N)$ is asymptotically optimal in heavy traffic, i.e. as the capacity κ approaches the expected aggregate demand μ^{agg}. Our proof is from first principles and does not use difficult approximations, unlike the Pena-Perez and Zipkin (1997) result for continuous time, mentioned earlier. Moreover, our asymptotic optimality result holds in the *strong* sense whereas Pena-Perez and Zipkin use it in the *weak* sense. We say that a policy π is asymptotically optimal in the weak sense along a sequence of systems indexed by n if the optimal cost approaches ∞ as n approaches ∞ and the ratio between the cost of π and the optimal cost approaches one. Furthermore, if the absolute difference between the cost of π and the optimal cost is bounded, we say that π is strongly asymptotically optimal.

To proceed with our asymptotic analysis, we first introduce some notation. Let $C^*(\mathbf{h}, \mathbf{b}, \kappa)$ be the optimal long run average cost of our inventory system when the holding cost vector is $\mathbf{h}$, the backorder cost vector is $\mathbf{b}$ and the capacity is $\kappa \in (\mu, \infty)$. Let $C^*(h, b, \kappa)$ be the same as $C^*(\mathbf{h}, \mathbf{b}, \kappa)$ when $\mathbf{h} = (h, h, \ldots, h)$ and $\mathbf{b} = (b, b, \ldots, b)$. Let $C^P(\mathbf{h}, \mathbf{b}, \kappa)$ denote the long run average cost of the priority policy, P, which assigns priority based on the order $(1, 2, \ldots, N)$ and uses the corresponding optimal base-stock levels. Following an argument identical to the proof of Lemma 2, it is easy to see that this optimal base-stock vector, say $\mathbf{S}^{P*}$, is given by

$$S^{P*,n} = \left(\Phi^{P,n}\right)^{-1}\left(\frac{b^n}{b^n + h^n}\right) \ \forall\ n,$$

where $\Phi^{P,n}$ denotes the distribution of the convolution of the steady-state shortfall $V_\infty^{P,n}$ and D^n. We present a preliminary lemma on the asymptotic behavior of the optimal cost $C^*(\mathbf{h}, \mathbf{b}, \kappa)$ using a well known result due to Kingman (1962) that a suitably scaled distribution of the waiting time in a single server queue converges to an exponential distribution in heavy traffic.

Lemma 10. *As the capacity κ approaches the expected aggregate demand μ, the optimal cost approaches ∞, i.e.* $\lim_{\kappa \downarrow \mu} C^*(\mathbf{h}, \mathbf{b}, \kappa) = \infty$.

Next, we present our assumption on the cost parameters formally before stating and proving our asymptotic result in Theorem 11.

Assumption 5. *The cost parameters satisfy the following conditions:* $h^1 \geq h^2 \geq \ldots \geq h^N$ *and* $b^N = \min\{b^j : 1 \leq j \leq N\}$.

Theorem 11. *Under Assumption 5, the following statement holds: There exists a finite constant* $\overline{M} < \infty$ *such that* $C^P(\mathbf{h}, \mathbf{b}, \kappa) - C^*(\mathbf{h}, \mathbf{b}, \kappa) \leq \overline{M}$ *for all* $\kappa > \mu^{agg}$ *. Therefore,* $\lim_{\kappa \downarrow \mu^{agg}} \frac{C^P(\mathbf{h},\mathbf{b},\kappa)}{C^*(\mathbf{h},\mathbf{b},\kappa)} = 1$.

7 Policy Performance and Results

Theorem 7 establishes that, as the backorder costs grow (or required service levels increase), the optimal cost under the symmetric allocation rule asymptotically approaches the optimal cost when the holding costs and demand distributions of all products are identical. While this result is of theoretical interest, it is also important to benchmark our policy.

Lower Bound for Benchmarking: Since the optimal cost is virtually impossible to calculate for a large set of problem instances due to the curse of dimensionality associated with dynamic programming, we require an easily computed lower bound on the optimal cost. Although we already have such a lower bound in Lemma 6 for the case of exchangeable demand distributions, we require a more generally applicable lower bound because other distributions are also included in our numerical investigation. We state such a lower bound in Lemma 12.

Let $G^n(x) = h^n \cdot E[(x - D^n)^+] + b^n \cdot E[(D^n - x)^+]$ be the expected single period newsvendor cost function for product n. We now develop a lower bound on the optimal long run average cost by using the free balancing relaxation (see, for example, Eppen and Schrage (1981) or Aviv and Federgruen (2001)). Let $F_1(y)$ be defined as follows: $F_1(y) = \min_{\mathbf{y}} \sum_{n=1}^{N} G^n(y^n)$ s.t. $\sum_{n=1}^{N} y^n = y$. We can now construct a lower bound on the optimal long run average cost using the function $F_1(\cdot)$. Let V_∞^{agg} denote the limiting random variable of the stochastic process representing aggregate ending shortfalls, i.e. $\{V_t^{agg}\}$. We employ V_∞^{agg} to derive a lower bound on the optimal cost.

