Inventory Pooling to Deliver Differentiated Service

Aydın Alptekinoğlu
Cox School of Business, Southern Methodist University, Dallas, TX 75275
aalp@cox.smu.edu

Arunava Banerjee
Department of Computer and Information Science and Engineering
University of Florida, Gainesville, FL 32611, arunava@cise.ufl.edu

Anand Paul
Warrington College of Business Administration
University of Florida, Gainesville, FL 32611, anand.paul@warrington.ufl.edu

Nikhil Jain
Servigistics Inc., 2300 Windy Ridge Parkway, 450 North Tower
Atlanta, GA 30339, njain@servigistics.com

March 15, 2012

Abstract

Inventory pooling is at the root of many celebrated ideas in operations management. Postponement, component commonality, and resource flexibility are some examples. Motivated by our experience in the after-market services industry, we propose a model of inventory pooling to meet differentiated service levels for multiple customers. Our central research question is the following: What are the minimum inventory level and optimal allocation policy when a pool of inventory is used in a single period to satisfy individual service levels for multiple customers? We measure service by the probability of fulfilling a customer’s entire demand immediately from stock. We characterize the optimal solution in several allocation policy classes, provide some structural results, formulas and bounds, and also make detailed inter-policy comparisons. We show that the pooling benefit is always strictly positive, even when an arbitrary number of customer demands are perfectly positively correlated.
1 Introduction

Inventory pooling, the practice of using a common pool of inventory to satisfy two or more sources of random demand, has been studied in the context of many operationally challenging situations. For example, there are large streams of literature that explore how pooling acts as an essential ingredient in containing the operational costs of high product variety, in mitigating supply chain disruptions, and in striking the right tradeoff between operational benefits and fixed costs of product-process flexibility in supply chains (Lee 2004).

In this paper, we analyze a single-period model that captures inventory pooling in an environment where customers' service expectations differ; hence, the policy by which inventory is allocated becomes critical, should one wish to reap the benefits of inventory pooling. We pose a fundamental question: When a pool of inventory is used to serve customers with varying service level requirements, what are the minimum inventory level and optimal allocation policy? We measure service by the probability of meeting a customer's entire demand immediately from stock (type-1 service measure; Silver et al. 1998, p. 245).

This is the essence of a problem that frequently occurs in after-market service operations, an industry sector estimated to make up 8% of GDP in the US (Cohen et al. 2006), when certain levels of service need to be maintained for a collection of current and relatively long-term contracts at minimum cost. In such settings, the total revenue from trade is fixed, because demands are eventually satisfied and prices are contractually set, even though they may vary from one customer to another. Often, service level requirements of customers differ. We frame the problem as follows: find the combination of inventory level and allocation policy that maintains a set of current contracts most efficiently.

Our model is highly stylized; it assumes a single period in which the firm uses a type-1 service measure and batches demands from multiple customers before attempting to fulfill them. We observed how HOLT CAT, Caterpillar's Texas dealership, manages spare parts inventories. The most important measure of service they monitor for each store is called On Time In Full (OTIF), the percentage of spare parts orders fully satisfied on time, because their customers often see no value in having only a portion of the parts required to perform a repair (Barry 2006). Moreover, they discourage urgent orders, because non-urgent orders for such parts as air filters are typically batch-processed overnight rather than immediately upon order receipt, which is more costly. The totality of HOLT CAT’s operation is, of course, much more complicated than what is suggested by our stylized model. For example, the time component of OTIF is only crudely captured, and inventory replenishment and demand-batching may not always be synchronized even for non-urgent items. Nevertheless, we hope that our model might serve as a building block for more complex and realistic models in this area.

What piqued our interest the most in industry practice is the decoupling of ordering and allocation decisions. In our model, we treat the ordering decision, which sets the spare parts inventory level, and
the allocation decision, which rations the available inventory among customers via a prioritization scheme of some sort, simultaneously. By optimizing jointly over allocation policies as well as inventory levels, we demonstrate the benefits of integrating these decisions.

Our main goal is to find analytical characterizations of the optimal inventory level and allocation policy for customers with different service level requirements. We define three classes of allocation policies, and obtain structural results and formulas that optimize jointly over inventory level as well as allocation policy. We find the optimal solution for two customers with arbitrary demand distributions, but require \textit{iid} (independent and identically distributed) demands for three or more customers. We demonstrate the advantages of interlinking inventory and allocation decisions, and give insights into when less sophisticated allocation policies are almost as good as the optimal policy. Finally, we have a result that is in contrast with backorder-cost models, in which the pooling benefit is zero when demands are perfectly positively correlated (Eppen 1979). In our service-level-constrained model, we show analytically that the pooling benefit is strictly positive even if demands are perfectly positively correlated. We relegate all proofs to Appendix A (see the Online Companion).

2 Literature Review

In positioning our paper, we find the broad framework presented by Ozer and Xiong (2008) useful. They identify four quadrants into which many inventory models can be slotted. The dimensions underlying their matrix are (1) backorder-cost or service-level models and (2) single or multiple demand points. Our paper fits into the fourth quadrant; for completeness we review representative papers in the two quadrants most relevant to our paper: backorder-cost and service-level models with multiple demand points. Even within a given quadrant, researchers make different modeling choices: continuous review versus periodic review, optimizing the parameters of an assumed allocation policy versus finding the form of the optimal allocation policy, and finite horizon versus infinite horizon.

We remark that the setting in the present paper and the settings in the bulk of the literature reviewed below are not directly comparable as they apply to different distribution environments. For instance, in many models, customers carry inventory; the warehouse may (or may not) carry inventory at a central location and allocates inventory to satisfy the replenishment requests from downstream retailers (who sell to end-consumers). Further, the literature generally does not assume allocation can be made after demand is realized; that is, demands are not batched during a period but are satisfied in real time. These models capture a context in which it would be too late to wait for demand realizations before making allocation decisions given the positive shipment lead times between the warehouse and the retailers.

Our paper is most closely related to Swaminathan and Srinivasan (1999) and Zhang (2003). Swaminathan
and Srinivasan develop an algorithm to compute the optimal ordering and allocation policies for the same problem that we study. The combinatorial complexity of the problem, and hence the difficulty of obtaining a practical solution efficiently, is evident from their paper. Switching iteratively between binary search and Monte Carlo simulation, their approach is necessarily computational and exponential-time, as they pose the problem in its most general form without structuring either the space of policies or the demand distributions. In contrast, we emphasize policies that are intuitive and easy to implement, and to that end we provide some structural results, formulas, and bounds. We also make detailed inter-policy comparisons. Zhang (2003) studies a specific class of allocation policies again in a single period, considering the special case when demand distributions and service levels are such that at most one customer’s demand can go unsatisfied.

Eppen and Schrage (1981) study a supplier-depot-multiple-warehouse system in which the warehouses face mutually independent normally distributed demands. At the end of every period, an aggregate replenishment order is placed with the supplier. Replenishment stock is routed through the depot, where an allocation rule has to be framed for distributing stock to the warehouses. The following allocation rule is assumed: stock is distributed so as to equalize type-1 service levels at the warehouses. This allocation rule is feasible when demands are stable, but may otherwise be infeasible. Assuming that the rule is feasible, the authors develop an expression for the value of the order that minimizes the sum of expected holding and backorder costs.

Schwarz et al. (1985) study a one-warehouse multiple-retailer system in which all the entities hold stock and follow continuous-review (Q, R) policies. Each retailer faces independent Poisson demand and receives replenishments from the warehouse, which is replenished by an uncapacitated source. If the retailers run out of stock, they place backorders with the warehouse. The backorders are filled on a first-come-first-serve (FCFS) basis from the warehouse. The problem is to determine lot-sizes and reorder points so as to maximize the fill rate at the warehouse subject to an upper bound on the system inventory.

Hopp et al. (1999) model a spare parts distribution system wherein a distribution center (DC) supports a number of customer facilities that generate Poisson demands. The DC as well as the facilities hold stock; the facilities follow a one-for-one replenishment strategy, and the DC follows a continuous-review replenishment strategy. The problem is to determine the parameters of the ordering policies at the DC and the facilities so as to minimize expected inventory-related costs across the system subject to service level constraints that place upper bounds on the order frequency at the DC and the average delay experienced by each facility.

Caglar et al. (2004) study a distribution system with a similar topology to that in Hopp et al. (1999), but for repairable parts. A fixed number of depots serve customers, each of whom owns a machine with multiple parts subject to failure. Each depot sees a Poisson arrival process of failed parts. Each failed part is replaced by a spare part from stock or backordered. All failed parts are transported to a central warehouse, where they are repaired. Repair times at the warehouse and transportation times between the central warehouse
and depots are modeled. The problem is to determine basestock levels at the central warehouse and depots so as to minimize the total system-wide inventory holding cost subject to service level constraints in the form of bounds on average response time. A computationally efficient heuristic is presented to solve the problem.

Deshpande et al. (2003) study service level differentiation for two demand classes, each following a Poisson process with different rates. They assume a continuous-review \((Q, R)\) policy for inventory replenishment, and a threshold policy for inventory allocation, which stipulates that lower-priority customers (those with lower shortage cost) are not served at all if inventory-on-hand falls below a threshold level. They study optimal policy parameters and backlog clearing mechanisms.

Arslan et al. (2007) study a problem that is quite close to ours, but they model it differently. They aim to find the optimal parameters of a continuous-review \((Q, R)\) inventory policy for a single stocking point and an allocation policy for a number of customers with differentiated service level requirements. The allocation policy is a natural adaptation of threshold policy to multiple customers. Customer demands are Poisson with different rates. The problem is to find threshold levels and an optimal value of \(R\) (for a given \(Q\)) such that the probability of a strictly positive inventory level exceeds a certain minimum acceptable level, which varies from customer to customer. The authors present an efficient heuristic to solve the problem.

Ozer and Xiong (2008) study a distribution system comprising a warehouse replenishing multiple retailers, each of which operates a continuous-review basestock (one-for-one replenishment) policy. All locations carry inventory. The demand process at each retailer is Poisson; unsatisfied demand is backordered. The warehouse fills requests from the retailers on an FCFS basis. The problem is to determine basestock levels that minimize the system inventory holding cost subject to the following service level constraint: the probability that a demand at each retailer can be filled from existing stock must exceed a threshold level. Bounds and heuristics are developed to determine optimal basestock levels at each location and the ensuing average cost.

