Submitted to *Management Science*
manuscript MS-14-00384

Resource Pooling and Allocation Policies to Deliver Differentiated Service

Yuanguang Zhong
School of Business Administration, South China University of Technology, Guangzhou, China

Zhichao Zheng
Lee Kong Chian School of Business, Singapore Management University, Singapore

Mabel C. Chou
Department of Decision Sciences, NUS Business School, National University of Singapore, Singapore

Chung-Piaw Teo
Department of Decision Sciences, NUS Business School, National University of Singapore, Singapore

Resource pooling strategies have been widely used in industry to match supply with demand. However, effective implementation of these strategies can be challenging. Firms need to integrate the heterogeneous service level requirements of different customers into the pooling model and allocate the resources (inventory or capacity) appropriately in the most effective manner. The traditional analysis of inventory pooling, for instance, considers the performance metric in a centralized system and does not address the associated issue of inventory allocation. Using Blackwell's Approachability Theorem, we derive a set of necessary and sufficient conditions to relate the fill rate requirement of each customer to the resources needed in the system. This provides a new approach to study the value of resource pooling in a system with differentiated service requirements. Furthermore, we show that with "allocation flexibility" the amount of safety stock needed in a system with independent and identically distributed demands does not grow with the number of customers but instead diminishes to zero and eventually becomes negative as the number of customers grows sufficiently large. This surprising result holds for all demand distributions with bounded first and second moments.

Key words: Blackwell's Approachability Theorem; Inventory Pooling; Service Levels; Fill Rates
History: This paper was first submitted on Mar 4, 2014 and has been with the authors for 2 years for 3 revisions.

1. Introduction

Sales in global retail e-commerce increased more than 20% worldwide in 2014–to almost $840 billion–and they are expected to hit close to $1 trillion by 2015, according to a recent Global Retail E-Commerce Index published by A.T. Kearney. At the same time, both brick-and-mortar and pure-play online retailers are competing to create omni-channel offerings that link online and physical shopping. Brick-and-mortar leaders (such as Walmart and Nordstrom) continue to expand

their online offerings, while online players (such as Amazon and Singapore's Zalora) are setting up physical retail channels.

However, the boom in online sales has led to a proliferation of digital distribution channels for the manufacturers and suppliers. P&G China, for instance, employs four distinct business models in the e-commerce market space: The Open Market Model supports sales through small distributors on platforms such as TMall or Taobao; the Pure Player Model supports big distributors of P&G products (such as JD.com, Yihaodian, Walmart, and Suning.com); the Bricks-and-Clicks Model and the Direct-to-Customer Model support many of the other players in the e-commerce sector in China. Many of these e-distributors do not have an extensive network of offline inventory to support their online sales. Therefore, they rely on their manufacturers and suppliers to maintain inventory in the supply network and use the technique of "drop shipping"[1] to deliver goods to customers. A key challenge for such suppliers and manufacturers (such as P&G) is to maintain enough inventory required by each and every one of such e-distributors. This has led to a surge in the size of inventory maintained at the warehouses of manufacturers. Instead of locking in a fixed amount of inventory for each distributor in its warehouse, can the manufacturer execute a type of virtual transshipment of inventory within its warehouse to support sales generated from different e-distributors? Intuitively, this approach would enable the manufacturer to achieve higher profitability with less stock by pooling the inventory for different e-distributors together. However, how much can the manufacturer save using this approach?

To make the discussion concrete, we consider a manufacturer supporting a number of e-distributors, each facing uncertain demand that is met by filling orders with the inventory available to each e-distributor in a virtual inventory[2] system (cf. Figure 1). Each e-distributor i fills orders by drawing from a pile of inventory, say Q_i, in the virtual inventory system to satisfy its own demand. This is followed by a shipment from the manufacturer to the customer using drop shipping.

We note that with a dedicated pile of inventory in the system, each e-distributor can be guaranteed a certain service level, as a function of the inventory level. However, given the unpredictable

[1] Drop shipping is a supply chain management technique in which the retailer does not keep goods in stock but instead transfers customer orders and shipment details to a product supplier (either a manufacturer, distributor, or wholesaler) who then ships the goods directly to the customer. A 2009 report from the American Society of Business and Behavioral Sciences estimated that 22–33% of Internet retailers had adopted drop shipping as their primary method of order fulfillment. Forrester reported in 2012 that 34% of the products sold on Amazon.com in 2011 came from drop shipping, which accounted for orders worth $14.2 billion dollars (see e-Commerce Drop Shipping Standards (e-DSS), http://www.e-dss.org/drop-shipping). The advantage of drop shipping is clear: Retailers can broaden their product offerings to consumers and grow revenues while avoiding inventory-carrying costs and warehouse expenses.

[2] A virtual inventory includes all of the products available to a company for its sales, whether on a retail floor, in a back room or at a warehouse. When a customer requests a specific product which is available in the virtual inventory, the company either retrieves the product or sends a request for the product to be shipped to the point of sales (say the retail store) or customer's house. Note that a virtual inventory lists not only all of the products currently in the possession of the company, but also those the company can order from other sources and in turn sell to the customer (Symes, 2011).

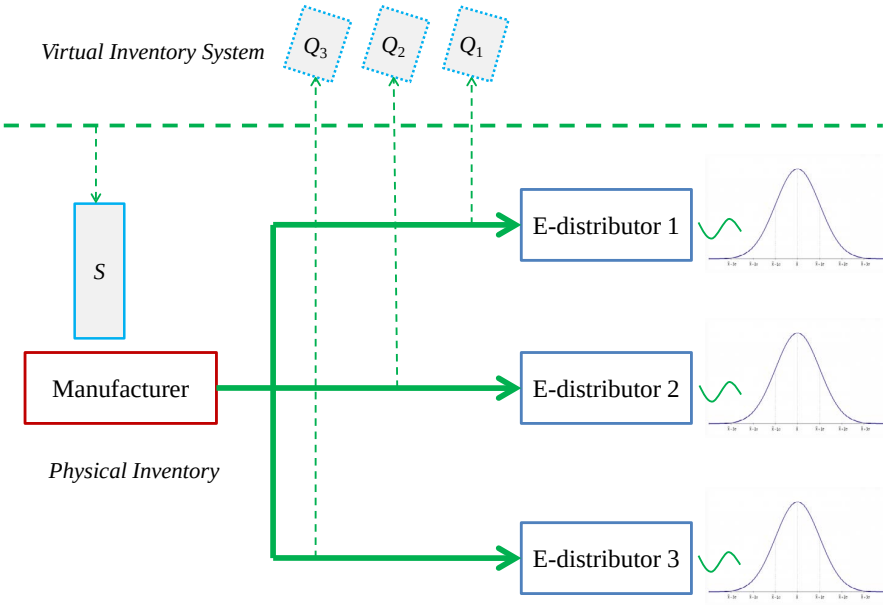


Figure 1 Supply Chain of the E-distributors and Manufacturer

nature of online sales, some of the inventory may not be utilized. Instead of maintaining an aggregate inventory (the sum of all the dedicated inventories) in its warehouse, the manufacturer can exploit its *allocation flexibility* to reduce the total inventory held for these e-distributors, while maintaining the service levels delivered to each e-distributor (as with dedicated virtual inventories). To this end, the manufacturer must address the following question:

> What is the minimum amount of inventory needed at the warehouse, such that it can be allocated to each e-distributor in a way that ensures each will get the required service level?

A common service level used in practice measures (and aims to increase) the *fill rate*, also called *type 2 service level*, obtained by subtracting the proportion of lost sales from the orders received. In other words, it measures the proportion of demand that can be fulfilled. When the manufacturer is able to pool the demand from the e-distributors together and allocate inventory appropriately, how would the safety stock level scale? Interestingly, we show in this paper that in a system with a sufficiently large group of e-distributors serving independent and identically distributed demands, as long as the fill rate requirement of each e-distributor is less than 1, the system does not need to stock more than the mean demand (i.e., the sum of the expected demand from all the e-distributors) to meet the fill rate obligations. This implies that there is actually no need to maintain any safety

stock[3] at all in the system! This adds further to the appeal of drop shipping for the suppliers in the booming e-commerce supply chain operations.

The rest of this paper is organized as follows. We review the relevant literature in the next section. In Section 3, we formally define the problem with a single pool of resources and develop our model and solution. Numerical analysis is presented to analyze the performance and robustness of our proposed policy. In Section 4, we analyze the safety stock phenomenon by exploring the structural properties of our model. In Section 5, we extend our analysis to incorporate issues with (i) 100% fill rate, (ii) minimal fill rate and (iii) finite-horizon fill rate requirements. Section 6 concludes the paper and discusses potential future research directions.

2. Literature Review

Inventory pooling has been studied in many settings. Here, we restrict our attention to papers that are closely related to ours. With regard to the newsvendor framework, researchers have mainly focused on studying the effect of inventory pooling under different settings. For example, Eppen (1979) was the first to demonstrate that the pooling effect (on cost) is always positive by aggregating all the individual demands that are assumed to be independent and identically distributed (i.i.d.) normal random variables. Van Mieghem and Rudi (2002) further analyzed the pooling effect in the newsvendor network setting. Corbett and Rajaram (2006) generalized the results found by Eppen (1979) to arbitrary demand distributions. Recently, Mak and Shen (2014) considered the risk pooling benefit in a general two-tiered supply chain structure, where both the demand and supply are stochastic. Other papers in this area have considered inventory pooling by integrating new factors. For instance, Özer (2003) studied a periodic-review distribution system with advance demand information and investigated the joint value of pooling demand and advance demand information in the system. Swinney (2012) introduced strategic customer behavior into inventory pooling and studied the impact of this feature on the benefit of inventory pooling.

The above traditional analysis derives a well-known "Square-Root Law": *The safety stock required for a pool of customers with i.i.d. demands grows in the order of the square root of the number of customers.* This rule is derived with a *centralized* system, when the demands of all the customers are pooled together. However, the traditional analysis is silent on how the centralized stock can be allocated to affect the service level experienced by each customer, and how individual service level requirement in turn affects the amount of safety stock needed. This is a core problem in the management of distribution systems with a central depot serving multiple locations/warehouses. In this paper, we explicitly address these gaps and associated issues.

[3] Safety stock is defined as the stock exceeding average demand that is maintained in the system in order to achieve a required service level.

Eppen and Schrage (1981) used the classical multi-echelon, multi-period inventory model to solve this problem, under the assumption that each warehouse has the same type 1 service level. Federgruen and Zipkin (1984) derived a dynamic programming model to analyze a similar problem, using a class of myopic allocation policies. They proved that a single-location inventory model could be used to approximate this dynamic problem. Erkip et al. (1990) further generalized the model of the depot-warehouse system to allow for demand correlation across warehouses and over time. They derived a closed-form expression for the optimal safety stock by assuming that each warehouse has the same coefficient of variation of demand. Benjaafar et al. (2008) studied a network system, with multi-inventory locations and multi-demand sources. For reviews of this line of research, please refer to Hopp et al. (1999), Caglar et al. (2004), and Özer and Xiong (2008). The focus of the above-mentioned models is mainly on cost minimization, as they ignore service differentiation concerns.

In this paper, we follow another stream of the literature, which classifies customers by their service level requirements. The literature in this area has traditionally focused on type 1 service level (stock out probability) requirements. Swaminathan and Srinivasan (1999) were the first to consider the problem of a single firm serving multiple customers with an individual type 1 service level requirement in a single period. Since the chance-constrained optimization model is hard to solve without specifying an allocation policy, they reformulated the model (through partitioning the space of the allocation outcomes) and proposed an algorithm to solve the problem. However, the computational time increases exponentially with the numbers of customers. Zhang (2003) studied the same problem as that in the paper by Swaminathan and Srinivasan (1999) and derived a closed-form expression for the optimal stocking level, under the assumption that at most one customer is not served. Recently, Alptekinoğlu et al. (2013) presented a systematic study of this problem. They identified the optimal policy to be of the priority type: Customers are served in accordance to a priority list. To be precise, an allocation policy belongs to the class of priority policies if it operates as follows: First, customers are ordered in a priority list; second, customer demands are filled from the available inventory in a decreasing order of priority; the sequential allocation process stops when all demands are filled or when the available inventory is exhausted, whichever occurs first. In this paper, we adopt their classification of priority-based allocation policies.

We obtain our inspiration for this study from a different field–wireless communication and networks. Wireless networks refer to computer networks that are connected using standard protocols but without cables of any kind. They have many applications, such as video streaming, online gaming, VoIP, and so on. In the area of wireless communication and networks, the theory of quality of service (QoS) has been well studied. The requirements include end-to-end deadline delay constraints, loss probabilities, and so forth. There are two core problems: One is deciding whether

the demands of a set of clients with given QoS constraints can be fulfilled; and the other is to find an optimal scheduling policy to meet the demands of all clients. There have been many studies on scheduling polices for wireless networks with QoS constraints. For instance, Hou et al. (2009) proposed an analytical framework for the problems; the authors used Blackwell's Approachability Theorem to prove the optimality of a proposed scheduling policy. Their fascinating paper paved the way for an important line of research on the applications of Blackwell's Approachability Theorem to scheduling problems. Hou and Kumar (2009) further extended the model in order to handle variable bit rate applications, including MPEG variable-bit-rate (VBR) video streaming, VoIP with differentiated quality, and wireless sensor networks (WSN). For more applications of their analytical framework, please refer to Hou and Kumar (2013).

Interestingly, the QoS requirements in wireless communication and networks are very similar to the service level requirements in inventory pooling. To the best of our knowledge, this paper is the first to exploit such connection in order to study the inventory pooling problem with fill rate requirements from customers.