Lemma 12. *Let* $LB_1 = \min_S E[F_1(S - V_\infty^{agg})]$. *Then,* LB_1 *is a lower bound on the optimal long run average cost over* Π*, the class of all non-anticipatory policies.*

7.1 Existing Heuristics

We now briefly describe the heuristics of DeCroix and Arreola-Risa (1998) and Aviv and Federgruen (2001), and compare their heuristics with our weighted balancing approach.

The heuristic of DeCroix and Arreola-Risa (1998) is a stationary base-stock policy. Let $(S^1, \ldots, S^N)$ denote the vector of base-stock levels for the N products. Let x^n be the net-inventory of product n at the beginning of the period. The symmetric resource allocation policy (SRAP) used in every period with insufficient capacity (for the inventory levels of all products to reach their base-stock levels) is to "balance" the ratios $\{x^n/S^n\}$. That is, allocate capacity to the product with the lowest ratio until that ratio equals the next highest ratio; from then, allocate capacity to these two products until their ratios equal the next highest ratio and so on, until the capacity is exhausted. Thus, our weighted balancing approach can be considered as a generalization of SRAP. Finally, the base-stock vector is chosen as follows. For every $n \in \{1, \ldots, N\}$, let z^n denote the newsvendor level for product n. That is, $z^n = \arg\min_y G^n(y)$. For products $n \in \{2, \ldots, N\}$, let $\gamma^n = z^n/z^1$. Let $f(S^1)$ denote the long run average cost of using the policy with the base-stock vector $(S^1, \gamma^2 S^1, \gamma^3 S^1, \ldots, \gamma^N S^1)$ and the allocation rule described above. The prescribed value of S^1 is that which minimizes $f(\cdot)$ and the prescribed value of S^n for any $n \neq 1$ is $S^1 \cdot \gamma^n$. Note that the evaluation of $f(S^1)$ for a given value of S^1 requires the computation of the steady state distribution of the shortfall vector. The computational effort for our weighted balancing approach is just the effort required to obtain this distribution. However, the heuristic above requires evaluating $f(S^1)$ over an entire search set for S^1, whereas we compute the steady state distribution of the shortfall vector *only once*.

The heuristic of Aviv and Federgruen (2001) is also a stationary base-stock policy. Let x^n be the net inventory of product n at the beginning of the period. In a period with insufficient capacity, the vector of inventory levels after ordering, $(y^1, \ldots, y^N)$, is chosen to solve the following myopic optimization problem: $\min_{\mathbf{y}} \sum_{n=1}^N G^n(y^n)$ s.t $y^n \geq x^n \ \forall \ n$ and $\sum_{n=1}^N (y^n - x^n) = \kappa$. The base-stock vector $(S^1, \ldots, S^N)$ is chosen as the solution to the optimization problem $\min \sum_{n=1}^N G^n(S^n)$ s.t. $\sum_{n=1}^N S^n = s$. The base stock levels add upto an "aggregate" basestock level solution to the following problem $s = \arg\min_S E[F_1(S - V_\infty)]$. Recall that $F_1(y) = \min_{\mathbf{y}} \sum_{n=1}^N G^n(y^n)$ s.t. $\sum_{n=1}^N y^n = y$. The AF heuristic requires the computation of the steady state distribution of the aggregate shortfall, in order to obtain the function F_1; thus, the AF method is comparable to our weighted balancing policies in terms of computational effort. However, the AF heuristic is *not* guaranteed to be optimal even in the symmetric case, whereas our policy is optimal in the symmetric case.

7.2 Policy under Weighted Balancing

Recall that we have established the optimality of the symmetric policy (a weighted balancing policy with weights of 1) and of the priority policy (a weighted balancing policy with extremely different weights) in the asymptotic regimes of high service levels and heavy traffic, respectively. Motivated by the fact that these two policies are very different in terms of their weight vectors, we propose searching over the space of weight vectors. While an exhaustive search for the weights would involve searching over the $N-1$ dimensional space of positive reals, we design a one dimensional search using a weight vector which is prescribed by m similar to Lemma 5, to find the best weighted balancing policy (i.e. the policy with the lowest cost). In our tables, we will refer to this policy as the "Search" policy or simply as *our* policy.