Gallego et al. (2007) study allocation mechanisms whereby a central control point (a manager who has access to system-wide inventory levels and costs) makes stock placement decisions for a set of downstream demand points facing Poisson demand with the objective of minimizing expected cost. The same theme of central-versus-local control is explored in Chen (1998), which studies optimal inventory placement in a serial \(N\)-stage system and compares echelon stock (central) and installation stock (local) policies.

In closing, we review a few inventory pooling models. Eppen (1979) shows that there is benefit to inventory pooling in the face of iid normal demands, and studies how this benefit varies as a function of demand correlation and the number of demand points. Erkip et al. (1990) extend the Eppen-Schrage model to the case of correlated demands, both across locations and across time at a given location. Ozer (2003) explores the interplay between advance demand information and inventory pooling. Alptekinoglu and Tang (2005) consider arbitrary numbers of depots and demand locations facing multivariate normal demand.
More broadly, two prominent methods of containing operational costs due to high variety are based on pooling: postponement, also known as delayed product differentiation (Lee and Tang 1997, Aviv and Federgruen 2001), and component commonality (Mirchandani and Mishra 2002, Van Mieghem 2004). Many models of assemble-to-order systems (Akçay and Xu 2004) and resource flexibility (Van Mieghem 1998) have some form of pooling at the core.

### 3 Problem Formulation

A firm supplies a single product to \( N \) customers from a centralized pool of inventory over the duration of a single period. Customer \( i \) has a random demand \( X_i \) and requires a minimum service level of \( \beta_i \in (0, 1) \); the probability that \( X_i \) is fully satisfied must be \( \beta_i \) or more. The \( X_i \)'s are continuous positive-valued random variables with distribution functions \( F_i(\cdot) \) and their sum has a distribution function \( G(\cdot) \).

Events unfold as follows: (1) the firm orders \( S \) units of the product in advance so as to receive them at the beginning of the period; (2) actual customer demands, denoted by \( x_i \), realize throughout the period; (3) at the end of the period, the firm allocates the available pool of inventory (\( S \) units) among \( N \) customers according to an allocation policy, and makes the appropriate shipments. Any leftover inventory is discarded.

An allocation policy in general is a mapping \( A : \mathbb{R}_+^{N+1} \rightarrow \mathbb{R}_+^{N} \) from inventory level and demand realizations \((S, x_1, x_2, \ldots, x_N)\) to inventory allocations \((y_1, y_2, \ldots, y_N)\) resulting in \( y_i = A_i(S, x_1, x_2, \ldots, x_N) \) such that \( y_i \leq x_i \) (no customer receives more inventory than needed) and \( \sum_{i=1}^{N} y_i = \min \left\{ S, \sum_{i=1}^{N} x_i \right\} \) (the firm either depletes its inventory or satisfies all customers), where \( \mathbb{R}_+ \) denotes the set of non-negative real numbers. Let \( \Omega \) be the set of all such mappings.

The firm wants to find the minimum inventory \( S \) coupled with an allocation policy \( A \) that together meet the service level requirements. Both of these decisions are made at the beginning of the period, at which point the outcome of \( A \) in terms of allocating actual quantities to the customers is \textit{a priori} uncertain. That is, at the time of selecting \( S \) and \( A \), demands \( X = (X_1, X_2, \ldots, X_N) \) as well as inventory allocations \( Y = (Y_1, Y_2, \ldots, Y_N) \) that result from applying \( A \) are uncertain; \( Y_i = A_i(S, X) \), the amount of inventory to be allocated to customer \( i \) is a random variable, and customer \( i \)'s demand is fully satisfied if and only if (iff) the event \( Y_i = X_i \) occurs. Service level requirements are therefore in the form of chance constraints.

The firm’s problem can be formally stated as follows (let \( P \{ \cdot \} \) denote probability):

\[
\begin{align*}
\text{Minimize} & \quad S \\
\text{subject to} & \quad P \{ A_i(S, X) = X_i \} \geq \beta_i \quad \text{for all } i = 1, 2, \ldots, N
\end{align*}
\]

where \( \Omega \equiv \{ A : \mathbb{R}_+^{N+1} \rightarrow \mathbb{R}_+^{N} \mid y_i = A_i(S, x_1, x_2, \ldots, x_N) \text{ and } y_i \leq x_i \text{ for } i = 1, \ldots, N, \text{ and } \sum_{i=1}^{N} y_i = \} \).
min\{S, \sum_{i=1}^{N} x_i\}\) is the set of mappings that each specify an allocation of available inventory to customers up to their demands. Note that the mapping \(A\) has to be derived at the beginning, before observing demands, because the firm cannot evaluate the feasibility of \(S\) without specifying \(A\).

## 4 Allocation Policies

We first define a class of allocation policies and show that an optimal policy belongs to this class. A *priority policy* is an allocation policy that leaves at most one customer partially satisfied, i.e., the set \(\{i \in \{1, \ldots, N\} : 0 < y_i < x_i\}\) is either empty or a singleton, for all demand realizations.

**Theorem 1** An optimal allocation policy is a priority policy.

We now offer an alternative definition of a priority policy that is more convenient to work with than the definition based on inventory allocations (\(y\) variables). An allocation policy belongs to the class of priority policies if it operates as follows. First, customers are ordered in a priority list — the sequence by which inventory is ‘doled out’ — before or after demand realizations are observed. Let \(\pi : \{1, \ldots, N\} \rightarrow \{1, \ldots, N\}\) be a one-to-one correspondence between priority list positions and customers. Each priority list \(\Pi = (\pi(1), \ldots, \pi(N))\) is defined by one such correspondence \(\pi\), with \(\pi(j)\) representing the customer in the \(j\)-th position of the priority list. Second, customer demands are filled from the available inventory in decreasing order of priority; demand from customer \(\pi(1)\) is filled first, customer \(\pi(2)\) second, and so on. This sequential allocation process stops when all demands are filled or when the available inventory is exhausted, whichever occurs first.

In this paper, we define and analyze two main classes of allocation policies that are differentiated by whether or not they make use of actual demand information when forming the priority list.

**Responsive Priority Policies:** The priority list \(\Pi\) is constructed using the demand realization information; e.g., smaller demand is filled first (say customer \(A\)'s), and then larger demand (customer \(B\)'s) is filled if there is any stock left over. Since actual demand information is used to determine the priority list, such allocation policies are said to be *responsive*. The set of rules involved in mapping the demand information to a priority list can be quite general. Intuitively, it seems more efficient to fill smaller demands first. At the same time, one needs to recognize that customer demand distributions and service level requirements may differ, so a simple rank ordering based on magnitude of demand is in general unlikely to work. This basic tension between efficient use of inventory and ability to differentiate service levels is a recurring theme of our paper.

**Anticipative Priority Policies:** The priority list \(\Pi\) is constructed without using the demand realization information. We study two particular variations of *anticipative* policies. The first has a deterministic priority list, fixed *a priori* independently of demand realizations; e.g., customer \(A\) always has higher priority than
customer B. Because the priorities are assigned on the basis of a fixed list, we call such a policy a fixed list policy. The second uses a randomized priority list, again independently of demands. One of the N! possible permutations, which corresponds to a unique one-to-one correspondence π, is chosen according to a discrete probability distribution over the set of all possible priority lists; e.g., a coin is tossed before the demands are realized, and if it falls heads (tails), customer A’s (B’s) demand is filled first. In contrast to a fixed list policy, the priority list is decided randomly, so we call such a policy a randomized list policy.

The optimal inventory levels within each policy class (indicated with subscripts) are ordered as follows.

**Theorem 2** $S_{\text{responsive}}^* \leq S_{\text{r-list}}^* \leq S_{\text{f-list}}^*$

In practice it is surely simpler to implement a fixed list policy than either a randomized list or a responsive policy, and it is cleaner and less information-intensive to operate a randomized list policy than a responsive policy. Further, there may be exogenous reasons (such as building long-term relationships with customers) that dictate the adoption of a fixed list policy. For these reasons, we conduct a detailed study of each of these classes of allocation policies, beginning with responsive policies.

### 5 Responsive Priority Policies

Using a responsive priority policy amounts to allowing the priority list to freely depend on demand realizations. One responsive policy that is intuitively appealing, and straightforward to compute and implement is to serve the customers in ascending order of demand realizations. We call this allocation policy the Greedy Policy (GP).

Given a fixed inventory level and any set of demand realizations, there is no allocation rule that completely satisfies more customers than GP does.

Based on GP, we first develop a lower bound on the optimal inventory level for the general problem with an arbitrary set of customer demand distributions and service level requirements. Let the order statistics corresponding to demands (ordered from smallest to largest) be $X_{[1]}, \ldots, X_{[N]}$. We define partial convolutions of the order statistics as follows: $Z_n = \sum_{i=1}^n X_{[i]}$ for $n \in \{1, \ldots, N\}$. Thus, $Z_n$ represents the sum of the $n$ smallest demands; we denote its distribution function by $H_n(.)$.

**Theorem 3** The unique solution $\mathcal{S}$ of the equation $\sum_{n=1}^N H_n(S) = \sum_{i=1}^N \beta_i$ is a lower bound on the optimal inventory level, i.e., $S^* \geq \mathcal{S}$.

If it is not already optimal, this bound represents a good starting point for solving the general problem. The proof uses GP, which is the most efficient policy in using the limited inventory, but GP ignores how service level requirements of customers are dispersed. Therefore, GP may over-serve some customers (the ones who
tend to have low demand values) and under-serve others (high-demand customers). Intuitively speaking, and based on our experience with numerical examples, $S$ is either optimal or near-optimal for problem instances where demand distributions and service level requirements do not differ drastically among customers.

In principle, $H_n(\cdot)$ can be computed analytically by using known facts about the distribution functions of order statistics and their convolutions. But in practice, it is easier to use Monte Carlo simulation to compute it, which is what we did to obtain the lower bound $\bar{S}$.

Without imposing some structure on the demand distributions or a limit on the number of customers, the general problem is difficult due to combinatorics of inventory allocation. In order to glean some structural insights into the problem and obtain analytical characterizations of the optimal solution, we first assume that the demands are iid, but otherwise arbitrary. We then analyze the two-customer problem with arbitrary (possibly non-iid) demands.

### 5.1 IID Demands

We first take the simpler case of undifferentiated service levels.