There is another stream of literature, which attempts to integrate the service differentiation issue into the inventory rationing model by assuming that there are different priority classes among the demands. Topkis (1968) first derived the optimal rationing policy to satisfy the requirements of different customer classes over multiple time periods. The optimal policy is determined by a set of critical rationing levels such that at a given time demand of a given class is to be satisfied only if no demand of a higher priority class remains unsatisfied and the stock level does not fall below the critical rationing level for that class at that time. Ha (1997a, 1997b) analyzed the optimal rationing policy in a multi-class system with lost sales and backorders. He found that the optimal rationing policy is of a threshold type. Other related articles can be found in de Véricourt et al. (2001, 2002), Ding et al. (2006), and Yu et al. (2013). In several studies, researchers have tried to integrate service-level constrained models with traditional cost-minimization models by establishing some equivalence results (e.g., van Houtum and Zijm 2000, Boyaci and Gallego 2001, Axsäter 2003, and Zhang and Sobel 2012.) In these papers, the authors have either assumed that there is a single customer class or have studied backorder models, where unmet demands can be backlogged. However, in our problem, there are multiple customer classes with differentiated service level requirements and any unmet demand is assumed to be lost. Therefore, to the best of our knowledge, there is no previous study mapping our problem to any traditional cost minimization model.

In addition, there have been many studies on queuing control that discuss the scheduling policies for allocating demand among multiple servers. Many of these policies are shown to be asymptotically optimal (e.g., Gans and van Ryzin 1997, Harrison 1998, and Maglaras 2000). The objectives

in these models are mainly on minimizing the average system time, the delays that customers experience, or related costs etc. These queuing systems can be viewed as make-to-order systems in which no inventory is held in anticipation of future demand (Benjaafar et al. 2008).

3. Resource Pooling

In this section, we introduce our model and describe the main results. We first describe the single-period problem, and then map it to an equivalent problem using stochastic linear programming framework. We next develop a solution approach to the stochastic linear programming problem, and then translate it into a solution approach for the single-period problem. We end this section by presenting some numerical analysis on the benefit of pooling and allocation flexibility.

3.1. Model Formulation

We consider a system with one firm (manufacturer) and N customers (e-distributors), where the firm supplies a common item to all the customers from a centralized pool of resources in a single period. The resource can be viewed as inventory or various capacities in manufacturing or service systems. Every customer faces a nonnegative random demand X_i $(i=1,\ldots,N)$ and their demand distributions can be different and/or correlated. Throughout the paper, we use bold face letters to denote vectors; for example, the demand vector is denoted as $\boldsymbol{X} := (X_1,\ldots,X_N)$. At the beginning of a period, the firm needs to determine the resource capacity level in the common pool (denoted as S) before knowing the actual demand from the customers. Next, the demand realizes for each customer, who then orders from the firm. After learning the customers' demand, the firm allocates $D_i(\boldsymbol{X},S)$ units of resources to customer i, from its pre-configured capacity S. Let β_i denote the fill rate requirement of customer i, which measures the expected proportion of demand from customer i that is fulfilled immediately. The allocation rule $D_i(\boldsymbol{X},S)$ must satisfy the following inequality:

$$\mathbf{E}\left[D_i(\boldsymbol{X},S)\right] \geq \beta_i \mathbf{E}\left[X_i\right] \ \forall i=1,\ldots,N. \tag{1}$$

For a single-period problem that we consider in this section, the above inequality basically says: the expected proportion of demand from each and every customer i that is fulfilled immediately is at least β_i. While it cannot be measured directly, this imposes indirectly constraints on the class of allocation rules that can be used, for a given level of capacity S. In existing inventory management literature, Inequality (1) is also widely used to define expected fill rate (e.g., Chen et al. 2003, Choi et al. 2004, Thomas 2005, Katok et al. 2008, and Bensoussan et al. 2010).

More formally, our problem can be formulated as the following Problem (P):

$$\begin{aligned} (\mathrm{P}) \quad & \min_{S,\,\boldsymbol{D}(\boldsymbol{X},S)} \ S \\ & s.t. \quad \mathbf{E}\left[D_i\left(\boldsymbol{X},S\right)\right] \geq \beta_i \mathbf{E}\left[X_i\right], \forall i=1,\ldots,N \end{aligned}$$

$$D_i(\boldsymbol{X}, S) \le X_i,\ \forall i = 1, \ldots, N,\ \forall \boldsymbol{X} \in \Omega$$
$$\sum_{i=1}^{N} D_i(\boldsymbol{X}, S) \le S,\ \forall \boldsymbol{X} \in \Omega$$
$$S \ge 0$$

where Ω represents the set of all possible realizations for demand $\boldsymbol{X}$. The second and third constraints in Problem (P) require that, under any demand realization, the amount of resources allocated to customer i cannot exceed the demand quantity X_i and the total amount of resources allocated cannot exceed the capacity level S. Note that there are two sets of decision variables for this problem: capacity level S and allocation decisions $\boldsymbol{D}(\boldsymbol{X}, S) = (D_1(\boldsymbol{X}, S), \ldots, D_N(\boldsymbol{X}, S))$. This problem is challenging, in particular because there are infinitely many ways to characterize the allocation function $\boldsymbol{D}(\boldsymbol{X}, S)$, which in turn has to be jointly optimized with the capacity decision S. Next, we discuss some common heuristics and provide counterexamples to illustrate why they fail to solve the problem.

A standard heuristic to estimate the minimum capacity level is to translate the problem into a single centralized system that aggregates all the demands and service level requirements. Unfortunately, the following example shows that this approach fails if the service level requirements are different even when the demand distributions are i.i.d.

EXAMPLE 1. Suppose the demand from two customers, denoted by X_1 and X_2, are independent and take the values 50 or 150 with equal probability. The fill rate requirements are $\beta_1 = 0.9$ and $\beta_2 = 0.1$, respectively. We need to serve the customers, using a single pool of resources with capacity level S. One common approach is to pool the service requirements of the two customers together and then divide this pooled service requirement by the pooled demand, leading to a pooled fill rate requirement β_0 as follows:

$$\beta_0 := \frac{\beta_1 \mathbf{E}[X_1] + \beta_2 \mathbf{E}[X_2]}{\mathbf{E}[X_1] + \mathbf{E}[X_2]} = 0.5.$$

To ensure that $\mathbf{E}[\min\{S, X_1 + X_2\}] \ge \beta_0 \mathbf{E}[X_1 + X_2]$, the minimum level required is $S = 100$, since $\mathbf{E}[\min\{100, X_1 + X_2\}] = 100 = 0.5(\mathbf{E}[X_1] + \mathbf{E}[X_2])$. However, to achieve a fill rate of 0.9 for customer 1 we need $S = 130$, since $\mathbf{E}[\min\{130, X_1\}] = 0.5 \times 50 + 0.5 \times 130 = 90 = \beta_1 \mathbf{E}[X_1]$. Therefore, the right capacity profile for this problem is $S = 130$ but not $S = 100$ (as inferred from a pooled system). We therefore cannot solve the problem with differentiated service requirements simply by employing a system with a pooled service requirement derived this way. ■

Intuitively, the term $\beta_i \mathbf{E}[X_i]$ can be viewed as the minimum amount of resources we need to allocate for customer i. This can be seen by observing (1): In order to achieve fill rate β_i, we need

to have at least $\beta_i \mathbf{E}[X_i]$ units of resource for customer i. Therefore, we can view $\sum_{i=1}^{N} \beta_i \mathbf{E}[X_i]$ as a simple lower bound for S needed to achieve the required fill rate β_i for every customer i. The next example shows that this simple lower bound for S is insufficient even when there is only one customer in the system.

EXAMPLE 2. Suppose the firm has one customer with demand X, which takes values 10 or 20 with equal probability. The fill rate requirement is $\beta = 0.8$. The simple lower bound gives a capacity level of $S = 12$ $(= \beta \mathbf{E}[X])$. Using this amount of resources, the expected amount allocated to the customer is 11 $(= 0.5 \times 10 + 0.5 \times 12)$, and the firm can guarantee a fill rate of only 0.73 $(= 11/15)$, not 0.8. For this example, using reasoning similar to that in Example 1, it is possible to verify that the minimum resource capacity needed to deliver the required fill rate is 14, which is strictly above the lower bound of 12. The issue here arises from the situations when the demand is small, the allocated resources cannot exceed the demand quantity and thus any unused resources will be wasted. Therefore, the firm has to raise the capacity level above the lower bound in order to accommodate the situations with low resource utilization. ■

Note that another simple policy such as $D_i(\boldsymbol{X}, S) = \beta_i X_i$ does not work either, because the total resources available for allocation are bounded by S. In other words, none of the constraints in Problem (P) is redundant and all of them have to be considered when we design the allocation policies.

3.2. Priority Policies and the Stochastic Programming Approach

The standard approach to solve problem (P) is to translate this into a stochastic linear programming problem, by sampling demand scenarios $\{\boldsymbol{X}(t)\}_{t=1}^{T}$ and solving the following model:

$$
\begin{aligned}
\min_{S,\, \boldsymbol{D}(\boldsymbol{X}(t),S)} \quad & S \\
s.t. \quad & \frac{\sum_{t=1}^{T} \left[D_i \left(\boldsymbol{X}(t), S \right) \right]}{T} \geq \beta_i \frac{\sum_{t=1}^{T} \left[X_i(t) \right]}{T}, \forall i = 1, \ldots, N, \\
& D_i \left(\boldsymbol{X}(t), S \right) \leq X_i(t), \ \forall i = 1, \ldots, N, \forall\ t = 1, \ldots, T, \\
& \sum_{i=1}^{N} D_i \left(\boldsymbol{X}(t), S \right) \leq S, \forall\ t = 1, \ldots, T, \\
& S \geq 0.
\end{aligned}
$$

This approach focuses on finding a solution for S, but ignores the structure of the allocation policy $D_i(\cdot, \cdot)$, since the latter is usually computed with brute force from the mathematical programming model.

On the other hand, we can construct explicit allocation policy, say a priority rule, and then determine the minimal S needed to deliver the required service with this allocation policy. Note that, under any priority policy, at most one customer's demand will be partially served. The rest of the customers, depending on their priority in the list, either have their demands fully satisfied or not served at all. The priority rules can be further differentiated by whether or not they make use of actual demand information $\boldsymbol{X}$ when forming the priority list (cf. Alptekinoğlu et al. 2013).

- *Responsive priority policies*: The priority list is constructed using the realized demand information $\boldsymbol{X}$. For instance, a *greedy allocation policy* based on a smaller-demand-filled-first rule is responsive since the priority list is based on the actual demand realization.
- *Anticipative priority policies*: The priority list is constructed without using the demand realization information. These lists may be either deterministic or randomly generated.

The novelty of our approach is that we explicitly construct the allocation rule in the stochastic linear programming model. More specifically, for each demand sample, we construct a priority rule and serve the customers accordingly, allocating the available resources to meet the full demand from each customer in turn if possible. The priority rule we construct for the t^{th} sample depends on the demand realization and priority rules constructed for the previous $(t-1)$ samples, but not on the demand realized in the current sample. Hence our priority rules are anticipative and not responsive in nature. In this way, we can use the priority rules we constructed for the stochastic linear programming model to compose a randomized priority policy for the single period problem (P). The priority list is picked randomly from a set of priority lists generated by solving the above stochastic linear programming problem. We formally describe our analysis in the rest of this section.

We consider the case when the allocation problem is solved using an infinite number of sample scenarios, with each customer i facing a nonnegative random demand $X_i(t) \sim X_i$ in the t^{th} sample scenario. At every sampling epoch t, the firm has S units of resources to be allocated to the customers; and the allocation happens after the demand is realized for each customer. We develop an anticipative allocation policy $\boldsymbol{D}(t)$, for each t, with capacity level S, so that the long-run proportion of demand met for customer i is at least β_i, i.e.,

$$\liminf_{T\to\infty} \frac{\sum_{t=1}^{T} D_i(t)}{\sum_{t=1}^{T} X_i(t)} \geq \beta_i, \; a.s. \tag{2}$$

For notational convenience, we suppress the dependence on the history. Note that $D_i(t)$ is bounded above by $X_i(t)$. Furthermore, because the allocation mechanism can be history-dependent, the existence of $\lim_{T\to\infty} \left[\sum_{t=1}^{T} D_i(t) / \sum_{t=1}^{T} X_i(t)\right]$ is not assured. Therefore, we use the operator $\liminf$ and require β_i to be exceeded almost surely when we model the stochastic linear programming reformulation of Problem (P) using infinite number of samples:

$$(\mathrm{P}') \quad \min_{S,\, \boldsymbol{D}(t)} \; S$$

$$
\begin{aligned}
s.t.\quad & \liminf_{T\to\infty} \frac{\sum_{t=1}^{T} D_i(t)}{\sum_{t=1}^{T} X_i(t)} \geq \beta_i,\ a.s.,\ \forall i = 1,\ldots,N \\
& D_i(t) \leq X_i(t),\ \forall t,\ \forall i = 1,\ldots,N \\
& \sum_{i=1}^{N} D_i(t) \leq S,\ \forall t \\
& S \geq 0
\end{aligned}
$$

To solve Problem (P′), we derive a set of necessary and sufficient conditions involving only $\boldsymbol{X}$ and $\boldsymbol{\beta}$ for the capacity level S to be feasible when delivering the required service level performance. Therefore, we can eliminate the allocation decisions $\boldsymbol{D}(t)$ from Problem (P′) and solve it as a single-variable optimization problem. By exploring the structural properties of the reformulated problem, we are able to develop an efficient algorithm for Problem (P) and derive interesting managerial insights for the original problem, which are elaborated in detail in Section 4.