We conducted several computational experiments and compared the performance of our policy and the lower bound, with those of the heuristics of DeCroix and Arreola-Risa (1998) ("DA" in the tables) and Aviv and Federgruen (2001) ("AF" in the tables) and also, against the priority policy (represented as "Pri").

7.3 Computational Design

In the computational study, we first analyzed the three-product case, i.e. $N = 3$, systematically. We calibrated the performance of our policy, by starting with the symmetric case, and making cost and demand parameters gradually. In the addendum file titled "Computations: Simple Policies for Managing Flexible Capacity", available from the authors, we demonstrate how the performance of policy improves, as the demand distributions, holding costs and penalty costs each become asymmetric among the products. Our approach outperforms extant policies as the demand and product costs become more asymmetric. (For brevity, we say "asymmetric demands" to refer to the case when the demand distributions for all the products are not identical, and therefore, the distribution of the demand vector is not exchangeable). We then show that our approach to allocating capacity outperforms other extant approaches, and improvingly so, as capacity becomes scarce (See §7.4). We test our policy against the optimal policy (§7.5). Our approach is also applicable and performs well for correlated demands. We present only independent demands for the sake of brevity. Finally, we examine our policies on the same set of test instances considered in the previous literature, on a wider product portfolio ($N > 3$) in §7.6. In our tests, we use Erlang (k, λ) with appropriate k and λ, in order to match the first two moments (mean and variance) of any continuous demand distribution.

7.4 The Effect of Capacity

We now explore the effect of capacity κ on asymmetric backorder costs, holding costs, and demands. In Tables 1 and 2, we sequentially decrease the capacity such that the utilization increases from 73.3% ($\kappa = 60$) to 97.78% (for $\kappa = 45$) and demonstrate that our policy is very efficient in allocating scarce capacity among the products.

$k = 12, \lambda = (1.5, 1, 0.5)$ $\mathbf{b} = (15, 6, 3)$,$\mathbf{h} = (1.1, 1, 0.9)$	Costs					% gap			
Capacity	LB	Pri	DA	AF	Search	LB	Pri	DA	AF
60	78.7	88.7	86.3	83.8	83.6	-5.9%	6.1%	3.3%	0.3%
58	78.9	91.0	89.5	85.8	85.2	-7.4%	6.8%	5.0%	0.6%
56	79.4	94.1	94.4	88.6	87.7	-9.5%	7.4%	7.7%	1.1%
54	80.3	98.5	99.9	92.9	91.2	-12.0%	8.0%	9.6%	1.9%
52	82.1	104.7	108.4	98.8	96.4	-14.8%	8.6%	12.4%	2.4%
50	86.1	114.3	121.5	109.0	104.8	-17.8%	9.1%	15.9%	3.9%
48	95.2	130.1	144.9	125.4	118.9	-19.9%	9.4%	21.9%	5.5%
46	120.2	161.6	197.3	156.1	147.9	-18.7%	9.3%	33.4%	5.5%
45	150.0	194.0	262.5	185.7	179.3	-16.3%	8.2%	46.4%	3.5%

Table 1: Policy cost behavior for asymmetric demand and costs as the total capacity decreases.

In Table 1, product 1 has the lowest variability and product 3 has the highest variability. We reverse the ordering in Table 2. We note that as the capacity gets tighter, the relative difference between our policy and the lower bound increases. This is due to the weakened nature of the lower bound under high utilization. When capacity is unlimited, the multi-product problem decomposes into N individual newsvendor problems, and the lower bound coincides with the optimal cost. As the capacity gets tighter, the issue of allocating capacity is paramount and the lower bound benefits from the fact that it allows for costless redistribution of inventories in each period. In any case, the relative performance of our policy is strong when the capacity is limited.

$k = 12, \lambda = (0.5, 1, 1.5)$ $\mathbf{b} = (20, 10, 5)$, $\mathbf{h} = (1.2, 1, 0.8)$	Costs					% gap			
Capacity	LB	Pri	DA	AF	Search	LB	Pri	DA	AF
60	86.6	99.3	96.3	93.9	93.3	-7.2%	6.4%	3.2%	0.7%
58	86.9	102.4	100.1	96.7	95.6	-9.1%	7.1%	4.8%	1.2%
56	87.5	106.6	104.3	100.3	98.9	-11.5%	7.8%	5.5%	1.4%
54	88.8	112.4	110.4	105.9	103.5	-14.2%	8.6%	6.7%	2.2%
52	91.5	120.7	119.7	114.2	110.3	-17.0%	9.4%	8.5%	3.6%
50	97.2	133.4	133.5	127.3	120.8	-19.5%	10.4%	10.5%	5.4%
48	110.1	154.0	159.0	149.5	138.3	-20.4%	11.4%	14.9%	8.1%
46	144.5	195.5	216.7	190.6	175.5	-17.7%	11.4%	23.5%	8.6%
45	184.3	238.6	290.1	234.9	216.8	-15.0%	10.1%	33.8%	8.8%

Table 2: Cost behavior of our policy as the total capacity becomes scarcer. The asymmetric demands are *reversed* from the previous table.