**Theorem 4** If the demands $X_1, \ldots, X_N$ are iid random variables and service levels are undifferentiated so that $\beta_1 = \cdots = \beta_N = \beta$, then GP is an optimal allocation policy, and the optimal inventory level is the unique solution $S_{GP}(\beta)$ to the equation $\sum_{n=1}^{N} H_n(S) = N\beta$.

Next we analyze the case of customers with iid demands requiring differentiated service levels. We shall see that GP plays an important role in this case also. Suppose $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_N$ with at least one strict inequality, and let $S_{GP}(\bar{\beta}_n)$ be the stock level required by GP to deliver a service level of exactly $\bar{\beta}_n \equiv (\beta_1 + \cdots + \beta_n)/n$ to customers $1, \ldots, n$ for all $n = 1, \ldots, N$. Theorem 3 implies that, with iid demands, $S_{GP}(\bar{\beta}_N)$ is a lower bound for the stock level required by the optimal responsive policy. We shall show that, barring a theoretical degenerate case to be spelled out in the next paragraph, the lower bound is in fact attained; the optimal stock level is $S_{GP}(\bar{\beta}_N)$. Further, the optimal allocation policy involves applying GP to demand realizations after first scaling each demand realization $x_i$ by a fixed scale factor $K_i$.

We assume that the stock level needed to serve a set of customers is a strictly increasing function of the number of customers. A degeneracy arises when this assumption fails to hold: some customers have service levels so low that they can free ride on the remaining stock after all the other customers are served, and still have their service level requirements fulfilled. These free-riders are of no practical interest in our model because we are concerned with customers with contractually committed service levels.

We call an allocation policy a *Cardinal Greedy Policy* (CGP) if, with a given stock level, it satisfies the demands of exactly as many customers as GP would satisfy for every set of demand realizations.
Theorem 5 Suppose customer demands are iid, service levels need to be differentiated, and there exist no free-riders, i.e., $S_{GP}(\tilde{\beta}_{n+1}) > S_{GP}(\tilde{\beta}_n)$ for $n = 1, \ldots, N - 1$. Then, (a) a CGP is an optimal allocation policy, the optimal inventory level is $S_{GP}(\tilde{\beta}_N)$, and (b) there exists an $N$-vector $(K_1, \ldots, K_N)$ such that an optimal allocation policy for each demand realization $(x_1, \ldots, x_N)$ is to prioritize customers either in increasing order of $K_i x_i$ or in increasing order of $x_i$.

This result implies that service level differentiation does not impose an additional inventory burden when demands are iid; servicing a set of customers with distinct service levels $\beta_1, \ldots, \beta_N$ and servicing the same set of customers with a service level of $\bar{\beta}$ for every customer requires an identical inventory level.

In more detail, the following allocation policy is optimal: (i) observe the demand realizations $(x_1, \ldots, x_N)$; (ii) allocate stock to customer $i$ in increasing order of $K_i x_i$ ($i = 1, \ldots, N$) while passing on to the next customer in the list if the current customer has a demand realization that exceeds the remaining stock; (iii) count the number of customers $N_K$ whose demands are completely satisfied with this allocation policy, and compare it with the number of customers $N_G$ whose demands would be completely satisfied by GP; (iv) if $N_K = N_G$, use the allocation policy in (ii) above; otherwise, allocate according to GP.

This allocation policy is in the class of CGP policies, and it is feasible for inventory level $S_{GP}(\tilde{\beta}_N)$ and a given set of service levels, and is therefore optimal. We note, however, that Theorem 5b is an existence result; it asserts that there is an optimal scaling, but does not give us a recipe for finding the optimal scale factors.

5.2 The Two-Customer Case

In this subsection, we show that a particular subclass of responsive policies contains the optimal solution in the two-customer case for any set of service level requirements and demand distributions (possibly non-iid). We treat the special case of bivariate normal demands in Appendix B (see the Online Companion), focusing on how the optimal inventory level ($S^*$) and magnitude of the pooling benefit ($S_1 + S_2 - S^*$) behave as a function of demand correlation and demand variability.

When $N = 2$, the firm’s allocation policy just needs to pick for each demand realization the customer that has the first priority (recall from Theorem 1 and the following discussion that it is sufficient to work with priority lists). Let $\hat{A} : \mathbb{R}_+^3 \to \{1, 2\}$ be a mapping from inventory level and demand realizations $(S, x_1, x_2)$ to a customer identity, with $\hat{A}(S, x_1, x_2)$ specifying the customer who gets the first priority. As in the general formulation (§3), $\hat{A}$ has to be decided before demand realizations are known, hence the customer with the first priority $\hat{A}(S, X_1, X_2)$ is a priori uncertain.

There are five possibilities for demand realizations: (i) if $x_1 + x_2 \leq S$, who gets priority makes no difference, because both customers can be fully satisfied; (ii) if $x_1 \leq S$ and $x_2 > S$, only customer 1 can be fully satisfied;
(iii) if \( x_1 > S \) and \( x_2 \leq S \), only customer 2 can be fully satisfied; (iv) if \( x_1 \leq S, x_2 \leq S \) and \( x_1 + x_2 > S \), inventory level is high enough to satisfy either customer individually but not both; and (v) if \( x_1 > S \) and \( x_2 > S \), neither customer can be fully satisfied. We assume without loss of optimality that \( \hat{A}(S, x_1, x_2) = 1 \) when (ii) happens, and \( \hat{A}(S, x_1, x_2) = 2 \) when (iii) happens (any allocation policy that fails to satisfy these properties for some \( S \) can be improved, in the sense of increasing the service level that it delivers to either customer or both using the same inventory). Let \( \hat{\Omega} \) be the set of all such mappings. Considering who gets fully satisfied in each of these possibilities, the firm’s problem with two customers can be formally stated as:

\[
\begin{align*}
\text{Minimize} & \quad S \\
\text{subject to} & \quad P \{ X_1 + X_2 \leq S \} + P \{ X_1 \leq S, X_2 > S \} + P \left \{ \omega(S), \hat{A}(S, X_1, X_2) = 1 \right \} \geq \beta_1 \quad (SL_1) \\
& \quad P \{ X_1 + X_2 \leq S \} + P \{ X_1 > S, X_2 \leq S \} + P \left \{ \omega(S), \hat{A}(S, X_1, X_2) = 2 \right \} \geq \beta_2 \quad (SL_2)
\end{align*}
\]

where \( \omega(S) \) represents the event that the firm can fully satisfy one of the customers but not both, i.e., \( X_1 \leq S, X_2 \leq S \), and \( X_1 + X_2 > S \). It is only when \( \omega(S) \) occurs that the firm’s choice of allocation policy matters.

We define a linear knapsack policy with parameters \((k_1, k_2)\) and \((t_1, t_2)\), where \( k_i \geq 0 \) and \( t_i \) are scalars, to be the following procedure for allocating inventory between two customers: (1) apply the linear transformation \( \tilde{x}_i = k_i x_i + t_i \) to each of the demand realizations; (2) prioritize customers in increasing order of \( \tilde{x}_i \) and allocate \( S \) accordingly. The \( x_i \)’s can be interpreted as the volume, and \( \tilde{x}_i \)’s as the cost (linear in volume) of a set of items that could potentially be packed in a knapsack with a total volume \( S \); hence the name linear knapsack. Note that the capacity of the knapsack is also a decision variable here.

To assign the first priority to customer 1 (2) if \( x_1 \leq S \) and \( x_2 > S \) (\( x_2 \leq S \) and \( x_1 > S \)), a linear knapsack policy must have a tie for \( x_1 = x_2 = S \), i.e., \( k_1 S + t_1 = k_2 S + t_2 \). This requires the intercepts to be linked in a certain fashion: \( t_2 - t_1 = S (k_1 - k_2) \). Without loss of generality, we set \( k_1 = 1, t_1 = 0, \) and \( t_2 = S (1 - k_2) \). A linear knapsack policy can thus be specified more parsimoniously by one scalar, \( k_2 \), and the linear transformations \( \tilde{x}_1 = x_1 \) and \( \tilde{x}_2 = k_2 x_2 + S (1 - k_2) \). In particular, it gives priority to customer 1 over customer 2 iff \( x_1 < k_2 x_2 \) and \( S (1 - k_2) \), i.e., \( \hat{A}(S, x_1, x_2) = 1 \) iff \( x_1 < k_2 x_2 \).

We are now ready to state our main result concerning the two-customer problem.

**Theorem 6** With two customers, the optimal inventory level is \( S^* \), and the linear knapsack policy with \( k_1 = 1 \) and \( k_2 = k^* \) is an optimal allocation policy. The optimal policy parameters are:

<table>
<thead>
<tr>
<th>Case</th>
<th>( \beta_1 &gt; \alpha_1 ) and ( \beta_2 &gt; \alpha_2 )</th>
<th>( S_0 )</th>
<th>( k_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>( \beta_1 &gt; \alpha_1 ) and ( \beta_2 \leq \alpha_2 )</td>
<td>( F_1^{-1}(\beta_1) )</td>
<td>0</td>
</tr>
<tr>
<td>Case 2</td>
<td>( \beta_1 \leq \alpha_1 ) and ( \beta_2 &gt; \alpha_2 )</td>
<td>( F_2^{-1}(\beta_2) )</td>
<td>( \infty )</td>
</tr>
</tbody>
</table>
with $S_0$ and $k_0$ uniquely determined by two implicit expressions:

$$P \{X_1 + X_2 \leq S_0\} = \beta_1 + \beta_2 - 1 + P \{X_1 > S_0, X_2 > S_0\}$$

$$P \{\omega(S_0), X_1 < k_0 X_2 + S_0 (1 - k_0)\} = \beta_1 - P \{X_1 + X_2 \leq S_0\} - P \{X_1 \leq S_0, X_2 > S_0\}$$

and the threshold service levels, $\alpha_1$ and $\alpha_2$, defined as:

$$\alpha_1 \equiv P \{X_1 + X_2 \leq S_0\} + P \{X_1 \leq S_0, X_2 > S_0\}$$

$$\alpha_2 \equiv P \{X_1 + X_2 \leq S_0\} + P \{X_1 > S_0, X_2 \leq S_0\}$$

Case 1 represents the mainstream situation without free-riders, whereas in Cases 2 and 3, one of the customers (customer 2 and 1, respectively) is able to free ride in the sense that he is satisfied even if he never gets priority in the event of $\omega(S_0)$. These two are extreme cases, where setting the inventory level as if there was only one customer is optimal. The threshold service levels $\alpha_i$, the probability that customer $i$ faces no contest from the other customer at inventory level $S_0$, let us precisely specify when a customer’s required service level is low enough to qualify him as a free-rider. Note that the customers cannot both be free-riders.