3.3. Necessary and Sufficient Conditions

In this subsection, we first present the main result before getting into the logic and intuition behind it. We relegate the proofs of all technical results in this paper to Appendix I.

THEOREM 1. *(1) Necessary conditions for the existence of a feasible allocation policy with capacity S for Problem (P′) are:*

$$\sum_{i\in U} \beta_i \mathbf{E}[X_i] \leq \mathbf{E}\left[\min\left\{S, \sum_{i\in U} X_i\right\}\right],\ \forall U \subseteq \{1,2,...,N\}. \tag{3}$$

(2) Moreover, the largest-debt-first policy[4] *is feasible when (3) holds, to deliver the fill rate of β_i to each customer i, for problem (P′)*

Based on the first part of Theorem 1, we can simplify Problem (P′) to the following optimization problem with a single decision variable, S:

$$
\begin{aligned}
(\mathrm{P}'')\ \min_{S}\ & S \\
s.t.\ & \sum_{i\in U} \beta_i \mathbf{E}[X_i] \leq \mathbf{E}\left[\min\left\{S, \sum_{i\in U} X_i\right\}\right],\ \forall U \subseteq \{1,2,...,N\} \\
& S \geq 0
\end{aligned}
$$

Let $\hat{S}$ denote the optimal solution to (P″). Note that the right-hand side of (3) is monotone in S. Therefore, the optimal capacity $\hat{S}$ can be found by bisection search, even for a large number of customers. In Section 3.5 we will describe in detail the algorithm for computing the optimal capacity level.

[4] The largest-debt-first policy is defined in section 3.3.2.

3.3.1. Necessary Conditions To see that (3) gives a set of necessary conditions, for any given subset U of customers at any sampling epoch t, the amount of resources allocated to U cannot exceed the total demand for all customers in U nor the total supply. That is, for any given U, $\sum_{i\in U} D_i(t) \le \sum_{i\in U} X_i(t)$ and $\sum_{i\in U} D_i(t) \le S$, which implies

$$\sum_{i\in U} D_i(t) \le \min\left\{S, \sum_{i\in U} X_i(t)\right\}.$$

Then (3) can be obtained by taking the average over an infinite number of sampling epochs of both sides of the above inequality and using the following lemma to complete the inequality.

LEMMA 1. *The fill rate of customer i is at least β_i in (P′) if and only if the average amount of resources allocated to customer i over an infinite number of sampling epochs is at least $\beta_i\mathbf{E}[X_i]$, i.e.,*

$$\liminf_{T\to\infty} \frac{\sum_{t=1}^{T} D_i(t)}{T} \ge \beta_i\mathbf{E}[X_i], \ a.s.$$

Note that the average amount of resources allocated to customer i over an infinite number of sampling epochs is actually the expected amount of resources allocated to customer i in a single period. Therefore, intuitively, an allocation policy meeting the customized service level requirement of each and every customer is feasible as long as the expected allocated resources to each customer i exceeds the average *implied workload*, $\beta_i\mathbf{E}[X_i]$. Lemma 1 allows us to translate the desired fill rates into the desired properties of an allocation policy. Observe that the amount of resources allocated to customer i increases as the required service level β_i and expected demand $\mathbf{E}[X_i]$ increase.

3.3.2. Sufficient Conditions To prove that (3) is a sufficient condition, we construct an anticipative allocation policy, called *largest-debt-first* policy (denoted by $\boldsymbol{A}^{DF}$), that requires only $\hat{S}$ units of resources to deliver the desired fill-rate of β_i to customer i.

Define $r_i(t+1)$ to be the *debt* of customer i at the beginning of sampling epoch $(t+1)$ as follows:

$$r_i(t+1) := t\beta_i\mathbf{E}[X_i] - \sum_{s=1}^{t} D_i(s). \tag{4}$$

The term $t\beta_i\mathbf{E}[X_i]$ represents the expected amount of resources needed in the first t sample scenarios to attain a fill rate of β_i from Lemma 1, and $\sum_{s=1}^{t} D_i(s)$ is the actual amount received by customer i for the first t samples. Therefore, the gap between these two terms can be viewed as the debt, based on the required fill rates, that the firm owes each customer. Define

$$\boldsymbol{R}(t) := \Big(\beta_1\mathbf{E}[X_1(t)] - D_1(t),\ \beta_2\mathbf{E}[X_2(t)] - D_2(t), \ldots, \beta_N\mathbf{E}[X_N(t)] - D_N(t)\Big),$$

as the *debt* accrued in the t^{th} sample. Note that $r_i(t+1) = \sum_{s=1}^{t} R_i(s)$. Furthermore, let $\boldsymbol{\rho}(t)$ denote the vector of the average debts for the first t samples, i.e.,

$$\boldsymbol{\rho}(t) := \frac{1}{t}\sum_{s=1}^{t} \boldsymbol{R}(s) = \frac{\boldsymbol{r}(t+1)}{t}.$$

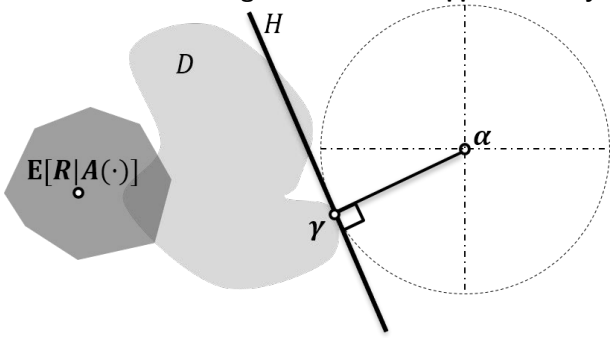
We use the *largest-debt-first* priority policy $\boldsymbol{A}^{DF}$ to allocate the resources: the priority list at epoch $(t+1)$ is formed according to the ordering of $\rho_i(t)$, from largest to smallest. Note that this is an anticipative policy since the demand information in epoch $(t+1)$ is not used to construct the priority list. Next, we show that, with capacity level $\hat{S}$, the above scheduling policy satisfies the service level requirements of all customers. This result relies on Blackwell's Approachability Theorem, which we briefly review next.

DEFINITION 1. A set $\mathcal{D} \subseteq \mathbb{R}^N$ is *approachable* by the sequence $\boldsymbol{\rho}(t)$ if and only if there is an allocation policy $\boldsymbol{A}(\cdot)$ such that the sample-average debt $\boldsymbol{\rho}(t)$ can be made to stay in or as close to $\mathcal{D}$ as we wish, for large enough t.

In 1956, Blackwell established the sufficient condition for approachability, as follows.

THEOREM 2. [Blackwell (1956)] *Suppose that $\mathcal{D}$ is any closed set, and for every $\boldsymbol{\alpha} \notin \mathcal{D}$ there is an action $\boldsymbol{A}(\cdot)$ such that the mean debt $\mathbf{E}[\boldsymbol{R}|\boldsymbol{A}(\cdot)]$ lies on the other side of the hyperplane H passing through $\boldsymbol{\gamma}$ (the point in $\mathcal{D}$ closest to $\boldsymbol{\alpha}$) and perpendicular to the line segment $\boldsymbol{\alpha\gamma}$ (see Figure 2). Then $\mathcal{D}$ is approachable with the strategy $\boldsymbol{A}(\cdot)$, where any arbitrary action can be taken when $\boldsymbol{\alpha} \in \mathcal{D}$.*

Figure 2 Schematic Drawing for Blackwell's Approachability Theorem



Intuitively, the classical approachability theorem can be explained as follows: If, whenever the current sample-average debt vector is outside the set $\mathcal{D}$ there is a policy to move the updated sample-average debt vector towards $\mathcal{D}$, then we can say that the time-average debt vector will

eventually get (arbitrarily close) to $\mathcal{D}$. Indeed, if the debts are bounded, by the Hoeffding-Azuma inequality together with the Borel-Cantelli lemma, the distance between $\boldsymbol{\rho}(t)$ and $\mathcal{D}$ converges to zero almost surely as t approaches infinity.

We exploit Blackwell's Theorem to establish the result in the second part of Theorem 1, that is, the necessary conditions for the resource pooling problem with fill rate constraints are sufficient, using our largest-debt-first policy as the allocation rule.

According to Lemma 1, to guarantee a fill rate of β_i for each customer i, the long-run average resource allocated to customer i must be at least $\beta_i \mathbf{E}[X_i]$. This immediately translates into the requirement on long-run average debt, which must lie in the nonpositive orthant

$$\mathcal{D} := \mathbb{R}_-^N = \{\boldsymbol{z} = [z_1, \ldots, z_N] : z_i \leq 0, \forall i = 1, \ldots, N\}.$$

In other words, we must show that the set $\mathcal{D}$ is approachable by the sequence $\boldsymbol{\rho}(t)$ with our allocation policy $\boldsymbol{A}^{DF}(\cdot)$ at capacity level $\hat{S}$, i.e.,

$$\liminf_{T\to\infty} \frac{r_i(T+1)}{T} \leq 0, \ \forall \ i = 1, \ldots, N. \tag{5}$$

Therefore, we can follow Blackwell's result in Theorem 2 to prove the approachability of $\mathcal{D}$. Specifically, we show that whenever the average debt vector $\boldsymbol{\rho}(t)$ falls outside the set $\mathcal{D}$, the largest-debt-first policy will give higher priority to customers with positive entries in $\boldsymbol{\rho}(t)$; and with an $\hat{S}$ that satisfies (3) we are able to guarantee that the expected debt in epoch $(t+1)$ moves towards $\mathcal{D}$, as required in Theorem 2.

More precisely, if the priority list at epoch t is ordered according to the sequence $([1], [2], \ldots, [N])$, then our priority-based allocation mechanism $\boldsymbol{A}^{DF}(\cdot)$ ensures that

$$\sum_{k=1}^{n} D_{[k]}(t) = \min\left\{\hat{S}, \sum_{k=1}^{n} X_{[k]}(t)\right\}, \ \forall \ n = 1, \ldots, N.$$

Since $\hat{S}$ satisfies $\sum_{k=1}^{n} \beta_{[k]} \mathbf{E}[X_{[k]}] \leq \mathbf{E}[\min\{\hat{S}, \sum_{k=1}^{n} X_{[k]}(t)\}]$, we have

$$\begin{aligned}\sum_{k=1}^{n} \mathbf{E}\left[R_{[k]}(t) | \boldsymbol{A}^{DF}(\cdot)\right] &= \sum_{k=1}^{n} \beta_{[k]} \mathbf{E}\left[X_{[k]}\right] - \mathbf{E}\left[\sum_{k=1}^{n} D_{[k]}(t)\right] \\ &\leq \mathbf{E}\left[\min\left\{\hat{S}, \sum_{k=1}^{n} X_{[k]}(t)\right\}\right] - \mathbf{E}\left[\sum_{k=1}^{n} D_{[k]}(t)\right] = 0, \forall n = 1, \ldots, N. \end{aligned} \tag{6}$$

This is a crucial property that we exploit in our analysis of the performance of the allocation mechanism.

3.4. Randomized Priority Policy

In this subsection, we show that the desired fill rate performance can be attained in our original single-period model as in Problem (P) by using a randomized priority policy constructed from the largest-debt-first policy in the stochastic linear programming model. The specific randomizing procedure to generate the priority list is given as follows.

ALGORITHM 1. Randomized Priority List for Problem (P)

1. We first simulate the demands $\boldsymbol{X}(t)$ over T samples, where T is a sufficiently large number.
2. In each sample t $(t=1,\ldots,T)$ we allocate the resources S according to the largest-debt-first policy $\boldsymbol{A}^{DF}(\cdot)$, as described earlier.
3. Let $L(t)$ denote the priority list in epoch t. Note that $L(t)$ is a random list that depends on the demand realized in the first $(t-1)$ samples, but is independent of $\boldsymbol{X}(t)$.
4. Finally, we randomly draw a priority list from $L(t), t=1,\ldots,T$, with equal probability.

Let L represent the priority list generated by the above procedure. We refer to this allocation rule as the *randomized largest-debt-first policy*, denoted by $\boldsymbol{A}^{L}(\cdot)$. It follows from the previous analysis that the randomized priority policy inherits the desired properties of the largest-debt-first policy.

THEOREM 3. *If the resource capacity level S satisfies the conditions in (3) and resources are allocated according to the randomized largest-debt-first policy $\boldsymbol{A}^{L}(\cdot)$ generated from Algorithm 1 (with sufficiently large T), then the expected fill rate for customer i is at least β_i as defined in (1).*

The proof of Theorem 3 follows easily from Theorem 1 and the randomized mechanism. Theorem 3 implies that the optimal resource capacity level for (P″) is also optimal for the single-period problem. In other words, the optimal capacity level for solving Problem (P) can also be obtained by solving Problem (P″).

3.5. Numerical Analysis

In this subsection, we perform numerical analysis to assess the value of resource pooling in our model. Solving Problem (P″) is still computationally prohibitive, due to the exponential number of constraints involved. However, we can exploit the existence of a simple allocation mechanism that can determine whether the fill rate targets are attainable with the existing resource capacity level. We summarize our algorithm in detail below.

ALGORITHM 2. Optimal Capacity Level for Solving Problem (P″)

1. Compute a lower and upper bound for the optimal resource capacity level. For instance, an upper bound for the optimal capacity can be constructed from Theorem 4, derived in Section 4.

2. Use the mid-point of the lower bound and upper bound as the initial capacity level and simulate the performance of the randomized largest-debt-first policy, using Algorithm 1 with a given T.