Capacity	Priority			DA			AF			Search		
60	60	35	24	60	29	16	60	29	16	60	32	20
58	60	36	26	60	29	16	60	29	16	60	33	21
56	60	37	28	64	31	17	60	30	16	60	34	22
54	60	39	32	68	33	18	60	30	17	60	35	26
52	60	41	38	72	35	19	60	31	18	63	37	29
50	60	43	47	82	40	22	60	33	20	66	40	34
48	60	47	64	93	45	25	60	37	24	70	44	47
46	60	52	101	122	59	33	60	46	44	75	49	77
45	60	56	139	150	73	40	60	48	79	77	52	113

Table 3: Base-stock levels under different policies for instances in Table 2.

In Table 3, we show the base-stock levels for the scenarios reported in Table 2. In general, it appears that the priority policy assigns a significantly higher base-stock for the product 3. On the other hand, the DA heuristic chooses inventories such that a significantly higher base stock is assigned to Product 1. Our Search policy and the AF heuristic both choose base-stock levels that are in between those chosen under the Priority and the DA heuristics. It appears that the AF heuristic chooses weakly lower base-stocks for the products, compared to our policy - Note both AF and DA policies do not propose optimal basestock levels for asymmetric problems. These differences are more pronounced as the capacity becomes tighter. It also appears that our policy significantly outperforms other policies when the capacity is scarce, by setting up the base-stock parameters appropriately. The difference in the base-stock levels in our policy and those in the other heuristics may be possibly due to the better allocation approach used in our policy.

7.5 Comparison with the Optimal Policy

In this section, through Tables 4 through 7, we explore our performance against the optimal cost evaluated from the dynamic program (DP) over a set of computational experiments with limited state space, such that the DP solutions can be achieved within a few hours in each instance.[2] In all computational tables that follow, we fix the same capacity ($\kappa = 10$), and report the costs of all the heuristics, the lower bound and the optimal cost along with optimality gap for our policy.

$(k^1, k^2, k^3) = (3, 4, 3)$ $(\lambda^1, \lambda^2, \lambda^3) = (1, 1, 1.5)$, $\mathbf{h} = (1, 1, 1)$			Costs						% Optimality Gap
b^1	b^2	b^3	LB	DP	Pri	DA	AF	Search	Search-DP
2	2	2	5.90	7.08	8.71	7.43	7.42	7.32	3.37%
3	3	3	7.17	8.61	10.88	9.01	9.82	8.93	3.68%
4	4	4	8.06	9.76	12.53	10.20	10.65	10.14	3.89%
6	6	6	9.46	11.41	15.06	11.86	12.43	11.81	3.51%
10	10	10	11.27	13.56	18.48	14.57	15.84	14.07	3.80%
12	12	12	11.90	14.33	19.63	14.90	15.73	14.89	3.94%
15	15	15	12.65	15.26	21.09	15.98	17.69	15.85	3.86%

Table 4: Cost behavior of our policy and the lower bound against the optimal solution. The costs are symmetric and the demands are asymmetric. The symmetric penalty costs increase progressively down the column.

In Table 4, we study symmetric cost instances, and the demand distributions are asymmetric. Both our policy and the lower bound grow weaker as the backorder costs increase. However, we find that that our search policy is better than the other heuristics suggested in the literature, deviating only about $< 4\%$ from the optimal costs. In Table 5, we study the effect of varying demand asymmetry for fixed symmetric backorder costs.

$(k^1, k^2, k^3) = (3, 4, 3)$ $\mathbf{b} = (10, 10, 10), \mathbf{h} = (1, 1, 1)$			Costs						% Optimality Gap
λ^1	λ^2	λ^3	LB	DP	Pri	DA	AF	Search	Search-DP
1	1	1.1	11.38	20.35	35.85	28.91	32.05	28.91	42.08%
1	1	1.2	17.31	18.65	29.89	23.40	25.74	23.37	25.32%
1	1	1.3	13.46	16.21	23.80	18.11	19.78	18.11	11.72%
1	1	1.4	12.00	14.64	20.62	15.79	16.95	15.64	6.88%
1	1	1.5	11.27	13.56	18.48	14.57	15.84	14.07	3.80%

Table 5: Cost behavior of our policy and the lower bound against the optimal solution. Penalty costs are symmetric, and demands progressively asymmetric.