Cases 1-3 are mutually exclusive, and also exhaustive for all practical purposes. There remain two other possibilities: $\{\beta_1 \leq \alpha_1$ and $\beta_2 \leq \alpha_2$ with at least one inequality strict$\}$ cannot happen, because $\beta_1 + \beta_2 = \alpha_1 + \alpha_2 + P \{\omega(S_0)\}$ by definition; $\{\beta_1 = \alpha_1$ and $\beta_2 = \alpha_2\}$ is a pathological case with $P \{\omega(S_0)\} = 0$, which makes all three solutions equivalent and optimal. (We ignore the latter for ease of exposition.)

Building on Theorem 6, we now establish for the general problem with any number of customers that there is always some benefit to pooling. Let $S_i = F_i^{-1}(\beta_i)$ be dedicated inventory levels in the absence of pooling.

**Theorem 7** The pooling benefit is always strictly positive, i.e., $S^* < S_1 + S_2 + \cdots + S_N$.

The proof first uses Theorem 6 to show that the pooling benefit is always strictly positive in the two-customer case. It then rests on the following observation: Theorem 6 can be used to develop upper bounds for the general problem. Suppose that the firm pairs customers and solves the ordering and allocation problems for each pair in isolation. The sum of inventory levels obtained for pairs, plus dedicated inventories for non-paired customers (if any), would be an upper bound on the globally optimal inventory level.

While it is commonly known that the pooling benefit vanishes as correlation approaches +1 in newsvendor models (Eppen 1979), Theorem 7 holds for perfectly positive correlation also. The intuitive reason is that the responsive policy is able to respond to variations in demands efficiently. To see this in a concrete example, take two customers, let demands be perfectly positively correlated $P\{X_1 = X_2\} = 1$, and assume symmetric service level requirements $\beta_1 = \beta_2 = \beta$. When demand is moderately high ($S/2 < X_1 \leq S$) the responsive policy can satisfy one of the customers fully, whereas the comparable no-pooling policy with the total inventory $S$
divided into two dedicated piles of size $S/2$ would not be able to fully satisfy any of the customers. So, with the same amount of inventory, no-pooling always achieves less in terms of service. This symmetric two-customer example makes the argument especially transparent, but a similar dynamic drives the result in the asymmetric case also. In fact, the pooling benefit is generally larger in problem instances with asymmetry in demand distributions and/or service level requirements. In closing, we note that there are resource flexibility and component commonality models that show pooling benefit under perfectly positive correlation (Van Mieghem 1998, Van Mieghem 2004). Their rationale is distinct from ours as it rests on some form of asymmetry, e.g., differences in profitability between products.

6 Anticipative Priority Policies

Anticipative priority policies ignore demand realizations when making up the priority list. This operational simplicity may come at the expense of carrying higher inventory. Anticipative policies are still worthwhile to analyze, as they are observed in industry often, especially the fixed list policies.

6.1 Fixed List Policies

A fixed list policy is the simplest allocation policy to design and operate. Customers are put in a fixed priority list, and their demands are filled from the pool of inventory one after the other in the order dictated by the list until there is no more stock left, or until all the demands are completely filled.

For a given set of demand distributions and service levels, each of the $N!$ distinct priority lists is associated with a distinct inventory level. The following result identifies those inventory levels and finds the optimal inventory level with its corresponding optimal fixed list policy.

**Theorem 8** The optimal fixed list policy ranks the customers in decreasing order of their required service levels. Relabel customers such that $\beta_1 \geq \cdots \geq \beta_N$. Set $\pi(k) = k$ for all $k \in \{1, \ldots, N\}$. The optimal priority list is $\Pi^*_{\text{fixed list}} = (1, \ldots, N)$. The optimal inventory level is $S^*_{\text{fixed list}} = \max \{G_1^{-1}(\beta_1), \ldots, G_N^{-1}(\beta_N)\}$, where $G_k$ is the distribution function of $X_1 + \cdots + X_k$ for $k \in \{1, \ldots, N\}$.

It is surprising that the highest-service-level-first rule is optimal without any conditions on demand distributions. For instance, whether the highest-service-level customer has a low or high demand on average compared with the other customers, it is optimal to give that customer top priority in allocation. This is true even when customer demands are correlated. Hence, a fixed list policy may be the policy of choice in practice, especially when distributional information about customer demands is lacking.
Despite their popularity in practice, fixed list policies do not necessarily guarantee a positive pooling benefit. We show this by counterexample in §7; pooling benefit can be strictly negative for optimal fixed list policies. For customer demands with multivariate normal distribution, however, the optimal fixed list policy does ensure a non-negative pooling benefit. Recall that \( S_i = F_i^{-1}(\beta_i) \).

**Theorem 9** The optimal fixed list policy yields a positive pooling benefit, i.e., \( S^*_\text{list} \leq S_1 + \cdots + S_N \), if the demands \( (X_1, \ldots, X_N) \) follow an arbitrary multivariate normal distribution with means \( (\mu_1, \ldots, \mu_N) \), standard deviations \( (\sigma_1, \ldots, \sigma_N) \), and correlation coefficients \( \rho_{ij} \in [0,1] \) between the demands of customers \( i \) and \( j \).

### 6.2 Randomized List Policies

A randomized list policy involves a randomization step to generate the priority list, which can be specified by a set of \( N! \) positive fractional weights placed on all possible priority lists (\( N! \) permutations of \( N \) customers) that sum to unity. In this section, we show how to compute the optimal randomized list policy. The case of iid demands is easier to solve, so we analyze it first and then move on to arbitrary demand distributions.

Consider iid demand random variables \( X_1, \ldots, X_N \). Let the distribution function of the sum of any \( n \) of these random variables be \( G_n(\cdot) \). Let the column vectors \( (G_1(S), \ldots, G_N(S))^T \) and \( (\beta_1, \ldots, \beta_N)^T \) be denoted by \( C(S) \) and \( B \), respectively. Let \( w_{ij} \) be the probability that customer \( i \) is assigned priority position \( j \), and let \( W \) be the \( N \times N \) matrix with \( w_{ij} \) in row \( i \) and column \( j \). Note that \( W \) is a doubly stochastic matrix, and by Birkhoff’s Theorem it can be written as a convex combination of \( N \times N \) permutation matrices (Marshall and Olkin 1979, p. 19). Hence, \( W \) constitutes a randomized list policy; the permutation matrices and the positive fractional weights summing to 1, which make up the convex combination, determine the priority list.

**Theorem 10** Suppose the demands are iid random variables. (a) The optimal randomized list policy can be found by solving the following problem: minimize \( S \) subject to \( W \cdot C(S) \geq B \), where \( S \) and the elements of the matrix \( W \) are the decision variables. (b) The unique solution \( S_c \) of the equation \( \sum_{n=1}^{N} G_n(S) = \sum_{i=1}^{N} \beta_i \) is a lower bound for the optimal stock. (c) All the service levels are exactly satisfied if and only if \( C(S_c) \) majorizes \( B \). The optimal inventory in this case is precisely \( S_c \).

Solving the optimization problem in (a) and the equation in (b) are both easy because \( G_n(S) \) are monotone increasing in \( S \). Further, for a fixed value of \( S \), the mathematical program in (a) is an LP. Also note that the solution to an equation like \( \sum_{n=1}^{N} G_n(S) = \sum_{i=1}^{N} \beta_i \) can be estimated using Monte Carlo simulation software. We have found that problem instances with service levels upwards of 70% almost invariably have exact solutions. Once the optimal stock has been found, finding the optimal doubly stochastic matrix, and hence the optimal allocation policy parameters, is a matter of solving linear equations.
When demands are not iid, the problem is significantly more complex. We outline a solution procedure to handle this case. Let $P(\pi_k)$ denote a discrete probability distribution over all $N!$ priority lists $\Pi_k = (\pi_k(1), \ldots, \pi_k(N))$ for $k = 1, \ldots, N!$, such that $\sum_{k=1}^{N!} P(\pi_k) = 1$.

**Step 1.** Compute the optimal fixed-list inventory level (from Theorem 8), which serves as an upper bound on the optimal randomized-list inventory level (by Theorem 2): $S_{\text{f-list}}^* = \max\{G_1^{-1}(\beta_1), \ldots, G_N^{-1}(\beta_N)\}$. A lower bound is $\max\{F_1^{-1}(\beta_1), \ldots, F_N^{-1}(\beta_N)\}$.

**Step 2.** Set $S = S_{\text{f-list}}^*$ and attempt to find a probability distribution $P(\cdot)$ over all possible priority lists so that the following inequality is satisfied for all customers $i = 1, \ldots, N$ (let $I \{\cdot\}$ denote the indicator function):

$$\sum_{j=1}^{N} \sum_{k=1}^{N!} I \{\pi_k(j) = i\} P(\pi_k) F_{\pi_k(1), \ldots, \pi_k(N)}(S) \geq \beta_i$$

**Step 3.** Perform a binary search for the smallest feasible $S$ between the upper and lower bounds computed in Step 1, repeating Step 2 as many times as needed, and stopping when we reach an $S$ for which the system of linear inequalities has no solution. The last feasible inventory level is optimal.

The procedure converges because we employ binary search, or interval bisection, between finite upper and lower bounds to find the optimal stock level. The bisection is guaranteed to converge, because $\sum_{j=1}^{N} \sum_{k=1}^{N!} I \{\pi_k(j) = i\} P(\pi_k) F_{\pi_k(1), \ldots, \pi_k(N)}(S)$ is a continuous and monotone function of $S$. Although the binary search itself is logarithmic, the algorithm is exponential time - $O(c^N)$ - because we need to solve for $N!$ variables in Step 2 of the algorithm.

### 7 Numerical Comparisons of Allocation Policies

In this section, we present numerical examples with the express purpose of making comparisons between the allocation policy classes analyzed in §5 and §6. Figures 1 and 2 report the optimal inventory levels that correspond to the no-pooling, fixed list, randomized list, and responsive policies when there are three customers with iid demands and differentiated service levels. The underlying demand distributions are normal with a CV (coefficient of variation = standard deviation / mean) of 0.2 and 0.3, and lognormal with a CV of 0.5 and 1. We vary $\bar{\beta}$ from 75% to 95% in increments of 5%, and introduce higher service level differentiation by starting with a uniform set of $\beta$’s and simultaneously reducing $\beta_1$ and increasing $\beta_3$ by 2.5% at a time.