3. Perform a binary search for the minimum S between the lower and upper bounds obtained in Step 1; repeat Step 2 until all customers' service level requirements are satisfied.

We consider a system with three customers. We assume that the demand distributions among these customers are correlated and that τ is the common coefficient of correlation. We analyze three cases, where $\tau = -0.4$, $\tau = 0$, and $\tau = 0.4$. The optimal resource capacity levels are obtained after running the above algorithm, with $T = 1000$. To measure the numerical error of our solution algorithms, we define the approximation rate as follows:

$$AR = 1 - \frac{\sum_{i=1}^{N} \max\left\{\beta_i - \bar{\beta}_i, 0\right\}}{\sum_{i=1}^{N} \beta_i},$$

where $\bar{\beta}_i$ represents the simulated fill rate for customer i. Even for $T = 100$, the approximation rates in most sample paths already exceed 99.6%.

Table 1 shows the optimal resource capacity levels in the system (with and without pooling) and the associated pooling effect (defined as the percentage reduction in resource capacity level). In the table, "no pooling S" is the sum of the minimum capacity levels required to serve each retailer. Our results show that the pooling effects can be very significant as the demand variability increases for all three demand correlation cases–both when the variance of individual demand increases and when the distribution becomes more skewed (from normal to log-normal). Moreover, the pooling effect also increases when there is more differentiation in the service level requirements of all three demand correlation cases. The main reason for this is that customers with higher service level requirements need to be served with a higher capacity buffer, which can be partially used to cushion the demands from other customers with lower service levels. In addition, from Table 1, we observe that the pooling effect always decreases when demand becomes more positively correlated. Table 2 shows that our findings continue to hold even when the demand distributions are not identical.

4. Safety Stock Effects

Recall that safety stock is defined as the stock exceeding average demand that is maintained in the system in order to achieve a required service level. To see the implication of our results on safety stock, we first consider the case when demands are normally distributed. It is well known that (Jensen and Bard, 2002)

$$\mathbf{E}\left[\min\left\{S, \sum_{i=1}^{N} X_i\right\}\right] = \mathbf{E}\left[\sum_{i=1}^{N} X_i\right] - \mathbf{E}\left[\max\left\{S, \sum_{i=1}^{N} X_i\right\} - S\right]$$

Table 1 Pooling Effect in Type 2 Systems with 3 Customers, i.i.d. Demands, and Differentiated Service Level Requirements

					Optimal Capacity			Pooling Effect		
Demand	Required Service Level			No Pooling	$\tau=-0.4$	$\tau=0$	$\tau=0.4$	$\tau=-0.4$	$\tau=0$	$\tau=0.4$
Distribution	$\beta_1(\%)$	$\beta_2(\%)$	$\beta_3(\%)$	S	$\hat{S}$	$\hat{S}$	$\hat{S}$	(%)	(%)	(%)
$N(10, 2)$	80.0	80.0	80.0	24.60	24.00	24.06	24.24	2.44	2.20	1.46
	75.0	80.0	85.0	24.67	24.00	24.06	24.24	2.72	2.45	1.74
	70.0	80.0	90.0	24.88	24.00	24.06	24.24	3.54	3.28	2.57
	85.0	85.0	85.0	26.55	25.49	25.67	26.00	3.99	3.30	2.07
	82.5	85.0	87.5	26.59	25.49	25.67	26.00	4.14	3.42	2.22
	80.0	85.0	90.0	26.67	25.49	25.67	26.00	4.44	3.75	2.52
	90.0	90.0	90.0	28.88	27.01	27.47	28.00	6.48	4.88	3.05
	87.5	90.0	92.5	28.96	27.01	27.47	28.00	6.73	5.15	3.31
	85.0	90.0	95.0	29.16	27.01	27.47	28.00	7.37	5.79	3.98
	95.0	95.0	95.0	32.06	28.67	29.77	30.75	10.57	7.17	4.09
	92.5	95.0	97.5	32.35	28.67	29.77	30.75	11.38	7.98	4.95
$N(10, 3)$	80.0	80.0	80.0	25.93	24.00	24.36	24.95	7.44	5.97	3.78
	75.0	80.0	85.0	26.00	24.00	24.36	24.95	7.69	6.29	4.04
	70.0	80.0	90.0	26.35	24.00	24.36	24.95	8.92	7.53	5.31
	85.0	85.0	85.0	28.30	25.52	26.20	27.06	9.82	7.40	4.38
	82.5	85.0	87.5	28.34	25.52	26.20	27.06	9.95	7.56	4.52
	80.0	85.0	90.0	28.49	25.52	26.20	27.06	10.42	8.00	5.02
	90.0	90.0	90.0	31.27	27.13	28.35	29.55	13.24	9.29	5.50
	87.5	90.0	92.5	31.33	27.13	28.35	29.55	13.41	9.53	5.68
	85.0	90.0	95.0	31.72	27.13	28.35	29.55	14.47	10.60	6.84
	95.0	95.0	95.0	35.45	29.04	31.24	33.02	18.08	11.71	6.85
	92.5	95.0	97.5	35.83	29.04	31.24	33.02	18.95	12.74	7.84
$\mathrm{Log}N(10, 5)$	80.0	80.0	80.0	29.05	24.11	25.30	26.73	17.01	12.90	7.99
	75.0	80.0	85.0	29.32	24.11	25.30	26.73	17.77	13.71	8.83
	70.0	80.0	90.0	30.27	24.11	25.30	26.73	20.35	16.42	11.69
	85.0	85.0	85.0	32.98	25.82	27.73	29.71	21.71	15.92	9.92
	82.5	85.0	87.5	33.11	25.82	27.73	29.71	22.02	16.25	10.27
	80.0	85.0	90.0	33.52	25.82	27.73	29.71	22.97	17.27	11.37
	90.0	90.0	90.0	38.47	27.87	30.89	33.92	27.55	19.70	11.83
	87.5	90.0	92.5	38.75	27.87	30.89	33.92	28.08	20.29	12.46
	85.0	90.0	95.0	39.74	27.87	30.89	33.92	29.87	22.27	14.65
	95.0	95.0	95.0	47.81	30.88	35.77	40.50	35.41	25.14	15.29
	92.5	95.0	97.5	49.15	30.88	35.77	40.50	37.17	27.23	17.60
$\mathrm{Log}N(10, 10)$	80.0	80.0	80.0	44.53	26.08	30.56	35.32	41.43	31.38	20.68
	75.0	80.0	85.0	45.36	26.08	30.56	35.32	42.50	32.63	22.13
	70.0	80.0	90.0	48.52	26.08	30.56	35.32	46.25	37.01	27.21
	85.0	85.0	85.0	54.11	29.20	35.20	41.56	46.04	34.95	23.19
	82.5	85.0	87.5	54.52	29.20	35.20	41.56	46.44	35.44	23.77
	80.0	85.0	90.0	55.75	29.20	35.20	41.56	47.62	36.86	25.45
	90.0	90.0	90.0	68.98	33.71	42.05	50.57	51.13	39.05	26.69
	87.5	90.0	92.5	70.00	33.71	42.05	50.57	51.84	39.93	27.76
	85.0	90.0	95.0	73.67	33.71	42.05	50.57	54.24	42.92	31.36
	95.0	95.0	95.0	97.62	42.27	53.37	68.12	56.70	45.33	30.22
	92.5	95.0	97.5	103.62	42.27	53.37	68.12	59.27	48.50	34.27

$$= \mathbf{E}\left[\sum_{i=1}^{N} X_i\right] - \sqrt{\mathbf{Var}\left[\sum_{i=1}^{N} X_i\right]}\Big[\phi(k) + k\Phi(k)\Big],$$

Table 2 Pooling Effect in Type 2 Systems with 3 Customers, Non-Identical but Independent Demands, and Differentiated Service Level Requirements

Customer 1: $\mathcal{N}(10,2^2)$; Customer 2: $\mathcal{N}(10,3^2)$; Customer 3: Log-$\mathcal{N}(10,5^2)$									
				Optimal Capacity			Pooling Effect		
Required Service Level			No Pooling	$\tau=-0.4$	$\tau=0$	$\tau=0.4$	$\tau=-0.4$	$\tau=0$	$\tau=0.4$
$\beta_1(\%)$	$\beta_2(\%)$	$\beta_3(\%)$	S	S^*	S^*	S^*	(%)	(%)	(%)
80.0	80.0	80.0	26.08	24.01	24.41	25.01	7.95	6.38	4.12
75.0	80.0	85.0	26.81	24.01	24.41	25.01	10.42	8.95	6.61
70.0	80.0	90.0	24.88	24.01	24.41	25.01	14.59	13.08	10.84
85.0	85.0	85.0	28.72	25.57	26.35	27.25	10.97	8.30	5.09
82.5	85.0	87.5	29.17	25.57	26.35	27.25	12.40	9.85	6.46
80.0	85.0	90.0	29.88	25.57	26.35	27.25	14.39	11.72	8.82
90.0	90.0	90.0	32.07	27.31	28.74	30.05	14.85	10.40	6.33
87.5	90.0	92.5	32.99	27.31	28.74	30.05	17.22	12.92	8.88
85.0	90.0	95.0	34.43	27.31	28.74	30.05	20.56	16.35	12.64
95.0	95.0	95.0	37.30	29.71	32.27	34.54	20.29	13.49	7.43
92.5	95.0	97.5	39.98	29.71	32.27	34.54	25.51	18.61	13.75

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and distribution functions, respectively. $\mathbf{Var}[X]$ represents the variance of random variable X, and

$$k = \frac{\mathbf{E}\left[\sum_{i=1}^{N} X_i\right] - S}{\sqrt{\mathbf{Var}\left[\sum_{i=1}^{N} X_i\right]}}.$$

Let $G(x) := \phi(x) + x\Phi(x)$ denote the expected shortage function for a standard normal distribution. Then our system of necessary conditions in (3) can be rewritten as

$$G\left(\frac{\mathbf{E}\left[\sum_{i\in U} X_i\right] - S}{\sqrt{\mathbf{Var}\left[\sum_{i\in U} X_i\right]}}\right) \le \frac{\mathbf{E}\left[\sum_{i\in U}(1-\beta_i)X_i\right]}{\sqrt{\mathbf{Var}\left[\sum_{i\in U} X_i\right]}}, \forall U \subseteq \{1,2,\ldots,N\}.$$

Since $G(\cdot)$ is monotone and convex, we need

$$S \ge \max\left\{\mathbf{E}\left[\sum_{i\in U} X_i\right] - \sqrt{\mathbf{Var}\left[\sum_{i\in U} X_i\right]}\, G^{-1}\left(\frac{\mathbf{E}\left[\sum_{i\in U}(1-\beta_i)X_i\right]}{\sqrt{\mathbf{Var}\left[\sum_{i\in U} X_i\right]}}\right) : \forall U \subseteq \{1,2,\ldots,N\}\right\}. \quad (7)$$

Note that $G^{-1}(x) < 0$ only if $x < \sqrt{1/2\pi}$. Suppose the term $\mathbf{E}[\sum_{i\in U}(1-\beta_i)X_i]$ grows faster than $\sqrt{\mathbf{Var}[\sum_{i\in U} X_i]}$, as U grows larger; then we expect the term

$$G^{-1}\left(\frac{\mathbf{E}\left[\sum_{i\in U}(1-\beta_i)X_i\right]}{\sqrt{\mathbf{Var}\left[\sum_{i\in U} X_i\right]}}\right)$$

to be non-negative for large U. If one of these subsets attains the maximum value in (7), S does not exceed the mean demand of all the customers in the system, implying that there is no need for safety stock.

To see this in a more concrete manner, consider the case when the demands of the customers are independent and identically distributed. For a system with N customers–each with demand mean μ and variance σ^2, and a fill rate requirement of β–the optimal resource capacity level simplifies to

$$\hat{S} = \max\left\{ n\mu - \sigma\sqrt{n} \times G^{-1}\left(\frac{(1-\beta)\mu\sqrt{n}}{\sigma}\right) : n = 1, \ldots, N \right\}.$$

Figure 3 shows how the amount of safety stock needed ($\hat{S} - N\mu$) varies as N varies when

- $\mu = 5, \sigma = 1$, and
- $\beta = 0.99$.

Interestingly, from Figure 3, we observe that the amount of safety stock grows initially as N grows larger. However, once the size breaks the barrier of around 11, the pooling effect of allocation flexibility starts to dominate and the safety stock requirement starts to shrink, hitting zero at around $N = 64$. Note that in this case, the binding inequality in (3) is the one for $U = \{1, 2, \ldots, N\}$. The term $\mathbf{E}[\sum_{i=1}^{N}(1-\beta_i)X_i]$ grows in the order of N, but $\sqrt{\mathbf{Var}[\sum_{i=1}^{N} X_i]}$ grows in the order of $\sqrt{N}$. In the beginning, the "negative" effect of increasing total variance due to pooling–measured by the term $\sqrt{\mathbf{Var}[\sum_{i=1}^{N} X_i]}$–dominates, so the safety stock requirement increases. However, as the number of customers, N, increases, the benefit of allocation flexibility–measured by the term $\mathbf{E}[\sum_{i=1}^{N}(1-\beta_i)X_i]$–gradually takes over. Recall that $G^{-1}(x)$ is increasing, concave, and crosses zero at $x = \sqrt{1/2\pi}$. As a result, in a system with customers all facing i.i.d. and normally distributed demand, there is no need to maintain any safety stock for sufficiently large N!