In Table 5, we varied the asymmetric demand, holding the variance of product 1 at the smallest value, and product 3 at the highest value. We notice in this case, both the lower bound and our policy perform *better* as the demand asymmetry increases. In particular, the performance of our policy improves from 42.08% to 3.8% as the demand becomes more asymmetric and as capacity tightens.

In Table 6, we repeat the same scheme in Table 5, except that the penalty costs are also made asymmetric. Thus backorder costs, mean demands, and standard deviations of the demand

$(k^1,k^2,k^3)=(3,4,3)$ $\mathbf{b}=(15,6,3),\mathbf{h}=(1,1,1)$			Costs						% Optimality Gap
λ^1	λ^2	λ^3	LB	DP	Pri	DA	AF	Search	Search-DP
1	1	1.1	15.74	15.88	23.74	26.74	21.74	21.70	36.68%
1	1	1.2	13.25	14.82	20.34	21.55	18.76	18.36	23.85%
1	1	1.3	11.14	13.30	16.81	16.60	15.20	15.04	13.10%
1	1	1.4	10.34	12.29	14.94	14.36	13.42	13.32	8.35%
1	1	1.5	9.95	11.63	13.73	12.87	12.89	12.31	5.81%

Table 6: Cost behavior of our policy and the lower bound against the optimal solution. Both holding costs and the demands are now asymmetric.

distributions are all asymmetric in this set of experiments. Our policy improves as the demands become more asymmetric, as observed from the last column (from 36.68% to 5.81%),

$(k^1,k^2,k^3)=(3,4,3)$ $\mathbf{b}=(6,3,1),\mathbf{h}=(1,1,1)$			Costs						% Optimality Gap
λ^1	λ^2	λ^3	LB	DP	Pri	DA	AF	Search	Search-DP
1	1	1.1	9.77	10.27	13.78	18.70	13.98	13.28	29.35%
1	1	1.2	8.60	9.72	12.08	15.06	12.30	11.61	19.52%
1	1	1.3	7.65	8.92	10.34	11.87	10.63	9.90	11.06%
1	1	1.4	7.24	8.40	9.42	10.27	9.88	9.03	7.49%
1	1	1.5	7.04	8.04	8.82	9.47	8.52	8.45	5.10%

Table 7: Cost behavior of our policy and the lower bound against the optimal solution. The penalty costs and the demands are asymmetric. The asymmetry of the demands is increased progressively.

In Table 7, we increased the ratio of the highest backorder cost to the lowest backorder cost. Our policy further improves under asymmetry, as can be seen by comparing the corresponding optimality gaps from Tables 6 and 7, showing (net) gains in performance gap ranging from 0.7% to 7%.

To summarize, the performance of our Search policy improves with respect to the optimal cost (i) as demands become more asymmetric, and (ii) as the holding and penalty costs become more asymmetric. In all tested cases, the performance of our policy is superior to the other policies suggested in the literature.

7.6 Comparisons for Larger Product Portfolio

We extensively tested our policy for the same multiple-product instances tested in Aviv and Federgruen (2001) and DeCroix and Arreola-Risa (1998).

In Table 8, we report our policy performance for all relevant test cases in Aviv and Federgruen (2001), using three sub-tables in the increasing order of demand variability. Within each sub-table, we progressively decreased the available capacity. The scenarios tested in Aviv and Federgruen