The tables reveal an interesting insight. When demands are relatively stable with CV 0.3 or lower, there is virtually no difference between the inventory levels prescribed by the optimal responsive policy and the optimal randomized list policy. When demands are highly variable, however, the differences widen significantly, and the gulf is particularly marked for lower service levels. These observations suggest that when demands are modeled with stable distributions such as normal, it is sufficient to restrict the search to fixed and randomized
<table>
<thead>
<tr>
<th>Demand Distribution</th>
<th>N (10,2)</th>
<th>Optimal Inventory Pooling Benefit</th>
<th>Fixed Policy</th>
<th>Rand. Policy</th>
<th>Responsive Policy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Beta 1</td>
<td>Beta 2</td>
<td>Beta 3</td>
<td>Sum</td>
<td>No Pooling</td>
</tr>
<tr>
<td>75.0% 75.0% 75.0%</td>
<td>225%</td>
<td>34.05</td>
<td>32.35</td>
<td>27.69</td>
<td>27.66</td>
</tr>
<tr>
<td>72.5% 75.0% 77.5%</td>
<td>225%</td>
<td>34.06</td>
<td>32.10</td>
<td>27.69</td>
<td>27.66</td>
</tr>
<tr>
<td>70.0% 75.0% 80.0%</td>
<td>225%</td>
<td>34.08</td>
<td>31.82</td>
<td>27.69</td>
<td>27.66</td>
</tr>
<tr>
<td>67.5% 75.0% 82.5%</td>
<td>225%</td>
<td>34.13</td>
<td>31.60</td>
<td>27.69</td>
<td>27.66</td>
</tr>
<tr>
<td>65.0% 75.0% 85.0%</td>
<td>225%</td>
<td>34.19</td>
<td>31.35</td>
<td>27.69</td>
<td>27.66</td>
</tr>
<tr>
<td>80.0% 80.0% 80.0%</td>
<td>240%</td>
<td>35.05</td>
<td>32.92</td>
<td>29.13</td>
<td>29.13</td>
</tr>
<tr>
<td>77.5% 80.0% 82.5%</td>
<td>240%</td>
<td>35.06</td>
<td>32.60</td>
<td>29.13</td>
<td>29.13</td>
</tr>
<tr>
<td>75.0% 80.0% 85.0%</td>
<td>240%</td>
<td>35.11</td>
<td>32.35</td>
<td>29.13</td>
<td>29.13</td>
</tr>
<tr>
<td>72.5% 80.0% 87.5%</td>
<td>240%</td>
<td>35.12</td>
<td>32.10</td>
<td>29.13</td>
<td>29.13</td>
</tr>
<tr>
<td>70.0% 80.0% 90.0%</td>
<td>240%</td>
<td>35.30</td>
<td>31.82</td>
<td>29.13</td>
<td>29.13</td>
</tr>
<tr>
<td>80.0% 80.0% 80.0%</td>
<td>240%</td>
<td>35.05</td>
<td>32.92</td>
<td>29.13</td>
<td>29.13</td>
</tr>
<tr>
<td>77.5% 80.0% 82.5%</td>
<td>240%</td>
<td>35.06</td>
<td>32.60</td>
<td>29.13</td>
<td>29.13</td>
</tr>
<tr>
<td>75.0% 80.0% 85.0%</td>
<td>240%</td>
<td>35.11</td>
<td>32.35</td>
<td>29.13</td>
<td>29.13</td>
</tr>
<tr>
<td>72.5% 80.0% 87.5%</td>
<td>240%</td>
<td>35.12</td>
<td>32.10</td>
<td>29.13</td>
<td>29.13</td>
</tr>
</tbody>
</table>

Figure 1: The optimal inventory level and pooling benefit (% reduction in inventory due to pooling) in problem instances with N = 3, iid normal demands, and differentiated service levels.

list policies. However, in the case of highly unstable or long-tailed distributions, responsive policies may reap significant inventory savings over fixed and randomized list policies.

We also observe that higher service differentiation consistently leads to higher pooling benefit for all three policy classes. This is driven by the fact that the no-pooling inventory level is more sensitive to the maximum of the required service levels than the optimal pooling solution in any policy class. The optimal fixed list policy solution is driven by the minimum, whereas the optimal randomized list and responsive policy solutions are driven by the average of the required service levels. Especially within the latter two policy classes, the use of a single pool of inventory better absorbs the stochastic highs and lows in customer demands and allows for a response to typical rather than to extreme events. Note that the no-pooling solution requires more inventory
with higher differentiation; when inventory is pooled, the optimal fixed list policy requires less, whereas the optimal randomized list and responsive policies require the same amount. A caveat is in order: non-iid demands can alter these observations, because in that case how demands differ (e.g., which customer’s mean demand is higher) may interact with the nature of service differentiation in unpredictable ways. Contrary to our basic finding, Deshpande et al. (2003) observe that higher service level differentiation leads to a smaller inventory pooling benefit (Table 2, p. 696).

Finally, note that an optimal fixed list policy can suffer from a strictly negative pooling benefit; it can require more inventory than no-pooling. Our numerical experiments (not reported here) suggest that this happens for highly variable demands and relatively low service levels. For example, if three customers had iid lognormal demands with mean 10 and standard deviation 15, and $\beta_1 = \beta_2 = \beta_3 = 75\%$, the pooling benefit from the optimal fixed list policy would be $-6.33\%$.

8 Concluding Remarks

This paper was motivated by our observations of industry practice in the field of supply chain planning for after-market service operations. Firms often have agreements with their clients containing explicit service level clauses for delivering parts to support products. In this paper, we develop solutions that simultaneously determine the replenishment quantity and the priority rule for optimally allocating inventory to customers demanding different service levels.

We earmark three fundamental classes of allocation policies: fixed list, randomized list, and responsive. We obtain complete solutions for fixed and randomized list policies, and partial solutions for responsive policies in the form of bounds and solutions for special cases. We uncover a subclass of responsive policies, called linear knapsack, that is optimal generally for the two-customer case. We show for any number of customers that the pooling benefit is always strictly positive, even when demands are perfectly positively correlated. We find that when demands are independent random variables with low to moderate coefficients of variation, there is virtually no difference between the inventory levels prescribed by the optimal responsive policy and the optimal randomized list policy. Whereas, if demands are highly variable, the differences between the optimal prescriptions of the three policy classes become significant, which is when responsive policies are most helpful.

It is immediate that our single period solutions extend to a periodic-review infinite-horizon model with service defined as the long-run fraction of periods in which a customer’s demand is fully satisfied from stock, provided lead time is zero. In that simple multiperiod scenario, each period is effectively decoupled from the next, and the optimal one-period solution can be implemented in every period without any loss of optimality. An extension to a multiperiod model appears to be much more challenging if any of the following features
Figure 2: The optimal inventory level and pooling benefit (% reduction in inventory due to pooling) in problem instances with $N = 3$, iid lognormal demands, and differentiated service levels.

are incorporated into the model: significant supply or demand lead time, different frequencies for inventory replenishment and demand batching, finite planning horizon, and non-stationary allocation policies. We mark out such extensions as worthy problems for future research.

References


*Management Sci.* 43(1) 40-53.


Online Companion for
“Inventory Pooling to Deliver Differentiated Service”

Appendix A: Proofs

Proof of Theorem 1

Suppose $A_0$ is an optimal allocation policy and it does not belong to the class of priority policies. Then, there must exist demand realizations for which at least two customers, say $i_1$ and $i_2$ (the identity of such customers may vary from one realization to another), are partially satisfied, i.e., $0 < y_{ik} < x_{ik}$ for $k = 1, 2$. For those demand realizations, reallocating inventory from customer $i_2$ to customer $i_1$ so that $y_{i_1} \leftarrow \min\{y_{i_1} + y_{i_2}, x_{i_1}\}$ and $y_{i_2} \leftarrow \max\{y_{i_1} + y_{i_2} - x_{i_1}, 0\}$ is no worse than $A_0$, because it results in a feasible allocation policy that uses the same amount of inventory, and it may even fully satisfy customer $i_1$. The reason why feasibility is preserved is as follows. Before reallocation, since $i_1$ and $i_2$ are partially satisfied in that demand realization, its probability has zero contribution to the service level enjoyed by customers $i_1$ and $i_2$. After reallocation, there are two possibilities for each demand realization: (1) it still has zero contribution because $i_1$ is still partially satisfied; (2) it has a strictly positive contribution because $i_1$ can now be fully satisfied. As a result, the service level enjoyed by customer $i_1$ stays the same or strictly increases, whereas the service level enjoyed by customer $i_2$ stays the same. (Note that the policy remains intact for the rest of the realizations.)

Continuing with identifying such realizations and pairs of customers, and reallocating inventory in this fashion until it is no longer possible, one either finds a strictly better policy (a policy that satisfies all service level requirements using strictly less inventory than $A_0$, i.e., if it happens that all customers enjoy strictly higher service levels after all the reallocations are done, then the inventory level can be reduced without violating feasibility) or a policy that is no worse than $A_0$ (a policy that satisfies all service level requirements and uses the same amount of inventory as $A_0$). The former is a contradiction, whereas the latter indicates that the resulting priority policy and $A_0$ are alternative optima.

Proof of Theorem 2

The class of randomized list policies includes the class of fixed list policies; one could randomize with a probability distribution that concentrates all the probability at one priority list. Therefore, the optimal randomized list policy uses no more inventory than the optimal fixed list policy. Similarly, the class of responsive policies includes the class of randomized list policies since we may map each possible $N$-vector of demand realizations into a priority list generated randomly and independently of demands.
Proof of Theorem 3

Due to Theorem 1 it is sufficient to work with priority lists. Rewriting the service level constraints in terms of priority lists $\Pi = (\pi(1), \ldots, \pi(N))$, we obtain the following inequality. For customer $i$ to be fully satisfied when she occupies position $n$ in the priority list, the sum of her demand and the demands of all customers that come before her in the priority list must be less than the inventory level $S$.

$$\sum_{n=1}^{N} P \left\{ \pi(n) = i, \sum_{k=1}^{n} X_{\pi(k)} \leq S \right\} \geq \beta_i$$

In general $\pi(\cdot)$ itself is allowed to depend on demand realizations; not to clutter the notation, we suppress that dependence. Adding both sides of these $N$ inequalities we obtain a necessary condition for $S$ to be feasible:

$$\sum_{n=1}^{N} P \{ \sum_{k=1}^{n} X_{\pi(k)} \leq S \} \geq \sum_{i=1}^{N} \beta_i,$$

where the left-hand side follows from the law of total probability.