Our results can be sharpened further. We also compare the optimal inventory level to the lower bound, $N\beta\mu$, and plot the gap in Figure 4. In this case, we can use the property of $G^{-1}(\cdot)$ to argue that this gap will converge to zero for sufficiently large N. In fact, for $N \geq 3500$, our result essentially indicates that we need only $N\beta\mu$ units of capacity to ensure that each customer will receive fill rate of β from the pooled system.

We next examine the situation when we want to satisfy in full the demand of the customers. Intuitively, with a higher fill rate requirement for each customer, the chances of the customer facing shortfall will be reduced. Figure 5 shows that, indeed, as the fill rate β gets closer to 1, the average number of customers suffering shortfalls reduces to zero. In fact, for $N = 100$ and $\beta = 0.99$, we can fulfill the fill rate requirement for each customer using only 498 units of inventory, slightly less than the average total demand. At the same time, on average, only 1.3 customers will not get her demand met in full.[5] If we choose $\beta = 0.999$, then we need around 512 units of inventory to

[5] This may sound counter-intuitive since there is more than 50% chance that the total demand will be more than 498. To understand this result, suppose we serve the customers using the greedy policy, i.e., customers with smaller demands will have higher priority over customers with higher demand. In this case, we note that the probability mass for aggregate demand to be in the intervals $(-\infty, 498], (498, 506], (506, 514], (514, 522], (522, 530]$ and $(530, 538]$

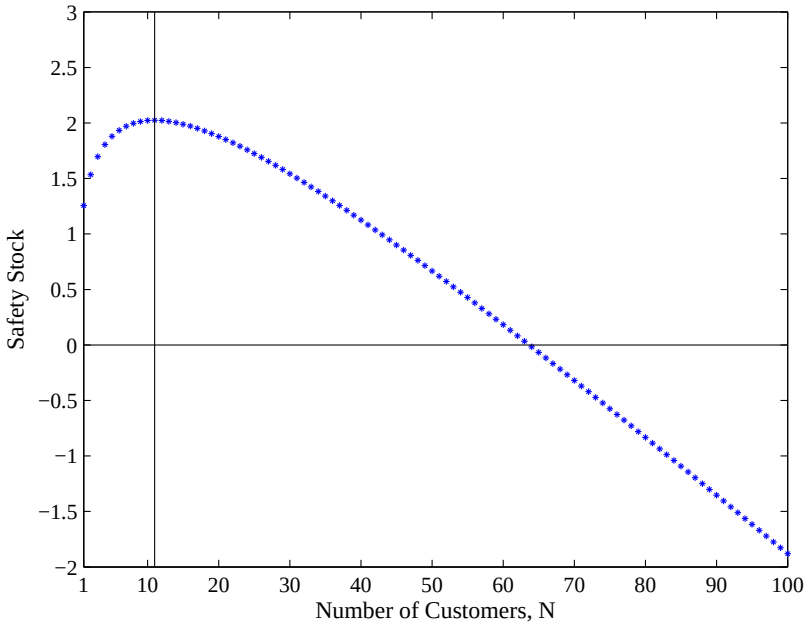


Figure 3 Safety Stock Requirement ($\hat{S} - N\mu$) as N varies

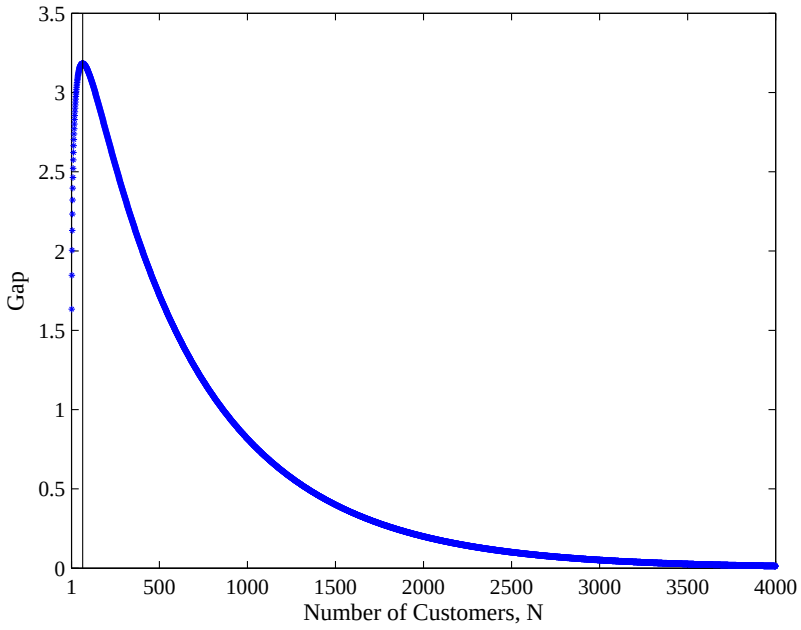


Figure 4 Gap Between the Optimal Inventory Level and Lower Bound ($\hat{S} - N\beta\mu$) as N varies

meet the fill rate obligations, but this time, on average, only less than 0.2 customers will not get

are $0.42, 0.305, 0.193, 0.067, 0.013$, and 0.0012 respectively. In the first interval, we can meet all demand in full. In the subsequent intervals, since the maximum demand from 100 customers is around 8, the number of customers with shortfall is roughly $1, 2, \ldots, 5$ in the intervals listed in that order. The average number of customers with shortfall is

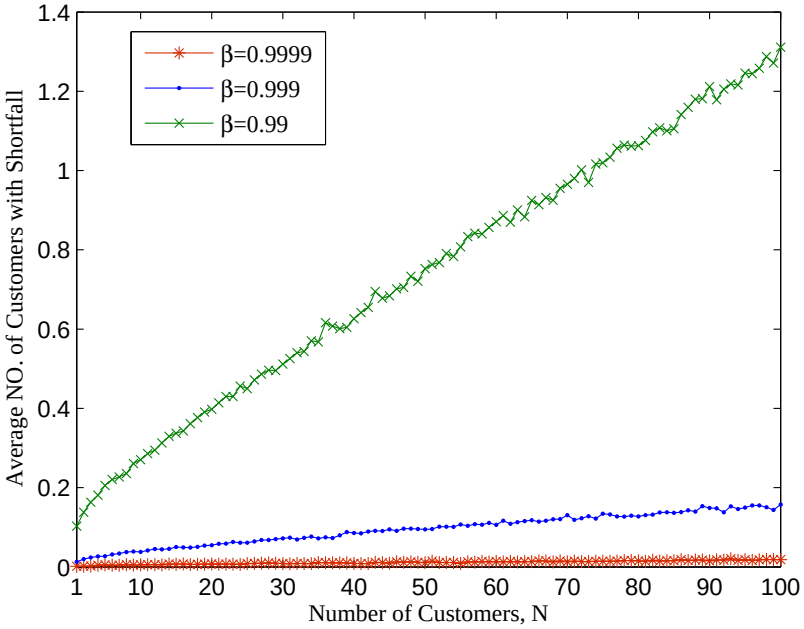


Figure 5 No. of customers who suffered shortfall on average as N varies

her demand met in full. By choosing $\beta = 0.9999$ and using around 522 units of inventory, we can virtually assure that all customers will be able to get their demand met in full.

Remark: The above also suggest a natural solution to the problem posed in the introduction. In order to serve the e-distributors as if Q_i units of inventory are reserved for each of them in the virtual inventory system, the manufacturer can do the following:

- Partition the demand of each customer into two groups: $\min\{Q_i, X_i\}$ and $(X_i - Q_i)^+$, where $(x)^+ := \max\{0, x\}$;
- Serve the demand group $\min\{Q_i, X_i\}$ with a very high fill rate, say 0.9999;
- Serve the demand group $(X_i - Q_i)^+$ with fill rate β_i such that

$$0.9999\mathbf{E}\left[\min\left\{Q_i, X_i\right\}\right] + \beta_i \mathbf{E}\left[(X_i - Q_i)^+\right] \geq \mathbf{E}\left[\min\left\{Q_i, X_i\right\}\right].$$

This is possible as long as

$$\mathbf{E}\left[(X_i - Q_i)^+\right] \geq 0.0001 \times \mathbf{E}[\min\{Q_i, X_i\}].$$

Our analysis shows that at a fill rate of 0.9999, all demands from the first group can be met in full with very high probability, i.e., each e-distributor can expect orders up to Q_i units to be filled

only around 0.9495 in this case. Our allocation policy focuses on fill rate performance, but nevertheless can indirectly keep the number of stock out to be small when fill rate is high.

with very high probability, using $(0.9999\sum_{i=1}^{N}\mathbf{E}[\min\{Q_i,X_i\}]+\mathcal{O}(1))$ units of inventory in the system. At the same time, the demand in the second group can be served using $(\sum_{i=1}^{N}\beta_i\mathbf{E}[(X_i-Q_i)^+]+\mathcal{O}(1))$ units of inventory. The total inventory needed is only $\sum_{i=1}^{N}\mathbf{E}[X_i]+\mathcal{O}(1)$. Compared to the case when the manufacturer reserves Q_i units for each e-distributor, this can translate into a significant amount of savings for the manufacturer when the number of e-distributors is large, and $Q_i > \mathbf{E}[X_i]$. ■

4.1. Distributionally Robust Solution

In the rest of this section, we prove that this insight concerning the safety stock requirement holds for all demand distributions with finite first and second moments. Note that the value of $\hat{S}$ depends on the distribution used to model the demand X_i. In practice, the full distributional information may be unavailable or difficult to put together, especially in the case when demands may be correlated. On the other hand, the means and covariances of the demands are often known or at least can be easily estimated. We next derive a bound for the resource pooling problem, using a well-known classical mean-variance bound (Scarf 1958). See also Natarajan et al. (2009) for an alternate derivation and generalization.

PROPOSITION 1. [Scarf (1958)] *Suppose that demands are independent and have known means μ_i's and covariance $\sigma_{i,j}$'s. Then*

$$\mathbf{E}\left[\min\left\{S,\sum_{i=1}^{N}X_i\right\}\right]\geq\frac{1}{2}\left[S+\sum_{i=1}^{N}\mu_i-\sqrt{\left(\sum_{i=1}^{N}\mu_i-S\right)^2+\mathbf{Var}\left[\sum_{i=1}^{N}X_i\right]}\right].$$

Building on Proposition 1, we can obtain a simple upper bound for the optimal resource capacity level in our model.

THEOREM 4. *Suppose that demands have known means and variances. Then*

$$S:=\max_{U\subseteq\{1,2,\ldots,N\}}\left\{\sum_{i\in U}\beta_i\mu_i+\frac{\mathbf{Var}\left[\sum_{i\in U}X_i\right]}{4\sum_{i\in U}(1-\beta_i)\mu_i}\right\}\ \textit{is feasible for (P).}$$

When the demands are independent, the optimal resource capacity level for any allocation policies lies in the following interval:

$$\left[\sum_{i=1}^{N}\beta_i\mu_i,\sum_{i=1}^{N}\beta_i\mu_i+\frac{1}{4}\max_i\frac{\sigma_i^2}{(1-\beta_i)\mu_i}\right].$$

Theorem 4 follows by choosing S to ensure that

$$\frac{1}{2}\left[S+\sum_{i\in U}\mu_i-\sqrt{\left(\sum_{i\in U}\mu_i-S\right)^2+\mathbf{Var}\left[\sum_{i\in U}X_i\right]}\right]\geq\sum_{i\in U}\beta_i\mu_i,\forall U\subseteq\{1,2,\ldots,N\}.$$

This result is interesting because for each customer i, a capacity of $\beta_i\mu_i$ is not sufficient to ensure a fill rate of β_i, since $\mathbf{E}[\min\{\beta_i\mu_i, X_i\}] < \beta_i\mu_i$. However, in a pooled system we need only an additional $(1/4)\max_i \sigma_i^2/[(1-\beta_i)\mu_i]$ units above the theoretical lower bound of $\sum_{i=1}^N \beta_i\mu_i$ to ensure that each customer receives a fill rate of at least β_i. The managerial implication of this bound is clear: If managers need to invest in a lot of capacity to meet the high fill rate demanded by a customer, then it pays to pool into the system the demand of "many" other customers whose fill rate requirement–measured by the term $\sigma_i^2/((1-\beta_i)\mu_i)$–may not be as stringent.

One slight drawback of the above bound is that, when β_i is close to 1, the "constant" term in the upper bound may be prohibitively large. However, as the Scarf's bound is tight for some demand distributions, it is in general impossible to improve on this upper bound. In the worst case, if there exists some $\beta_i = 1$, the upper bound will not be useful. However, we can easily extend our approach to obtain a reasonable upper bound for such an extreme case as shown in Section 5.1.

5. Extensions

In this section, we extend our basic model by incorporating additional service-level constraints, including 100% fill rate requirement, minimal (non-expected) fill rate requirement and finite-horizon fill rates.

5.1. 100% Fill Rate Requirement

Suppose there is a customer (denoted as customer 0) with random demand X_0 and fill rate requirement $\beta_0 = 1$. Similarly, let μ_0 denote the mean of X_0. Although we focus on the case with only one such customer, the results below can be easily generalized to the case in which more than one customer have 100% fill rate requirements.