Utilization	Costs ($\mu = 40, \sigma = 5$)					Cost Gap		
of capacity	LB	Pri	DA	AF	Search	Pri	DA	AF
0.5	2.814	2.814	2.814	2.814	2.814	0.0%	0.0%	0.0%
0.66	2.814	2.814	2.814	2.814	2.814	0.0%	0.0%	0.0%
0.80	2.814	2.814	2.814	2.816	2.814	0.0%	0.0%	0.1%
0.889	2.814	2.814	2.814	4.686	2.814	0.0%	0.0%	66.4%
0.952	2.81469	3.27728	2.88981	47.3965	2.88963	13.4%	0.0%	1540.2%
0.976	2.8146	4.28293	3.26035	215.012	3.25688	31.5%	0.1%	6501.8%
0.990	2.8149	7.20741	5.47431	1275.34	5.47156	31.7%	0.1%	23208.5%
Utilization	Costs ($\mu = 40, \sigma = 10$)					Cost Gap		
0.5	5.88411	5.88411	5.88411	5.88411	5.88411	0.0%	0.0%	0.0%
0.66	5.88411	5.88414	5.88412	5.89053	5.88412	0.0%	0.0%	0.1%
0.80	5.88411	5.91856	5.88781	8.06056	5.88781	0.5%	0.0%	36.9%
0.889	5.88411	6.42585	5.95338	31.0372	5.953	7.9%	0.0%	421.4%
0.952	5.88418	8.81775	6.72287	218.26	6.72232	31.2%	0.0%	3146.8%
0.976	5.88828	12.9048	9.49216	860.201	9.49108	36.0%	0.0%	8963.3%
Utilization	Costs ($\mu = 40, \sigma = 20$)					Cost Gap		
0.50	10.1326	10.1326	10.1326	10.1326	10.1326	0.0%	0.0%	0.0%
0.66	10.1326	10.1459	10.1341	10.8818	10.1341	0.1%	0.0%	7.4%
0.80	10.1326	10.6219	10.183	22.9873	10.183	4.3%	0.0%	125.7%
0.889	10.1328	13.9616	10.8966	130.44	10.897	28.1%	0.0%	1097.0%
0.952	10.1369	18.4581	13.5375	408.124	13.5372	36.4%	0.0%	2914.8%
0.976	10.1587	25.153	18.8255	1001.33	18.8254	33.6%	0.0%	5219.0%
0.990	10.2182	34.777	27.5254	2150.47	27.5252	26.3%	0.0%	7712.7%

Table 8: Cost behavior of our policy against other heuristics for the multiple product cases ($N = 5$) in Aviv and Federgruen (2001). The cases tested in Aviv and Federgruen (2001) are all symmetric with holding costs at 0.05 and backorder costs at 1.0 for all products. Demands as shown above.

(2001) were all symmetric and comprise the first four rows in each sub-table. When utilization is low, the performances are comparable, as capacity constraints do not bind, and the allocation issues do not arise. However, as utilization increases, our policy outperforms other policies; the performance of the AF degrades quickly for high-utilization (i.e., tight capacity).

Demand and Cost Description	Utilization	Cost Gap over		
	of capacity	Pri	DA	AF
(I) Symmetric Demand and Symmetric Costs	83.3%	17.7%	0.3%	275.0%
$k = 2, \lambda = 1$	91.6%	36.4%	5.7%	1347.9%
(II) Symmetric Demand and Asymmetric Costs	83.3%	3.20%	4.45%	255.35%
$k = 2, \lambda = 1$	91.6%	4.67%	13.40%	1012.7%
(III) Asymmetric Demand and Asymmetric Costs	83.7%	6.90%	9.09%	417.09%
$k_i = (1, 2, 3, 4, 5)$	90.9%	2.23%	42.8%	1090%
$\lambda_i = (1.5, 1.6, 0.68, 11, 1.49)$	98.9%	1.57%	169.2%	>2000%

Table 9: Summary of our policy performance against other heuristics for the multiple product cases ($N = 5$) in DeCroix and Arreola-Risa (1998).

We also tested our policy for over more than 160 multiproduct cost scenarios described in DeCroix and Arreola-Risa (1998). We summarize the computational results from the 5-product cases in Table 9. We test symmetric cost and symmetric demand scenario in (I), and asymmetric costs in (II) and asymmetric costs and asymmetric demands in sub-table (III). Just as in DeCroix and Arreola-Risa (1998), the unit holding costs are such that $h^i = \eta * c^i$ for $\eta \in \{0.01, 0.05, 0.10, 0.15\}$, and the backorder costs at $b^i = \pi * c^i$ for $\pi \in \{0.5, 1, 1.5, 3\}$ where, $c^i = 2, 4, 8, 12$. Following their approach, for each sub-table in Table 9, we generated 20-24 cost scenarios from the above test set, and calculated the average costs of all policies over identical instances. We varied utilization by adjusting the capacity κ, and repeated the scenarios to re-calculate the average cost savings of our policy. Our policy performance is consistently superior to other policies (our policy costs were lower in *every* tested instance). We note that our policy performance is significantly better under the cases when flexibility is most valuable, – scenarios of multiple asymmetric products sharing tight capacity.

7.7 Discussion of Policy Performance

Based on the extensive computational experiments, we note that: (i) In virtually all of the problem instances we computed, our policy *significantly outperforms* all the extant heuristics. Our policy is optimal for symmetric cases. (ii) When the capacity utilization is low (i.e., excess capacity), our performances are comparable. As capacity decreases, our policy consistently outperforms other heuristics. (iii) Finally, when the number of products increase, or when the product characteristics are asymmetric, our policy significantly outperforms all other policies.