If the firm applied GP, giving priority to smaller demands first, the left-hand side would equal $\sum_{n=1}^{N} H_n(S)$.

Proof of Theorem 4

Suppose the firm uses GP with inventory level $S$. For any customer $i$, and for any priority list position $n$, the probability that customer $i$ occupies position $n$ is $\frac{1}{N}$, and the service level that she enjoys in that position is $P\{X_{[1]} + \cdots + X_{[n]} \leq S\}$. So, the service level constraint for each and every customer is $\frac{1}{N}\sum_{n=1}^{N} P\{X_{[1]} + \cdots + X_{[n]} \leq S\} = \frac{1}{N}\sum_{n=1}^{N} H_n(S) \geq \beta$. The smallest feasible inventory level under GP then is $S_{GP}(\beta)$, the unique solution to the equation $\sum_{n=1}^{N} H_n(S) = N\beta$, which is also the equation that defines the universal lower bound $\underline{S}$ applied to this special case with iid demands and undifferentiated service levels (Theorem 3).

Optimality follows from the fact that the universal lower bound $\underline{S}$ is feasible.

Proof of Theorem 5a

We proceed by induction. When $N = 1$, the formula for optimal stock is trivially true. Let $n$ be an arbitrary positive integer. Suppose there exists a CGP policy for any set of $n$ customers that satisfies their required service levels exactly. Then we shall show that there exists a CGP policy for any set of $n + 1$ customers that satisfies their required service levels exactly. The theorem will then follow immediately.

We denote the customers by $c_1, c_2, \ldots, c_{n+1}$ and let their required service levels be ordered as $\beta_1 > \beta_2 > \cdots > \beta_{n+1}$ (we assume no ties for ease of exposition, this is without loss of generality). Let $S_{GP}(\beta_m)$ be
the stock level required by GP to deliver a service level of exactly $\beta_m \equiv \frac{\beta_1 + \cdots + \beta_m}{m}$ to each of $c_1, \ldots, c_m$ for $m = 1, \ldots, N$. By induction hypothesis there exists a CGP policy (call it $\Pi_0$) with stock level $S_{GP}(\beta_m)$ for $c_1, \ldots, c_m$. Note that every optimal allocation policy must necessarily move on to the next customer in the list if the current customer has a demand realization that exceeds the available stock.

**Step 1:** Escalate the stock to $S_{GP}(\beta_{n+1})$. Consider an allocation policy that uses this stock level and makes allocations as follows: we first let $c_{n+1}$ be served last and the other $n$ customers be served in accordance with $\Pi_0$. Provided this policy is CGP for the set of $n+1$ customers for the given demand realizations; otherwise we allocate stock according to GP. Note that this allocation policy is CGP by definition. Let $\hat{\beta}_i$ be the service level actually delivered to $c_i$ by this policy (we shall use this notation for actual service levels delivered by an allocation policy). Then we have $\hat{\beta}_i > \beta_i$ for $i = 1, \ldots, n$, and - assuming that customer $n+1$ is not a free-rider - $\hat{\beta}_{n+1} > \beta_{n+1}$.

**Step 2:** Consider maintaining $\hat{\beta}_i (i = 1, \ldots, n - 1)$ and modifying $\hat{\beta}_n$ and $\hat{\beta}_{n+1}$. By interchanging the roles of $c_n$ and $c_{n+1}$ in the allocation policy $\Pi_0$, which is possible only because the demands are iid, we see that there exist two distinct CGP policies (operating with a stock of $S_{GP}(\beta_{n+1})$) such that $\hat{\beta}_{n+1} < \beta_{n+1}$ in one policy (that is, $\Pi_0$) while $\hat{\beta}_{n+1} > \beta_{n+1}$ in another. Note that $\hat{\beta}_1, \ldots, \hat{\beta}_{n-1}$ take the same values in both policies.

**Step 3:** We shall deduce that there exists a CGP policy (call it $\Pi_1$) such that $\hat{\beta}_{n+1} = \beta_{n+1}, \hat{\beta}_i > \beta_i$ ($i = 1, \ldots, n - 1$) and $\hat{\beta}_n < \beta_n$. We modify allocation policy $\Pi_0$ as follows. At each point in the demand space of $c_1, \ldots, c_{n+1}$, we allocate exactly the same quantities to customers $c_1, \ldots, c_{n-1}$ as $\Pi_0$. We make changes only in the quantities allocated to $c_n$ and $c_{n+1}$ at certain points in the space. Consider the set $\Omega$ of all points at which either $c_n$ or $c_{n+1}$ but not both can be fulfilled, given the allocations of the other customers. In $\Pi_0$ customer $c_{n+1}$ was by design denied all these potential opportunities for complete fulfillment in the point set $\Omega$. Pick a small positive number $\epsilon$ and partition $\Omega$ into $n+1$-dimensional balls each of Lebesgue measure $\epsilon$. Then substitute $c_{n+1}$ for $c_n$ one by one in each $\epsilon$-ball until their assignments in these regions are completely reversed. The assumption of absolutely continuous demand distributions implies that consequent increase in the service level accorded to $c_{n+1}$ is a continuous function of $\epsilon$. Note that since the perturbations preserve the total number of customers serviced at every point in demand space, the perturbed allocation policies remain CGP. By choosing $\epsilon$ to be sufficiently small, we obtain an allocation policy in which $c_{n+1}$ exactly attains his required service level. This proves the existence of policy $\Pi_1$ that was claimed in the first sentence of Step 3.

**Step 4:** Implementing $\Pi_1$ results in $\hat{\beta}_i > \beta_i$ for $i = 1, \ldots, n - 1, \hat{\beta}_n < \beta_n$, and $\hat{\beta}_{n+1} = \beta_{n+1}$. We modify $\Pi_1$ by leaving its allocations for $c_1, \ldots, c_{n-2}$ and $c_{n+1}$ completely unchanged but modifying its allocations for $c_n$ and $c_{n-1}$. Consider the set $\Omega$ of all points at which either $c_n$ or $c_{n-1}$ but not both can be fulfilled, given the allocations of the other customers as per $\Pi_1$, but only $c_{n-1}$ is in fact fulfilled. At each realization, we first scan all the customers except $c_n$ and $c_{n-1}$ and either put them into the bucket of fulfilled customers or not as
per policy $\Pi_1$ (which is well defined at the end of Step 3). Then we note whether or not $c_n$ and $c_{n-1}$ are put into the bucket. Suppose at every point in demand space, we make the following change and no other change: replace $c_{n-1}$ by $c_n$ whenever there is room for each of these two individually but no room for both left in the bucket. Then we will end up giving $c_n$ a higher service level than the service level given to $c_{n-1}$ by policy $\Pi_1$, which is in turn higher than $\beta_{n-1}$ and therefore higher than $\beta_n$. So we make this interchange only over as much of demand space as needed to pull up the service level attained by $c_n$ to exactly $\beta_n$. This implies that there exists a CGP policy (call it $\Pi_2$) such that implementing it results in $\hat{\beta}_i > \beta_i$ for $i = 1, \ldots, n-2$, $\hat{\beta}_{n-1} < \beta_{n-1}$, and $\hat{\beta}_i = \beta_i$ for $i = n, n+1$. Repeating this step $(n-2)$ more times — the total number of loops being $n$ — we arrive at a CGP policy that results in $\hat{\beta}_i = \beta_i$ ($i = 1, \ldots, n+1$). This proves the theorem.

**Proof of Theorem 5b**

Suppose we have $N$ customers $c_1, \ldots, c_N$. Fix an arbitrary positive integer $n < N$. Our induction hypothesis is that there exist finite strictly positive scale factors $K_{ij}$ ($j = 1, \ldots, n$) such that the allocation policy described in the theorem (which we shall refer to as $\Pi^*$), operating with a stock of $S_{GP}(\beta_N)$, satisfies the service levels of a subset $\Gamma$ of $n$ customers exactly (for any choice of scale factors for the customers not in $\Gamma$). We shall prove that there exists a positive number $K_{i_0}$ such that when $K_{i_0}$ is used to scale the demand of a customer $i_0 \in \{1, \ldots, N\} - \Gamma$ then $\Pi^*$ satisfies the service levels of the customers in $\Gamma$ as well as that of $c_{i_0}$.

**Step 1:** Fix the set of $n$ optimal scale factors for the subset of customers $\Gamma$ corresponding to an arbitrary set of scale factors for the customers not in $\Gamma$. Then $S_{GP}(\beta_N)$ together with the given scale factors $K_i$ over-serves at least one customer (say $c_1$) and under-serves at least one customer (say $c_J$) because it is a CGP policy and therefore the sum of the service levels attained under it is $\sum_{i=1}^{N} \beta_i$. Assume without loss of generality that $\beta_1 > \beta_J$. So we have $\hat{\beta}_1 > \beta_1 > \beta_J > \hat{\beta}_J$.

**Step 2:** Pick a small positive number $\epsilon$. Replace $K_I$ by $K_I + \epsilon$ and $K_J$ by $K_J - \epsilon$. Let the scale factors of the remaining customers not in $\Gamma$ (if any) remain unchanged. Then invoke the induction hypothesis again to obtain a new set of optimal scale factors for the customers in $\Gamma$ so that their service levels are exactly satisfied.

We note that since the demands are iid, $\hat{\beta}_J$ decreases monotonically (from its value before modifying $K_I$) and $\hat{\beta}_J$ increases monotonically.

**Step 3:** Repeat Step 2. After finitely many iterations, $K_I$ will take the value $K_J$ and $K_J$ will take the value $K_I$ (to ensure that an exact interchange takes place, we need only take $\epsilon$ to be of the form $\frac{K_I - K_J}{M}$ for some arbitrary positive integer $M$). But this implies that $\hat{\beta}_I$ and $\hat{\beta}_J$ will have interchanged their values too, and it follows from Step 1 that $\hat{\beta}_I$ will have crossed $\beta_I$ from above. By a suitable choice of $M$ we can ensure that at some step in the procedure, $\hat{\beta}_I = \beta_I$. This proves the induction step. The base case for the induction is easily established for $n = 1$. 