Note that a 100% fill rate requirement is reasonable only if X_0 is bounded above by a constant, i.e., $\mathbf{P}(X_0 \le \Delta_0) = 1$ for some constant Δ_0. In this case, it is clear that Δ_0 units of capacity is enough in order to satisfy this customer. The question is then how many additional units of capacity, denoted as S, would be needed to meet the fill rate (β_i) requirement of customer i whose demand is X_i for the rest of the customers, $i = 1, 2 \ldots, N$. Note that although Δ_0 units of capacity are set aside for customer 0, the remaining $(\Delta_0 - X_0)$ units will be available for other customers if needed. We can modify our theory to show that the following conditions are both necessary and sufficient for the additional capacity level S in this setting:

$$\sum_{i\in U}\beta_i\mathbf{E}[X_i] \le \mathbf{E}\left[\min\left\{S + \Delta_0 - X_0, \sum_{i\in U} X_i\right\}\right], \forall U \subseteq \{1, 2, ..., N\}. \tag{8}$$

Similarly, an upper bound for the optimal capacity level needed, $(S + \Delta_0)$, is given by the following expression:

$$\max_{U\subseteq\{1,2,\ldots,N\}}\left\{\mu_0 + \sum_{i\in U}\beta_i\mu_i + \frac{\mathbf{Var}\left[X_0 + \sum_{i\in U} X_i\right]}{4\sum_{i\in U}(1-\beta_i)\mu_i}\right\}. \tag{9}$$

This can be used to strengthen the earlier bound when we have customers demanding a 100% fill rate in the system. The analysis for this extension follows exactly the same steps as before. We present the main differences in the derivation in Appendix II.

5.2. Minimal Fill Rate Requirement

Using the idea for the above extension, we can further generalize the results in this paper to capture more realistic conditions.

The fill rate condition discussed so far as stated in (1) is an *expected* fill rate condition. One may argue that it would be difficult to track such performance measures in practice for a single period problem. A natural fix to this issue is to consider an additional fill rate condition that requires a certain proportion of the realized demand to be satisfied in all scenarios, i.e.,

$$D_i(\boldsymbol{X}, S) \geq \beta_i^r X_i, \ \forall i = 1, 2, \ldots, N, \forall \boldsymbol{X} \in \Omega, \tag{10}$$

where $\beta_i^r \in [0,1]$. We refer to β_i^r as the *minimal* (non-expected) fill rate requirement. For these requirements to be feasible, we must have an upper limit on possible demand realization similar to the case with a 100% fill rate requirement, i.e., there exists $\Delta_{max} := \inf\{\Delta : \mathbf{P}\left(\sum_{i=1}^N \beta_i^r X_i \leq \Delta\right) = 1\}$. Obviously, to satisfy the minimal fill rate requirements, the resource capacity level must be above Δ_{max}. Again similar to the case with a 100% fill rate requirement, we are interested to find out the additional resource required to guarantee the expected fill rate requirements. To avoid trivial cases, we assume $\beta_i \geq \beta_i^r$ for every customer i. We can simply treat $\sum_{i=1}^N \beta_i^r X_i$ as X_0 in the case of a 100% fill rate, and then Δ_{max} is equivalent to Δ_0. The expected fill rate requirement for customer i's remaining demand of $(1-\beta_i^r)X_i$ is then $(\beta_i - \beta_i^r)$. The results in Section 5.1 can be applied to this setting. We thus have the following necessary and sufficient conditions for the additional capacity level S:

$$\sum_{i \in U} (\beta_i - \beta_i^r)\, \mathbf{E}[X_i] \leq \mathbf{E}\left[\min\left\{S + \Delta_{max} - \sum_{i=1}^N \beta_i^r X_i, \sum_{i \in U} (1-\beta_i^r) X_i\right\}\right], \forall U \subseteq \{1, 2, ..., N\}. \tag{11}$$

The upper bound on the optimal capacity level $(S + \Delta_{max})$ can be obtained as follows:

$$\max_{U \subseteq \{1,2,\ldots,N\}} \left\{ \sum_{i=1}^N \beta_i^r \mu_i + \sum_{i \in U} (\beta_i - \beta_i^r)\, \mu_i + \frac{\mathbf{Var}\left[\sum_{i=1}^N \beta_i^r X_i + \sum_{i \in U} (1-\beta_i^r) X_i\right]}{4 \sum_{i \in U} (1-\beta_i)\, \mu_i} \right\}. \tag{12}$$

It is easy to check that if $\beta_i^r = 0$ for all customers, then $\Delta_{max} = 0$ and the above results simply reduce to our previous results for the case with only expected fill rate requirements. Note that these results hold even if only partial customers have minimal fill rate requirements, i.e., $\beta_i^r = 0$ for a subset of customers. These extensions demonstrate the flexibility of our approach that allows us to capture more realistic features.

5.3. Finite-Horizon Fill Rate

So far, our analysis has focused only on the single-period model which is surrounded by the fill rate definition as the expected proportion (ratio) of the demand filled immediately. This ratio can be interpreted as the empirical fill rate observed when the resource allocation problem is solved repeatedly over an infinite horizon. However, in many applications, especially in the 3PL industry, fill rate performance is measured over a finite time horizon, i.e.,

$$\mathbf{E}\left[\frac{\text{Total demand filled within the planning horizon}}{\text{Total demand within the planning horizon}}\right].$$

Thomas (2005) investigated the distribution of finite-period fill rates, using a stationary order-up-to policy, and found that the length of the review periods affects the resource capacity required to meet the targeted (expected) fill rates. Chen et al. (2003) proved that the expected fill rate in a finite horizon setting is always greater than or equal to that in an infinite horizon setting with the same capacity. That is, suppose that the customers' demand $X(t)$'s are i.i.d. across t, and define X as any random variable with the same distribution as $X(t)$; then the following inequality holds for any positive integer T:

$$\mathbf{E}\left[\frac{\min\{X,S\}}{X}\right] \geq \mathbf{E}\left[\frac{\sum_{t=1}^{T}\min\{X(t),S\}}{\sum_{t=1}^{T}X(t)}\right] \geq \frac{\mathbf{E}\left[\min\{X,S\}\right]}{\mathbf{E}\left[X\right]}.$$

What are the benefits of allocation flexibility when the fill rate is measured by performance over a finite horizon of T periods? In this setting, when β_i is the target fill rate measured over T periods, we require

$$\mathbf{E}\left[\frac{D_i(\boldsymbol{X}^{[1]},\boldsymbol{S})+\cdots+D_i(\boldsymbol{X}^{[T]},\boldsymbol{S})}{X_i^{[1]}+\cdots+X_i^{[T]}}\right] \geq \beta_i,\ \forall i=1,2,\ldots,N, \tag{13}$$

where $X_i^{[t]}$ and $D_i(\boldsymbol{X}^{[t]},\boldsymbol{S})$ denote the demand and allocated amount of resources to customer i in period t, respectively, and $X_i^{[t]}$'s are i.i.d. random variables. To adjust for the finite horizon effect in our model, for each customer we map the target fill rate over T time periods onto a new target fill rate over an infinite horizon, so that the technique proposed in the earlier sections can be used to analyze the pooling effect. We also assume that the demands are normal, for ease of exposition.

Let $\mathcal{L}(x)=\mathbf{E}\left[\max\{Z-x,0\}\right]$ represent the standardized normal loss function, where Z is a standard normal random variable. Let $\theta(Q)=\mathbf{P}\left(X<Q\right)$, with X_i being i.i.d. copies of normal random variable X. We use the following approximation:

$$\min\{X_i,Q\} \approx \theta\left(Q\right)\left(X_i-\mu\right)+\mathbf{E}\left[\min\{X_i,Q\}\right].$$

This approximation, built on the work by Zheng et al. (2014), can be viewed as the least-square solution to the stochastic function $\min\{X_i, Q\}$. More details of the approximation procedure and numerical analysis are available in Appendix III. Note that now

$$\begin{aligned}\frac{\sum_{i=1}^{K}\min\{X_i,Q\}}{\sum_{i=1}^{K}X_i} &\approx \frac{\sum_{i=1}^{K}\{\theta(Q)(X_i-\mu)+\mathbf{E}[\min\{X_i,Q\}]\}}{\sum_{i=1}^{K}X_i}\\ &= \frac{\sum_{i=1}^{K}\left\{\theta(Q)X_i-\theta(Q)\mu+Q+\sigma\mathbf{E}\left[\min\left\{\frac{X_i-\mu}{\sigma}+\frac{\mu-Q}{\sigma},0\right\}\right]\right\}}{\sum_{i=1}^{K}X_i}\\ &= \theta(Q)+\frac{K\left[Q-\theta(Q)\mu-\sigma\mathcal{L}\left(\frac{\mu-Q}{\sigma}\right)\right]}{\sum_{i=1}^{K}X_i}.\end{aligned}$$

We need to determine the minimum Q such that

$$\mathbf{E}\left[\theta(Q)+\frac{K\left[Q-\theta(Q)\mu-\sigma\mathcal{L}\left(\frac{\mu-Q}{\sigma}\right)\right]}{\sum_{i=1}^{K}X_i}\right]\geq\beta.$$

This can be written as:

$$\min\left\{Q: K\mathbf{E}\left[\frac{1}{\sum_{t=1}^{K}X(t)}\right]\geq\frac{\beta-\theta(Q)}{Q-\theta(Q)\mu-\sigma\mathcal{L}\left(\frac{\mu-Q}{\sigma}\right)}\right\}. \tag{14}$$

For any given period K, we can solve (14) to obtain the optimal resource capacity level S_K needed. With resource level at S_K, however, we can achieve a long-run average fill rate of β' for the infinite horizon model. In this way, we can map any fill rate requirement for a finite horizon model onto a new fill rate requirement for the infinite horizon model and can invoke our method to determine the optimal resource capacity level needed for the pooled model.

We apply this heuristic to develop an approach to determining the capacity level for the pooled system with N customers, where customer i requires a minimum expected fill rate of β_i, measured over K periods. We demonstrate the idea, using the following numerical example.

Example 3. Consider three customers with i.i.d demands that are normally distributed, with mean 10 and standard deviation 3. They require expected fill rates of $\beta_1 = 0.85$, $\beta_2 = 0.90$, and $\beta_3 = 0.95$ over a finite horizon of K periods, where $K = 5$, 10, 20.

In Table 3, for each value of K, S_K for customer i is estimated using (14) and the value of β_i. Then, through simulation, we find the actual long-run average fill rate (β_i') that can be achieved for customer i. Using the values of β_1', β_2', and β_3' for each K, we can obtain the minimum capacity level of the resource pool S_K^* needed to achieve these long-run average fill rates, employing the method we derived in Section 3. We use the value of S_K^* to approximate the optimal capacity

Table 3 Pooling Results for a Finite Horizon

	$K=5$		$K=10$		$K=20$	
	S_5	β'	S_{10}	β'	S_{20}	β'
$\beta_1=0.85$	9.30	0.842	9.38	0.847	9.41	0.849
$\beta_2=0.90$	10.28	0.894	10.36	0.898	10.40	0.899
$\beta_3=0.95$	11.73	0.946	11.79	0.948	11.82	0.949
No Pooling	31.31		31.53		31.63	
Pooling	S_5^*	Realized SL	S_{10}^*	Realized SL	S_{20}^*	Realized SL
	28.06	0.849	28.23	0.850	28.30	0.851
		0.901		0.902		0.900
		0.952		0.949		0.951
Pooling Effect	10.38%		10.47%		10.75%	

level that meets the expected finite-horizon fill rates of each β_i, using the largest-debt-first policy for each K. Finally, to complete the analysis, we perform another simulation to find the realized expected service levels in K periods.

For instance, when $K=5$, without resource pooling there should–according to (14)–be a separate capacity of 9.30 for the customer with 85% fill-rate requirement, 10.28 for 90% fill rate, and 11.73 for 95% fill rate. Therefore, the total resource capacity level without pooling is 31.31. However, according to our analysis, using the corresponding β' as the new fill-rate requirement in the infinite horizon model–and using (3)–we need a capacity level of only 28.06 for the pool of resources, bringing a 10.38% pooling benefit. Interestingly, with resource pooling the realized expected fill rates over 5 periods are quite close to the required ones. Similar results hold for other values of K. ■

6. Concluding Remarks

In this paper, we consider the resource pooling problem with differentiated fill rate requirements. A challenging problem here is to find the optimal resource capacity level and the allocation mechanism that ensures each customer's service requirement is met. We obtain a set of necessary and sufficient conditions for the minimum capacity level and construct an anticipative allocation policy that can be easily implemented to deliver the targeted fill rate to each customer. The analysis uses Blackwell's Approachability Theorem, which may be of use in solving other classes of inventory problems. Furthermore, we obtain a distribution-free bound on the resource capacity needed, based on the first two moments of the demand distribution. The results imply that in a system with a sufficient number of customers whose demands are i.i.d., a firm does not need to hold any safety stock to deliver the required service levels as long as the fill rates all have an upper bound that is strictly less than one.

We note that resource allocation is a core problem in many supply chain planning models. The analytical results derived in this paper could find wider use in many other settings. For instance,

we can extend the main results derived in this paper to more general production network, to relate service level requirements to capacity configuration in the network. Finding a tractable closed form bound for this setting will be challenging though. It would also be interesting to explore further the connection between our approach and the threshold-type policies used for dynamic allocation of resources with multiple demand classes. The related issue of contract design in a supply chain operating under a service level agreement, using the optimal allocation rule, appears to be amenable to analysis using this approach. We leave these and other issues for future research.

Acknowledgments

We thank the Department Editor, the Associate Editor, and the three referees for their constructive and insightful comments on earlier versions of the paper. We are grateful to Professor Serguei Netessine and Professor Stephen Graves for their careful reading of our paper and the constructive comments they made for improving the paper. We would also like to thank seminar participants at Tsinghua University, New York University, Singapore Management University, Nanyang Technological University, and Singapore University of Technology and Design for their active participation in our seminar presentations. Their feedback has helped us to clarify many issues pertaining to our model and its connection with existing literature. This work was supported in part by the NUS Initiatives in Operations Research and Analytics, Agency for Science Technology and Research [Grant 1122904020, R314-000-091-305], National Natural Science Foundation of China [Grant 71501077, 71371075, 71420107024] and the Guangdong Natural Science Foundation [Grant 2014A030310212].