It is hard to fully characterize the optimal policy structure. However, we illustrate why our policy may perform well using a simple 3-product scenario with asymmetric products and tight capacity: $\kappa = 45$ with costs reported in the last row in Table 2, and the corresponding base stock levels in the last row of Table 3. Let the beginning inventory levels in some period be $(70, 70, 80)$ for the three products. Under the priority policy, the optimal base stock levels are $(60, 56, 139)$. Hence, we have to produce 59 units of product 3 and none for products 1 and 2. Due to limited capacity ($\kappa = 45$ units), shortfalls continue to exist (for product 3). Under the DA heuristic, the optimal base stock levels are $(150, 73, 40)$. Hence, we have to produce 80 units of product 1, 3 units of product 2 and none for product 3. Even in this case, shortfalls continue to exist, and surely for product 1, since capacity available is 45 but 80 units have to be produced. Under the AF heuristic, the optimal base stock levels are $(60, 48, 79)$. All items are above their base stock levels under the AF heuristic, so the entire capacity goes unused. Under our policy, the optimal base stock levels are $(77, 53, 113)$. 33 units of product 3 alone are produced. There is no shortfall. In this scenario, the DA and priority heuristics allow for too much shortfall for different products, and under the AF heuristic, the capacity may go unused (compared to the Search policy). Our policy searches over different weights and tries to achieve a balance between excessive shortfalls (due to high base-stock levels) and low capacity utilization (due to low base-stock levels).

The search policy picks the lowest cost policy by searching across weighted balancing policies (by choosing the best value of the parameter m). Observe that for appropriately chosen values of m, the corresponding weighted balancing policy is (a) optimal when all products are symmetric, (b) asymptotically optimal when the shortage cost to holding cost ratio approaches infinity and (c) asymptotically optimal as the capacity utilization (i.e., the ratio of the mean aggregate demand to the capacity) approaches 1. The values of m that achieve optimality (Theorem 4) or asymptotic optimality (Theorem 7) are quite different across these cases; for (a) and (b), the choice of $m = 1$ (i.e., all the weights being equal) achieves this whereas for (c) the choice of an extremely large value of m achieves this (i.e., the priority policy - see Theorem 11). Thus, the search policy can be loosely viewed as an appropriately chosen instance-specific convex combination of policies which are known to be optimal or asymptotically optimal for some special cases/asymptotic regimes of problem parameters.

8 Concluding Remarks

We have developed an intuitive, theoretically appealing and implementable policy for managing finite flexible capacity shared by multiple products. To implement our allocation policies, one just needs to examine their current shortfalls to determine the allocation of capacity amongst different products. In addition to being simple and intuitive to implement, our policies (a) have the theoretical appeal of being asymptotically optimal at high service levels and at high utilization levels, and (b) perform well when flexible capacity is most valuable (i.e., scarce capacity, varying demands and cost structures). Nevertheless, there are several challenging questions that are left unanswered. We note that set-up costs are absent in our model. Incorporating such costs and obtaining asymptotically optimal policies would be a worthwhile future research direction. Very little is known about the structure of the optimal policy when products are asymmetric. Even under our base-stock based policy, the structure of optimal allocation weight vector is not known. For the sake of computational ease, our heuristic calculates allocation vectors with a constant ratio between two successive products. A better understanding of the allocation structure would be beneficial in identifying some rules of thumb to employ flexible capacity better. We hope that our paper is a step towards achieving that objective.

Acknowledgements

The authors sincerely thank Prof. Steve Graves, an anonymous associate editor, and two anonymous reviewers for their detailed and constructive feedback on our theoretical results and proofs.

Notes

[1] We note that the conclusion of Theorem 3 of DeCroix and Arreola-Risa (1998) is not completely correct. For example, they claim the following: if, in some period, some products have inventory levels which exceed their optimal base-stock levels and if it is feasible to raise the inventory levels of the other products to their optimal base-stock levels, then the optimal policy is to not produce any of the products in the former category while bringing the other products' inventories to their optimal base-stock levels. This claim is incorrect because the optimal inventory level for a product after ordering depends non-trivially on the inventory levels of the other products since the cost-to-go function is *not* separable even though the single period cost function is separable with respect to the inventory levels of the products. However, we note that their claims are correct for every inventory vector in which every component is below its corresponding optimal base-stock level. The above comments also apply to Theorem 1 of their paper.