Proof of Theorem 6

Service level constraints \((SL_1)\) and \((SL_2)\) imply:

\[
2P\{X_1 + X_2 \leq S\} + P\{X_1 \leq S, X_2 > S\} + P\{X_1 > S, X_2 \leq S\} + P\{\omega(S)\} \geq \beta_1 + \beta_2
\]

This inequality must hold regardless of the specific responsive policy used. Because the demands are positive-valued, it is equivalent to: \(P\{X_1 + X_2 \leq S\} + 1 - P\{X_1 > S, X_2 > S\} \geq \beta_1 + \beta_2\). Observe that the left-hand-side is strictly increasing in \(S\), equal to 0 at \(S = 0\), and goes to 2 in the limit as \(S\) approaches \(+\infty\). Hence, there exists a unique \(S\) that satisfies it with equality, which we define in (1) as \(S_0\). Any \(S\) lower than \(S_0\) would violate one or both of the service level constraints. Therefore, \(S_0\) must be optimal as long as we can find a feasible allocation policy that satisfies the service level constraints:

\[
P\left\{\omega(S_0), \hat{A}(S, X_1, X_2) = 1\right\} \geq \beta_1 - \alpha_1
\]

\[
P\left\{\omega(S_0), \hat{A}(S, X_1, X_2) = 2\right\} \geq \beta_2 - \alpha_2
\]

(2)

where \(\alpha_1 \equiv P\{X_1 + X_2 \leq S_0\} + P\{X_1 \leq S_0, X_2 > S_0\}\) and \(\alpha_2 \equiv P\{X_1 + X_2 \leq S_0\} + P\{X_1 > S_0, X_2 \leq S_0\}\).

Since any portion of \(P\{\omega(S_0)\}\) can be assigned to customer 1 or 2 by a suitable choice of \(k_2 \in (0, \infty)\), \(S_0\) is indeed feasible and therefore optimal, when both right-hand-sides in (2) are strictly positive (Case 1). In this case there exists a unique feasible \(k_2\), denoted by \(k_0\), at which both service level constraints are binding, i.e., satisfying either service level guarantee exactly gives us a characterization of the optimal \(k_2\). When either one of the right-hand-sides in (2) is negative (both cannot be strictly negative, as their sum is equal to \(P\{\omega(S_0)\}\)), the firm is sure to satisfy the corresponding customer with an inventory level of \(S_0\) or higher. For instance, if \(\beta_2 - \alpha_2 \leq 0\) (and \(\beta_1 - \alpha_1 > 0\)), customer 2 is a free-rider, already assured a service level of \(\beta_2\) or better even if he never gets the priority in the event of \(\omega(S_0)\). Therefore, in this case (Case 2) it is optimal to set \(k_2 = 0\) in order to always give priority to customer 1, and the optimal inventory level (higher than \(S_0\) corresponds to the minimum \(S\) that makes \((SL_1)\) binding: \(P\{X_1 + X_2 \leq S^*\} + P\{X_1 \leq S^*, X_2 > S^*\} + P\{\omega(S^*)\} = P\{X_1 \leq S^*\} = \beta_1\). The optimal solution can be derived similarly when \(\beta_1 - \alpha_1 \leq 0\) (and \(\beta_2 - \alpha_2 > 0\)), which constitutes the other extreme case (Case 3) where customer 1 is a free-rider and the firm must always give priority to customer 2 by setting \(k_2\) at a sufficiently large number.

Proof of Theorem 7

Take any problem instance with \(N = 2\). In Cases 2 and 3 of Theorem 6, it is immediate that \(S^* < S_1 + S_2\). In Case 1, the optimal inventory \(S_0\) must satisfy: \(S_1 < S_0\) and \(S_2 < S_0\). To see this, suppose the opposite is true: \(S_0 \leq S_i\) for some customer \(i\). First, it must be that \(S_0 = S_i\), because lower inventory levels are definitely
infeasible for customer \(i\) irrespective of how allocation is done. Moreover, the only allocation policy feasible for customer \(i\) is to always give priority to him over customer \(-i\). But then, for this solution to be feasible for customer \(-i\), he must be a free-rider, i.e., \(\beta_{-i} \leq \alpha_{-i}\), which contradicts the definition of Case 1. In view of \(\beta_1, \beta_2 \in (0, 1)\), our observation that \(S_1 < S_0\) and \(S_2 < S_0\) in Case 1 implies:

\[
P\{X_1 \leq S_1 \text{ or } X_2 \leq S_2\} < P\{X_1 \leq S_0 \text{ or } X_2 \leq S_0\}
\]

Now, because \(\{X_1 \leq S_1 \text{ and } X_2 \leq S_2\}\) implies \(\{X_1 + X_2 \leq S_1 + S_2\}\) but not vice versa, \(P\{X_1 \leq S_1 \text{ and } X_2 \leq S_2\} \leq P\{X_1 + X_2 \leq S_1 + S_2\}\). Using this and the inequality above, we find that:

\[
P\{X_1 \leq S_1\} + P\{X_2 \leq S_2\} - P\{X_1 + X_2 \leq S_1 + S_2\} < 1 - P\{X_1 > S_0 \text{ and } X_2 > S_0\}
\]

Rearranging the terms, replacing \(P\{X_i \leq S_i\}\) with \(\beta_i\), and using the definition of \(S_0\) in (1), we obtain

\[
P\{X_1 + X_2 \leq S_0\} < P\{X_1 + X_2 \leq S_1 + S_2\},\text{ which implies that } S^* < S_1 + S_2.
\]

Now take any problem instance with \(N \geq 3\) customers and pair any two of the customers. Treat these two customers with their own pool of inventory to be allocated between the two of them by a linear knapsack policy (optimized using Theorem 6), and reserve dedicated inventory for all the remaining customers. Let \(\bar{S}\) be the resulting total inventory. Because it is accompanied with a feasible allocation policy, \(\bar{S}\) is an upper bound for the \(N\)-customer problem. Thus, \(S^* \leq \bar{S} < S_1 + S_2 + \cdots + S_N\), where the latter inequality comes from the fact that the pooling benefit is strictly positive for any two-customer problem (shown above).

**Proof of Theorem 8**

We claim that any list that is not in decreasing order of required service levels can be improved by an adjacent pairwise interchange. Take an arbitrary permutation of customers \(\Pi\), and for ease of exposition relabel all the customers such that \(\Pi = (1, 2, \ldots, N)\). Suppose \(\beta_k < \beta_{k+1}\) for some \(k \in \{1, \ldots, N-1\}\). Denote by \(\Pi'\) the permutation in which the positions of customers \(k\) and \(k + 1\) are interchanged in \(\Pi\), all the other positions remaining unchanged. Let \(\Gamma_1(x)\) denote the distribution function of \(\sum_{i=1}^{k} X_i\), \(\Gamma_2(x)\) the distribution function of \(\sum_{i=1}^{k+1} X_i\), and \(\Gamma_3(x)\) the distribution function of \(X_{k+1} + \sum_{i=1}^{k-1} X_i\). Define \(\tilde{S} = \max_{n=1,\ldots,k-1,k+2,\ldots,N} G_n(\beta_n)\), where \(G_n\) is the distribution function of \(X_1 + \cdots + X_n\) for \(n \in \{1, \ldots, N\}\).

Customer \(i\), who has position \(i\) in the priority list, receives a service level of \(G_i(S)\), which needs to be larger than \(\beta_i\) for the priority list to be feasible for customer \(i\). It follows that the optimal inventory given \(\Pi\) is \(\max\{\tilde{S}, \Gamma_1^{-1}(\beta_k), \Gamma_2^{-1}(\beta_{k+1})\}\) and the optimal inventory given \(\Pi'\) is \(\max\{\tilde{S}, \Gamma_3^{-1}(\beta_{k+1}), \Gamma_2^{-1}(\beta_k)\}\). Note that \(\Gamma_2^{-1}(\beta_{k+1})\) is greater than \(\Gamma_1^{-1}(\beta_k)\), \(\Gamma_2^{-1}(\beta_k)\) as well as \(\Gamma_3^{-1}(\beta_{k+1})\). Hence an adjacent pairwise interchange results in a decrease in the optimal inventory (at worst, the decrease is zero; this happens when \(\tilde{S}\) is the dominant term). The theorem follows after making finitely many pairwise interchanges.
**Proof of Theorem 9**

Without loss of generality, relabel customers if necessary to ensure $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_N$. Let $z_i = \Phi^{-1}(\beta_i)$ with $\Phi(\cdot)$ denoting the cdf of standard normal distribution. The optimal inventory under no pooling is $\bar{S} = \sum_{i=1}^N S_i$, where $S_i = \mu_i + z_i \sigma_i$. The optimal inventory that corresponds to the optimal fixed list policy, applying Theorem 8, is $S^*_{\text{list}} = \max_{n=1,\ldots,N} \{m_n + z_n s_n\}$, where $m_n = \sum_{i=1}^n \mu_i$ and $s_n^2 = \sum_{i=1}^n \sigma_i^2 + 2 \sum_{i=1}^n \sum_{j=i}^n \rho_{ij} \sigma_i \sigma_j$ are the mean and variance of $X_1 + \cdots + X_n$, respectively (i.e., $G_n(\cdot)$ used in the theorem is a normal distribution with mean $m_n$ and standard deviation $s_n$ in this case). Now observe that $S_1 + \cdots + S_n \geq m_n + z_n s_n$ if and only if $\sum_{i=1}^n \sigma_i^2 (z_i^2 - z_n^2) \geq 2 \sum_{i=1}^n \sum_{j=i}^n \rho_{ij} \sigma_i \sigma_j (\rho_{ij} z_n^2 - z_i z_j)$ for all $n = 1, \ldots, N$. This inequality holds for all $n$ because $z_1 \geq \cdots \geq z_N$. It follows that $\bar{S} \geq S^*_{\text{list}}$.