References

Alptekindğlu, A., A. Banerjee, A. Paul, N. Jain. 2013. Inventory pooling to deliver differentiated service. *Manufacturing & Service Operations Management*, **15(1)**, 33–44.

Axsäter, S. 2003. Note: Optimal policies for serial inventory systems under fill rate constraints. *Management Science*, **49(2)**, 247–253.

Benjaafar, S., Y. Li, D. Xu, S. Elhedhli. 2008. Demand allocation in systems with multiple inventory locations and multiple demand sources. *Manufacturing & Service Operations Management*, **10(1)**, 43–60.

Bensoussan, A., Feng, Q., Sethi, S.P. 2010. Achieving a Long-Term Service Target with Periodic Demand Signals: A Newsvendor Framework. *Manufacturing & Service Operations Management* **13(1)**: 73–88.

Blackwell, D. 1956. An analog of the minimax theorem for venctor payoffs. *Pacific Journal of Mathematics*, **6(1)**, 1–8.

Boyaci, T., G. Gallego. 2001. Serial production-distribution systems under service constraints. *Manufacturing & Service Operations Management*, **3(1)**, 43–50.

Caglar, D., C.L. Li, D. Simchi-Levi. 2004. Two-echelon spare parts inventory system subject to a service constraint. *IIE Transactions*, **36(7)**, 655–666.

Chen, J., D. K. Lin, D. J. Thomas. 2003. On the item fill rate for a finite horizon. *Operations Research Letters*, **31**, 119–199.

Choi, K. S, J. G. Dai, J. S. Song. 2004. On measuring supplier performance under vendor-manager-inventory programs in capacitied supply chains. *Manufacturing & Service Operations Management*, **6(1)**, 53–72.

Chuah, C., R. Katz. 1999. Network provisiong and resource management for IP telephony. Report No. UCB/CSD-99-1061, Augst 1999.

Corbett, C. J., K. Rajaram. 2006. A generalization of the inventory pooling fffect to nonnormal dependent demand. *Manufacturing & Service Operations Management*, **8(4)**, 351–358.

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

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

Ding, Q., P. Kouvelis, J. M. Milner. 2006. Dynamic pricing through discounts for optimizing multiple-class demand fufillment. *Operations Research*, **54(1)**, 169–183.

Eppen, G. D. 1979. Effects of centralization on expected costs in a multi-location newsboy problem. *Management Science*, **25(5)**, 498–501.

Eppen, G. D., L. Schrage. 1981. Centralized ordering policies in a multi warehouse system with lead times and random demand. In *Multi-level Production/Inventory Control System: Theory and Practice*, L. B. Schwarz (ed.), North-Holland, Amsterdam, 51–67.

Erkip, N., W. H. Hausman, S. Nahmias. 1990. Optimal centralized ordering policies in multi-echelon inventory systems with correlated demands. *Management Science*, **36(3)**, 381–392.

Federgruen, A., P. Zipkin. 1984. Approximations of dynamic, multilocation production and inventory problems. *Management Science*, **30(1)**, 69–84.

Gans, N., van Ryzin. 1997. Optimal control of a multiclass, flexible queueing system. *Operations Research*, **45(5)**, 677–693.

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

Ha, A. 1997b. Stock-rationing policy for a make-to-stock production system with two classes and backordering. *Naval Research Logistics*, **44(5)**, 458–472.

Harrison, M. J. 1998. Heavy traffic analysis of a system with parallel servers: asymptotic optimality of discrete-review policies. *Annals of Applied Probability*, **8(3)**, 822–848.

Hopp, W. J., R. Q. Zhang, M. L. Spearman. 1999. An easily implementable hierarchical heuristic for a two-echelon spare parts distribution system. *IIE Transactions*, **31(10)**, 977–988.

Hou, I. H., V. Borkar, P. R. Kumar. 2009. A theory of QoS for wireless. *Proceedings of IEEE INFOCOM 2009*, 486–494.

Hou, I. H., P. R. Kumar. 2009. Admission control and scheduling for QoS guarantees for variable- bit-rate applications on wireless channels. *Proceedings of MobiHoc 2009*, 175–184.

Hou, I. H., P. R. Kumar. 2013. Packets with deadlines: A framework for real-time wireless networks. *Foundations and Trends in Networking, Synthesis Lectures on Communication Networks*, **6(1)**, Morgan & Claypool Publishers, 2013.

Jensen, P. A., J. F. Bard. 2002. Chapter 25 Inventory Theory, in *Operations Research Models and Methods*. John Wiley and Sons, 2002.

Katok, E., D. Thomas, A. Davis. 2008. Inventory service-level agreements as coordination mechanisms: the effect of review periods. *Manufacturing & Service Operations Management*, **10(4)**, 609–624.

Maglaras, C. 2000. Discrete-review policies for scheduling stochastic networks: Trajectory tracking and fluid-scale asymptotic optimality. *Annals of Applied Probability*, **10(3)**, 897–929.

Mak, H. Y., Z. J. M. Shen. 2014. Pooling and dependence of demand and yield in multiple-location inventory systems. *Manufacturing & Service Operations Management*, **16(4)**, 263–269.

Natarajan., K., M. Song, C. P. Teo. 2009. Persistency model and its applications in choice modeling. *Management Science*, **27(3)**, 453–469.

Özer, Ö. 2003. Replenishment strategies for distribution systems under advance demand information. *Management Science*, **49(3)**, 255–272.

Özer, Ö, H. Xiong. 2008. Stock positioning and performance estimation for distribution systems with service constraints. *IIE Transactions*, **40(12)**, 1141–1157.

Scarf, H. 1958. A min-max solution of an inventory problem. In *Studies in The Mathematical Theory of Inventory and Production*, K. Arrow, S. Karlin and H. Scarf (ed.), 201–209.

Symes, S. Sep 20, 2011. The purpose of creating a virtual inventory. *Houston Chronicle*. Retrieved from `http://smallbusiness.chron.com/purpose-creating-virtual-inventory-24491`.

Swaminathan, J. M., R. Srinivasan. 1999. Managing individual customers service constrains under stochastic demand. *Operations Research Letters*, **24(3)**, 477–482.

Swinney, R. 2012. Inventory pooling with strategic consumers: Operational and behavior benefits. *Working Paper*, Duke University.

Thomas, D. J. 2005. Measuring item fill-rate performance in a finite horizon. *Manufacturing & Service Operations Management*, **7(1)**, 74–80.

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 Houtum, G.J., W.H.M. Zijm. 2000. On the relationship between cost and service models for general inventory systems. *Statistica Neerlandica*, **54(2)**, 127–147.

van Mieghem, J., N. Rudi. 2002. Newsvendor networks: inventory management and capacity investment with discretionary activities. *Manufacturing & Service Operations Management*, **4(4)**, 313–335.

Yu, Y., X. Chen, F. Zhang. 2015. Dynamic capacity management with general upgrading. *Operations Research*, **63(6)**, 1372–1389.

Zhang, J. 2003. Managing multi-customer service level requirements with a simple rationing policy. *Operations Research Letters*, **31(6)**, 477–482.

Zhang, J., M. Sobel. 2012. Interchanging fill rate constraints and backorder costs in inventory models. *International Journal of Mathematics in Operational Research*, **4(4)**, 453–472.

Zheng, Z., K. Natarajan, C. P. Teo. 2016. Least squares approximation to the distribution of project completion times with Gaussian uncertainty. *Working paper*. To appear in *Operations Research*.

Appendix I. Proofs of Results in Section 3

Proof of Theorem 1(1). We first prove that the conditions given in (3) are necessary. The sufficiency is proved later in Theorem 1(2) by explicitly constructing an allocation policy that achieves the fill rate requirement using the minimum S that satisfies (3).

Consider a feasible allocation policy; then for any subset $U \subseteq \{1, 2, \ldots, N\}$ and for any t, we have

$$\sum_{i \in U} D_i(t) \leq S,$$

On the other hand, it is obvious that

$$\sum_{i \in U} D_i(t) \leq \sum_{i \in U} X_i(t),$$

Therefore, for any subset U,

$$\sum_{i \in U} D_i(t) \leq \min \left\{ S, \sum_{i \in U} X_i(t) \right\},$$

By taking the long-run average, we have the following necessary condition for any subset U:

$$\sum_{i \in U} \liminf_{T \to \infty} \frac{\sum_{t=1}^{T} D_i(t)}{T} \leq \mathbf{E} \left[\min \left\{ S, \sum_{i \in U} X_i \right\} \right], \, a.s.$$

Note that, from Lemma 1, whose proof is given below, we obtain

$$\liminf_{T \to \infty} \frac{\sum_{t=1}^{T} D_i(t)}{T} \geq \beta_i \mathbf{E}\left[X_i\right], \, a.s.$$

Therefore, the necessary conditions in (3) hold. ∎

Proof of Lemma 1. According to the definition of a fill rate in (2), we have

$$\liminf_{T \to \infty} \frac{\sum_{t=1}^{T} D_i(t)}{\sum_{t=1}^{T} X_i(t)} \geq \beta_i, \, a.s.$$

Looking to the left-hand side (LHS) of the above inequality, we see that

$$\begin{aligned} \liminf_{T \to \infty} \frac{\sum_{t=1}^{T} D_i(t)}{\sum_{t=1}^{T} X_i(t)} &= \liminf_{T \to \infty} \frac{\sum_{t=1}^{T} D_i(t)}{T} \cdot \frac{T}{\sum_{t=1}^{T} X_i(t)} \\ &= \frac{1}{\mathbf{E}\left[X_i\right]} \liminf_{T \to \infty} \frac{\sum_{t=1}^{T} D_i(t)}{T}, \, a.s., \end{aligned}$$

Thus

$$\liminf_{T \to \infty} \frac{\sum_{t=1}^{T} D_i(t)}{\sum_{t=1}^{T} X_i(t)} \geq \beta_i, \, a.s.,$$

if and only if

$$\liminf_{T \to \infty} \frac{\sum_{t=1}^{T} D_i(t)}{T} \geq \beta_i \mathbf{E}\left[X_i\right], \, a.s.,$$

which is the result in Lemma 1. ■

Proof of Theorem 1(2). Recall that we use the debt, $r_i(t+1)$, to determine the priority policy in sample $(t+1)$. The debt for sample $(t+1)$, $R_i(t+1)$, is defined as $\beta_i \mathbf{E}[X_i] - D_i(t+1)$. The average debt from sample 1 up to sample $(t+1)$ is

$$\rho_i(t+1) = \frac{r_i(t+1) + R_i(t+1)}{t+1}.$$

Let $\mathcal{D}$ denote the set of nonpositive orthants in $\mathbb{R}^N$. According to Lemma 1, we want the average debt to approach $\mathcal{D} := \{\boldsymbol{z} = [z_1, z_2, ..., z_N] : z_i \le 0, \forall i = 1, 2, \ldots, N\}$, to ensure that customer i receives a service level of at least β_i.

Consider a point $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_n)$ not in $\mathcal{D}$, where

$$\delta_i = \rho_i(t) = \frac{r_i(t+1)}{t},$$

for some t. Without loss of generality, we assume that $\delta_1 \ge \delta_2 \ge \cdots \ge \delta_m > 0 \ge \delta_{m+1} \ge \delta_{m+2} \ge \cdots \ge \delta_N$. The key idea is to allocate the resources to those with highest debt first. The point in $\mathcal{D}$ closest to $\boldsymbol{\delta}$ is $\boldsymbol{\gamma} = (0, 0, \ldots, 0, \delta_{m+1}, \delta_{m+2}, \ldots, \delta_N)$. The hyperplane perpendicular to the line segment $\boldsymbol{\delta\gamma}$ is given by $H := \{\boldsymbol{z} \in \mathbb{R}^N : \sum_{i=1}^m z_i\delta_i = 0\}$. Note that $\boldsymbol{\delta} \in H^+ := \{\boldsymbol{z} \in \mathbb{R}^N : \sum_{i=1}^m z_i\delta_i > 0\}$. We need to show only that the mean debt $(\mathbf{E}[R_1(t+1)|\boldsymbol{A}^{DF}], \mathbf{E}[R_2(t+1)|\boldsymbol{A}^{DF}], \ldots, \mathbf{E}[R_N(t+1)|\boldsymbol{A}^{DF}])$ lies in $H^- \bigcup H := \{\boldsymbol{z} \in \mathbb{R}^N : \sum_{i=1}^m z_i\delta_i \le 0\}$, where $H^- := \{\boldsymbol{z} \in \mathbb{R}^N : \sum_{i=1}^m z_i\delta_i < 0\}$.