[2] We solve the finite horizon un-discounted dynamic program with a horizon length of 80 periods and take the minimum (over all starting states) average cost as a proxy for the optimal infinite horizon, average cost. The choice of the horizon length was based on our observation that our computed cost converges in the first two decimals. It can easily be shown that the cost we get is a lower bound on the optimal infinite horizon average cost. Thus the optimality gap that we report is numerically close to, but never smaller than the true optimality gap. In other words, our approach does not under-report the gap.

References

Asmussen, S. 2000. *Applied Probability and Queues 2/e.* Springer-Verlag.

Asmussen, S., and P. W. Glynn. 2007. *Stochastic simulation: algorithms and analysis*, Volume 57. Springer Science & Business Media.

Aviv, Y., and A. Federgruen. 2001. Capacitated Multi-Item Inventory Systems with Random and Seasonally Fluctuating Demands: Implications for Postponement Strategies. *Management Science* 47 (4): 512.

DeCroix, G., and A. Arreola-Risa. 1998. Optimal Production and Inventory Policy for Multiple Products under Resource Constraints. *Management Science* 44 (7): 950–961.

Eppen, G., and L. Schrage. 1981. Centralized Ordering Policies in a Multi-Warehouse System with Lead Times and Random Demand. *Multi-Level Production / Inventory Systems: Theory and Practice, ed. L. B. Schwarz*:51–67.

Esary, J., F. Proschan, and D. Walkup. 1967. Association of random variables, with applications. *The Annals of Mathematical Statistics* 38 (5): 1466–1474.

Evans, R. 1967. Inventory Control of a Multiproduct System with a Limited Production Resource. *Naval Research Logistics Quarterly* 14:173–184.

Federgruen, A., and P. Zipkin. 1986. An Inventory Model with Limited Production Capacity and Uncertain Demands I. The Average-Cost Criterion. *Mathematics of Operations Research* 11 (2): 193–207.

Glasserman, P. 1997. Bounds and Asymptotics for Planning Critical Safety Stocks. *Operations Research* 45 (2): 244–257.

Ha, A. 1997. Optimal dynamic scheduling policy for a make-to-stock production system. *Operations Research* 45 (1): 42–53.

Huh, W., G. Janakiraman, J. Muckstadt, and P. Rusmevichientong. 2009. Asymptotic Optimality of Order-up-to Policies in Lost Sales Inventory Systems. *Management Science* 55 (3): 404–420.

Huh, W., G. Janakiraman, and M. Nagarajan. 2016. Capacitated Multi-Echelon Inventory Systems: Policies and Bounds. *MSOM*:Forthcoming.

Huh, W. T., G. Janakiraman, and M. Nagarajan. 2011. Average Cost Inventory Models: An Analysis Using a Vanishing Discount Approach. *Operations Research* 59 (1): 143–155.

Kingman, J. 1962. On Queues in Heavy Traffic. *Journal of the Royal Statistical Society. Series B (Methodological)* 24 (2): 383–392.

Loynes, R. 1962. The stability of a queue with nonindependent inter-arrival and service. *Proc. Camb. Philos. Soc.* 58:497–520.

Pena-Perez, A., and P. Zipkin. 1997. Dynamic Scheduling Rules for a Multiproduct Make-To-Stock Queue. *Operations Research* 45 (6): 919–930.

Resnick, S. 1992. *Adventures in Stochastic Processes*. Birkhauser, Boston, MA.

Roundy, R. O., and J. A. Muckstadt. 2000. Heuristic Computation of Periodic-Review Base Stock Inventory Policies. *Management Science* 46 (1): 104–109.

Rubio, R., and L. Wein. 1996. Setting Base Stock Levels Using Product-form Queueing Networks. *Management Science* 42 (2): 259–268.

Schäl, M. 1993. Average Optimality in Dynamic Programming with General State Space. *Mathematics of Operations Research* 18 (1): 163–172.

Shaoxiang, C. 2004. The Optimality of Hedging Point Policies for Stochastic Two-Product Flexible Manufacturing Systems. *Operations Research* 52 (2): 312–322.

Veatch, M. H., and L. M. Wein. 1996. Scheduling a Make-To-Stock Queue: Index Policies and Hedging Points. *Operations Research* 44 (4): 634–647.

Wein, L. 1992. Dynamic Scheduling of a Multiclass Make-To-Stock Queue. *Opns. Res.* 40 (4): 724–735.

Zheng, Y., and P. Zipkin. 1990. A queueing model to analyze the value of centralized inventory information. *Oper. Res* 38 (2): 296–307.

Zipkin, P. 1995. Performance Analysis of a Multi-Item Production-Inventory System under Alternative Policies. *Management Science* 41 (4): 690–703.