**Proof of Theorem 10**

Since $X_i$ are iid random variables, $G_n(S)$ stands for the service level received by the customer with rank $n$. (Without the iid demand assumption, this service level would depend on the identities of the customers with ranks 1 through $n-1$, which would invalidate the analysis that follows.) After unconditioning by the probabilities generated by randomization, we obtain $G_1(S) w_{n1} + \cdots + G_N(S) w_{nN} \geq \beta_n$ for $n = 1, \ldots, N$. Part (a) of the theorem follows. Summing up these inequalities, we obtain $\sum_{n=1}^N G_n(S) \geq \sum_{i=1}^N \beta_i$, which must be satisfied by every feasible $S$. Since $G_n(S)$ are each continuous and monotone increasing in $S$, the smallest $S$ that satisfies the inequality is the unique solution $S_c$ of the equation $\sum_{n=1}^N G_n(S) = \sum_{i=1}^N \beta_i$. (Note that the left-hand-side is strictly increasing in $S$, equal to 0 at $S = 0$, and goes to $N$ in the limit as $S$ approaches $+\infty$, which is why $S_c$ is unique.) Hence, $S_c$ is a lower bound for the optimal stock, proving (b). Consider the system of equations $\mathbf{WA} = \mathbf{B}$. It is a standard result in linear algebra that $\mathbf{W}$ is doubly stochastic if and only if $\mathbf{A}$ majorizes $\mathbf{B}$. This proves part (c).

**Proof of Theorem 11**

Let $\phi(\cdot)$ and $\Phi(\cdot)$ denote the pdf and cdf of standard normal distribution, respectively. Define $\sigma_{12} \equiv \sqrt{\sigma_1^2 + 2 \rho \sigma_1 \sigma_2 + \sigma_2^2}$ and $\bar{z} \equiv (S_0 - \mu_1 - \mu_2) / \sigma_{12}$.

(1a) In Case 1, by the implicit function theorem and the chain rule,

$$
\frac{\partial S^*}{\partial \rho} = \left. \frac{\partial S_0}{\partial \rho} \right|_{S=S_0} - \left. \left( \frac{\partial}{\partial S} \left[ P \{X_1 + X_2 \leq S \} - P \{X_1 > S, X_2 > S \} \right] \right) \right|_{S=S_0}
$$
The denominator is positive. The second term of the numerator, \( \frac{\partial}{\partial \rho} P \{ X_1 > S, \ X_2 > S \} \bigg|_{S=S_0} \), is positive by Slepian’s inequality\(^1\). Finally, the first term of the numerator is negative by the following result, provided that \( S_0 > \mu_1 + \mu_2 \) (we omit the algebraic details of the derivation).

\[
\frac{\partial}{\partial \rho} P \{ X_1 + X_2 \leq S \} \bigg|_{S=S_0} = -\phi \left( \frac{\bar{z}}{\sigma_{12}} \right) \frac{\sigma_1 \sigma_2}{\sigma_{12}^2} < 0
\]

Therefore, \( S_0 \) is increasing in \( \rho \) under Case 1 as long as \( S_0 > \mu_1 + \mu_2 \). In Cases 2 and 3, \( S^* \) is invariant with \( \rho \), because it depends purely on one of the marginal demand distributions.

(1b) Since \( (S_1 + S_2) \) is invariant with \( \rho \), the result follows from (1a).

(2a) Recall that we are fixing \( \theta = \theta_1 = \theta_2 \) and setting \( \sigma_i = \theta \mu_i \) for \( i = 1, 2 \). In Case 1, by the implicit function theorem and the chain rule,

\[
\frac{\partial S^*}{\partial \theta} = \frac{\partial S_0}{\partial \theta} = -\frac{\frac{\partial}{\partial S} \left[ P \{ X_1 + X_2 \leq S \} - P \{ X_1 > S, \ X_2 > S \} \right]_{S=S_0}}{\frac{\partial}{\partial S} \left[ P \{ X_1 + X_2 \leq S \} - P \{ X_1 \leq S \} \right]_{S=S_0}}
\]

Since the denominator is positive, \( S_0 \) is increasing in \( \theta \), iff

\[
\frac{\partial}{\partial \theta} P \{ X_1 > S, \ X_2 > S \} \bigg|_{S=S_0} > \frac{\partial}{\partial \theta} P \{ X_1 + X_2 \leq S \} \bigg|_{S=S_0}
\]

Using the Leibniz Rule and the conditional distributions of \( X_1 \mid X_2 \) and \( X_2 \mid X_1 \), we obtain:

\[
\frac{\partial}{\partial \theta} P \{ X_1 > S, \ X_2 > S \} \bigg|_{S=S_0} = \frac{1}{\theta} \sum_{i=1}^{2} \left( \frac{S_0 - \mu_1}{\sigma_i} \right) \left[ 1 - \Phi \left( \frac{\frac{S_0 - \mu_{3-i}}{\sigma_{3-i}} - \rho \frac{S_0 - \mu_i}{\sigma_i}}{\sqrt{1 - \rho^2}} \right) \right]
\]

This is positive as long as \( S_0 > \mu_1 \) and \( S_0 > \mu_2 \). Furthermore,

\[
\frac{\partial}{\partial \theta} P \{ X_1 + X_2 \leq S \} \bigg|_{S=S_0} = -\phi \left( \frac{\bar{z}}{\theta} \right)
\]

which is negative iff \( S_0 > \mu_1 + \mu_2 \). Therefore, \( S_0 \) is increasing in \( \theta \) under Case 1 as long as \( S_0 > \mu_1 + \mu_2 \). In Cases 2 and 3, the result follows from \( \partial S_i / \partial \theta = \Phi^{-1} (\beta_i) \mu_i \) for \( i = 1, 2 \).

(2b) In Case 1, by the implicit function theorem and the chain rule, \( \frac{\partial}{\partial \theta} (S_1 + S_2 - S_0) > 0 \) iff

\[
\frac{\partial}{\partial \theta} [P \{ X_1 > S, X_2 > S \} - P \{ X_1 + X_2 \leq S \}] \bigg|_{S=S_0} < \left( \sum_{i=1}^{2} \Phi^{-1} (\beta_i) \mu_i \right) \frac{\partial}{\partial S} [P \{ X_1 + X_2 \leq S \} - P \{ X_1 > S, X_2 > S \}] \bigg|_{S=S_0}
\]

Some algebra yields the following equivalent condition:

\[
\phi \left( \frac{\bar{z}}{\sigma_{12}} \right) \frac{(S_0 - S_1 - S_2)}{\sigma_{12}^2} < 1 - \Phi \left( \frac{\frac{S_0 - \mu_2}{\sigma_2} - \rho \frac{S_0 - \mu_1}{\sigma_1}}{\sqrt{1 - \rho^2}} \right) (S_1 + S_2 - S_0 - \mu_2)
\]

\[
+ \frac{\sigma_1}{\sigma_2} \Phi \left( \frac{\frac{S_0 - \mu_1}{\sigma_1} - \rho \frac{S_0 - \mu_2}{\sigma_2}}{\sqrt{1 - \rho^2}} \right) (S_1 + S_2 - S_0 - \mu_1)
\]

The left-hand-side is strictly negative by Theorem 7. A sufficient condition for the right-hand-side to be positive is that the pooling benefit \((S_1 + S_2 - S_0)\) is larger than both \(\mu_1\) and \(\mu_2\). In Cases 2 and 3, the pooling benefit is \(S_2\) and \(S_1\), respectively; the result follows from \(\partial S_i/\partial \theta = \Phi^{-1}(\beta_i)\mu_i\) for \(i = 1, 2\).

**Appendix B: Bivariate Normal Demands**

Consider two customers with demands \(X_1\) and \(X_2\) following a bivariate normal distribution with respective means \(\mu_1\) and \(\mu_2\), standard deviations \(\sigma_1\) and \(\sigma_2\), and the correlation coefficient \(\rho \in [-1, 1]\). In conformance with standard practice in the field, we ignore the possibility of negative demand, or—in effect—assume such problem parameters that the probability of negative demand is negligibly small.

We explore how the optimal inventory level \((S^*\), derived in Theorem 6) and the magnitude of the pooling benefit \((S_1 + S_2 - S^*)\) behave as a function of the correlation coefficient and the coefficient of variation.

**Theorem 11** Suppose that the customer demands \((X_1, X_2)\) follow a bivariate normal distribution (with the parameters introduced above), and the customers’ required service levels are \(\beta_1\) and \(\beta_2\). Then, the optimal inventory level with pooling \((S^*)\), and the magnitude of the pooling benefit \((S_1 + S_2 - S^*)\) vary with the correlation coefficient and the coefficient of variation (ceteris paribus) as follows:

<table>
<thead>
<tr>
<th>Change in Parameter</th>
<th>Resulting Change in (S^*), Sufficient Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1a) (\rho \uparrow)</td>
<td>(\uparrow), if Case 1 holds and (S_0 &gt; \mu_1 + \mu_2)  (\leftarrow\rightarrow), if Case 2 or Case 3 holds</td>
</tr>
<tr>
<td>(2a) (\theta \uparrow)</td>
<td>(\uparrow), if Case 1 holds and (S_0 &gt; \mu_1 + \mu_2)  (\uparrow), if Case 2 holds and (\beta_1 \geq 50%)  (\uparrow), if Case 3 holds and (\beta_2 \geq 50%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change in Parameter</th>
<th>Resulting Change in ((S_1 + S_2 - S^*)), Sufficient Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1b) (\rho \uparrow)</td>
<td>(\downarrow), if Case 1 holds and (S_0 &gt; \mu_1 + \mu_2)  (\leftarrow\rightarrow), if Case 2 or Case 3 holds</td>
</tr>
<tr>
<td>(2b) (\theta \uparrow)</td>
<td>(\uparrow), if Case 1 holds and (S_1 + S_2 - S_0 &gt; \max(\mu_1, \mu_2))  (\uparrow), if Case 2 holds and (\beta_2 \geq 50%)  (\uparrow), if Case 3 holds and (\beta_1 \geq 50%)</td>
</tr>
</tbody>
</table>

\([\uparrow\): increase; \(\downarrow\): decrease; \(\leftarrow\rightarrow\): no change]
where $\theta_i \equiv (\sigma_i / \mu_i)$ is the coefficient of variation; and results (2a) and (2b) assume $\theta_i$ to be the same for both customers, i.e., $\theta_1 = \theta_2 \equiv \theta$, and vary standard deviations proportionally while fixing the means. The condition that $S_0 > \mu_1 + \mu_2$ is guaranteed to hold when $\beta_1 + \beta_2 \geq 150\%$.

There are versions of the first result in the pooling literature, and it is quite intuitive: the more positively correlated the demands, the less the pooling benefit. It is something of a surprise that the pooling benefit can decrease as $\rho$ drops, which tends to happen in low-service-level cases (e.g., when $\beta_1, \beta_2 \leq 75\%$ in our case). The limitation in the second result is also surprising: an increase in demand variability may not always lead to more pooling benefit, although it usually does.