According to our anticipative policy, the allocation process stops at the $(k+1)$th customer in the priority list once the remaining resources, $\hat{S} - \sum_{i=1}^k D_i(t+1)$, are less than 0. That is, $D_{k+1}(t+1) = 0$ if $\sum_{i=1}^{k-1} D_i(t+1) \le \hat{S}$ and $\sum_{i=1}^k D_i(t+1) \ge \hat{S}$. Therefore,

$$\sum_{i=1}^k D_i(t+1) \ge \min\left\{\hat{S}, \sum_{i=1}^k X_i(t+1)\right\}, \forall k = 1, 2, \ldots, m. \tag{15}$$

Since $\sum_{i=1}^k \beta_i \mathbf{E}[X_i] \le \mathbf{E}\left[\min\left\{\hat{S}, \sum_{i=1}^k X_i(t+1)\right\}\right]$ from (3), we get

$$\varpi_k := \sum_{i=1}^k \mathbf{E}\left[R_i(t+1)|\boldsymbol{A}^{DF}\right] = \mathbf{E}\left[\sum_{i=1}^k \beta_i \mathbf{E}[X_i] - \sum_{i=1}^k D_i(t+1)\right] \le 0, \forall k = 1, 2, \ldots, m. \tag{16}$$

Consequently, we have

$$\varpi_k \delta_k \le 0, \forall k = 1, 2, \ldots, m.$$

Therefore,

$$\begin{aligned}
\sum_{i=1}^m \mathbf{E}\left[R_i(t+1)|\boldsymbol{A}^{DF}\right]\delta_i &= \varpi_1\delta_1 + \sum_{i=2}^m (\varpi_i - \varpi_{i-1})\delta_i \\
&\le \varpi_1\delta_1 + \sum_{i=2}^m \varpi_i\delta_i - \sum_{i=2}^m \varpi_{i-1}\delta_{i-1} \\
&= \varpi_m\delta_m \\
&\le 0,
\end{aligned}$$

where the second inequality follows from $\delta_{i-1} \geq \delta_i$. By Blackwell's Approachability Theorem, our allocation policy approaches $\mathcal{D}$. Therefore, the long-run average performance of our policy attains the desired service level requirements for all customers. ■

Proof of Theorem 3. From the randomizing mechanism in Algorithm 1, we have

$$\mathbf{E}\left[D_i^L(\boldsymbol{X},S)\right] = \mathbf{E}\left[\frac{\sum_{t=1}^T D_i^{L(t)}(\boldsymbol{X},S)}{T}\right],$$

where the superscripts are used to emphasize the priority lists that are used to carry out the allocation. Since $\boldsymbol{X}$ follows the same distribution as $\boldsymbol{X}(t)$ and the latter is independent of $L(t)$,

$$\mathbf{E}\left[D_i^{L(t)}(\boldsymbol{X},S)\right] = \mathbf{E}\left[D_i^{L(t)}(\boldsymbol{X}(t),S)\right].$$

Therefore,

$$\mathbf{E}\left[D_i^L(\boldsymbol{X},S)\right] = \mathbf{E}\left[\frac{\sum_{t=1}^T D_i^{L(t)}(\boldsymbol{X}(t),S)}{T}\right] = \mathbf{E}\left[\frac{\sum_{t=1}^T D_i(t)}{T}\right] \geq \beta_i \mathbf{E}\left[X_i\right], \text{ for large enough } T,$$

and the proof is completed.

Proof of Theorem 4. To ensure that

$$\frac{1}{2}\left[S + \sum_{i=1}^n \mu_i - \sqrt{\left(\sum_{i=1}^n \mu_i - S\right)^2 + \mathbf{Var}\left[\sum_{i=1}^n X_i\right]}\right] \geq \sum_{i=1}^n \beta_i \mu_i,$$

we need

$$\left[S + \sum_{i=1}^n (1-2\beta_i)\,\mu_i\right]^2 \geq \left(\sum_{i=1}^n \mu_i - S\right)^2 + \mathbf{Var}\left[\sum_{i=1}^n X_i\right],$$

i.e.,

$$4\left(S + \sum_{i=1}^n \beta_i \mu_i\right)\sum_{i=1}^n (1-\beta_i)\,\mu_i \geq \mathbf{Var}\left[\sum_{i=1}^n X_i\right].$$

Therefore, to obtain an S that satisfies (3) for all U, we need

$$S \geq \sum_{i\in U} \beta_i \mu_i + \frac{\mathbf{Var}\left[\sum_{i\in U} X_i\right]}{4\sum_{i\in U}(1-\beta_i)\mu_i}, \forall\, U \subseteq \{1,2,\ldots,N\}.$$

This concludes the first part of the theorem.

When the demands are independent, it is obvious that

$$\sum_{i=1}^N \beta_i \mu_i + \frac{1}{4}\max_i \frac{\sigma_i^2}{(1-\beta_i)\mu_i} \geq \max_{U\subseteq\{1,2,\ldots,N\}}\left\{\sum_{i\in U}\beta_i\mu_i + \frac{\sum_{i\in U}\sigma_i^2}{4\sum_{i\in U}(1-\beta_i)\mu_i}\right\},$$

which gives us the upper bound on the optimal S. On the other hand, it is easy to see that $\sum_{i=1}^N \beta_i\mu_i$ is a lower bound for the resource capacity needed for any allocation policy. Thus, we complete the proof. ■

Appendix II. Extension for 100% Fill Rate Requirement

It is obvious that the conditions presented as (8) in Section 5.1 are necessary conditions for the additional capacity S. The proof follows exactly the same logic as Theorem 1 with the difference in notation only.

To show that these conditions are also sufficient, we can adopt the same allocation policy with a modification that X_0 always has the highest priority, while the rest of the customers are served according to the randomized largest-debt-first policy using the remaining capacity. We can follow the same setup as in the proof of Theorem 1(2) to show the sufficiency of (8), with only one concern that whether (16) still holds, i.e., $\varpi_k \le 0$, $\forall k = 1, 2, \ldots, m$. To check this, we first note that (15) is still true with the modified allocation policy, i.e., the allocation stops either when the remaining capacity is exhausted or all the remaining customers are served,

$$\sum_{i=1}^{k} D_i(t+1) \ge \min\left\{\hat{S} + \Delta_0 - X_0(t+1), \sum_{i=1}^{k} X_i(t+1)\right\}, \forall k = 1, 2, \ldots, m,$$

where $\hat{S}$ now denotes the minimum value of S that satisfies (8). Taking expectations on both sides of the above inequality and combing with (8), we get

$$\mathbf{E}\left[\sum_{i=1}^{k} D_i(t+1)\right] \ge \mathbf{E}\left[\min\left\{\hat{S} + \Delta_0 - X_0(t+1), \sum_{i=1}^{k} X_i(t+1)\right\}\right] \ge \sum_{i=1}^{k} \beta_i \mathbf{E}[X_i], \forall k = 1, 2, \ldots, m.$$

Consequently,

$$\varpi_k = \sum_{i=1}^{k} \mathbf{E}\left[R_i(t+1) | \boldsymbol{A}^{DF}\right] = \mathbf{E}\left[\sum_{i=1}^{k} \beta_i \mathbf{E}[X_i] - \sum_{i=1}^{k} D_i(t+1)\right] \le 0, \forall k = 1, 2, \ldots, m.$$

Then all the other steps in derivation follow easily, and therefore, we can conclude that the conditions in (8) are both necessary and sufficient.

The proof of the new upper bound in (9) follows the same algebra as in the proof of Theorem 4. For completeness, we present all the steps below. For every subset $U \subseteq \{1, 2, \ldots, N\}$,

$$\begin{aligned}
&\frac{1}{2}\left[S + \Delta_0 - \mu_0 + \sum_{i\in U} \mu_i - \sqrt{\left(S + \Delta_0 - \mu_0 - \sum_{i\in U} \mu_i\right)^2 + \mathbf{Var}\left[\sum_{i\in U} X_i + X_0\right]}\right] \ge \sum_{i\in U} \beta_i \mu_i \\
\iff &\left[S + \Delta_0 - \mu_0 + \sum_{i\in U} (1 - 2\beta_i)\mu_i\right]^2 \ge \left(S + \Delta_0 - \mu_0 - \sum_{i\in U} \mu_i\right)^2 + \mathbf{Var}\left[\sum_{i\in U} X_i + X_0\right] \\
\iff &4\left[\sum_{i\in U} (1 - \beta_i)\mu_i\right]\left(S + \Delta_0 - \mu_0 - \sum_{i\in U} \beta_i \mu_i\right) \ge \mathbf{Var}\left[\sum_{i\in U} X_i + X_0\right] \\
\iff &S + \Delta_0 \ge \sum_{i\in U} \beta_i \mu_i + \mu_0 + \frac{\mathbf{Var}\left[\sum_{i\in U} X_i + X_0\right]}{4\sum_{i\in U}(1 - \beta_i)\mu_i}.
\end{aligned}$$

Appendix III. Approximating Finite-Horizon Fill Rates

Consider a finite horizon order-up-to system, where the demand X_i's are i.i.d., with mean μ and standard deviation σ. Let Q denote the order-up-to level. We are interested in approximating the fill rate over K periods, given by

$$\frac{\min\{X_1,Q\}+\ldots+\min\{X_K,Q\}}{X_1+\ldots+X_K}.$$

Chen et al. (2003) established that

$$\mathbf{E}\left[\frac{\min\left\{\tilde{D}_1,Q\right\}}{X_1}\right]\geq\mathbf{E}\left[\frac{\sum_{i=1}^{K}\min\{X_i,Q\}}{\sum_{i=1}^{K}X_i,}\right]\geq\frac{\mathbf{E}\left[\min\{X_1,Q\}\right]}{\mathbf{E}\left[X_1\right]},\ \forall K=1,2,\ldots.$$

They argued that the distribution of the fill rate measurement affects the stocking decision.Therefore, the choice of the planning horizon in the fill rate measurement is an important consideration in the design of the fill rate target. Define $\theta(Q)=\mathbf{P}(X_1<Q)$. Assume X_i's are normally distributed (truncated to remove negative values). Let $\mathcal{L}(x)=\mathbf{E}[\max\{Z-x,0\}]$ denote the standardized normal loss function, where Z is the standard normal random variable. Note that $\mathcal{L}(x)=-\mathbf{E}[\min\{Z+x,0\}]$. Then

$$\begin{aligned}\frac{\sum_{i=1}^{K}\min\{X_i,Q\}}{\sum_{i=1}^{K}X_i}&\approx\frac{\sum_{i=1}^{K}\{\theta(Q)(X_i-\mu)+\mathbf{E}\left[\min\{X_i,Q\}\right]\}}{\sum_{i=1}^{K}X_i}\\&=\frac{\sum_{i=1}^{K}\left\{\theta(Q)X_i-\theta(Q)\mu+Q+\sigma\mathbf{E}\left[\min\left\{\frac{X_i-\mu}{\sigma}+\frac{\mu-Q}{\sigma},0\right\}\right]\right\}}{\sum_{i=1}^{K}X_i}\\&=\theta(Q)+\frac{K\left[Q-\theta(Q)\mu-\sigma\mathcal{L}\left(\frac{\mu-Q}{\sigma}\right)\right]}{\sum_{i=1}^{K}X_i},\end{aligned}$$

where the first approximation is based on Theorem 1 on least squares approximation from Zheng et al. (2016). Therefore,

$$\begin{aligned}\mathbf{E}\left[\frac{\sum_{i=1}^{K}\min\{X_i,Q\}}{\sum_{i=1}^{K}X_i}\right]&\approx\mathbf{E}\left[\theta(Q)+\frac{K\left(Q-\theta(Q)\mu-\sigma\mathcal{L}\left(\frac{\mu-Q}{\sigma}\right)\right)}{\sum_{i=1}^{K}X_i}\right]\\&=\theta(Q)+\left[Q-\theta(Q)\mu-\sigma\mathcal{L}\left(\frac{\mu-Q}{\sigma}\right)\right]\mathbf{E}\left[\frac{K}{\sum_{i=1}^{K}X_i}\right].\end{aligned}$$

Example 4. Consider the case when the i.i.d. demands X_i are normally distributed with mean $\mu=10$ and standard deviation $\sigma=3$. The mean fill rates obtained from simulation and the linear estimator developed above are summarized in Table 4.

The mean of the estimator obtained using our approximation approach is surprisingly close to the simulated fill rate performance, for all values of Q. Moreover, the effect of the review periods is more visible for small Q. For instance, when $Q=6$, the mean fill rate is around 61% when $K=2$, but it drops to 59% when K is around 10 to 20.

Table 4 Comparison between simulated and estimated finite-horizon fill rates

Q	$K=2$		$K=10$		$K=20$	
	Simulation	Estimation	Simulation	Estimation	Simulation	Estimation
6	0.6116	0.6135	0.5918	0.5919	0.5895	0.5895
7	0.6997	0.7023	0.6797	0.6798	0.6773	0.6774
8	0.7780	0.7812	0.7592	0.7593	0.7569	0.7570
9	0.8441	0.8477	0.8278	0.8279	0.8258	0.8258
10	0.8969	0.9004	0.8837	0.8838	0.8820	0.8821
11	0.9360	0.9392	0.9264	0.9264	0.9251	0.9251
12	0.9630	0.9656	0.9565	0.9566	0.9556	0.9556
13	0.9802	0.9821	0.9762	0.9762	0.9756	0.9756
14	0.9902	0.9914	0.9880	0.9880	0.9876	0.9876

This method can be used to determine the Q needed to meet a target (expected) fill rate of β, given the number of review periods K:

$$\theta(Q)+\left[Q-\theta(Q)\,\mu-\sigma\mathcal{L}\left(\frac{\mu-Q}{\sigma}\right)\right]\mathbf{E}\left[\frac{K}{\sum_{i=1}^{K}X_i}\right]=\beta,$$

i.e.,

$$K\mathbf{E}\left[\frac{1}{\sum_{i=1}^{K}X_i}\right]=\frac{\beta-\theta(Q)}{Q-\theta(Q)\,\mu-\sigma\mathcal{L}\left(\frac{\mu-Q}{\sigma}\right)}.$$

Figure 6 shows how the right-hand side (RHS) varies as a function of Q for selected β, for the instance in Example 1. For $K=2$ (with corresponding LHS value of 0.1053), to attain a service level of $\beta=0.7$ the graph indicates that we need Q to be only around 7. However, for a target of $\beta=0.99$ we need Q to be around 13.

Figure 6 RHS as a function of Q and β