A Model for Integrated Inventory and Assortment Planning

Victor Martínez-de-Albéniz[1] • Sumit Kunnumkal[2]

Abstract

Integrating inventory and assortment planning decisions is a challenging task that requires comparing the value of demand expansion through broader choice for consumers, with the value of higher in-stock availability. We develop a stockout-based substitution model for trading off these values, in a setting with limited store capacities and inventory replenishment, two features missing in the literature. Using the closed-form solution for the single product case, we develop an accurate approximation for the multi-product case. This approximated formulation allows us to optimize inventory decisions by solving a fractional integer program with a fixed-point equation constraint. When products have equal margins, we solve the integer program exactly by bisection over a one-dimensional parameter. In contrast, when products have different margins, we propose a fractional relaxation that we can also solve by bisection, and results in near-optimal solutions. Overall, our approach provides solutions within 0.1% of the optimal policy and finds the optimal solution in 80% of the random instances we generate.

Submitted: October 14, 2019. Revised: June 7, 2020.

Keywords: stockout-based substitution, choice models, Markov chain, integer programming.

1. Introduction

The general practice in retail management is to take assortment and inventory planning decisions separately. Indeed, retailers usually take assortment breadth decisions first, by offering a given assortment for each store cluster; they then set inventory levels for each product, within store capacity constraints. For instance, for fashion apparel, large stores receive the full assortment from the current collection, while small/medium stores only receive the most popular items. In fact, the reason small stores receive only a fraction of the assortment is to avoid stock-outs: the retailer typically prefers to offer high service levels for the popular products, at the price of sacrificing variety.

Similarly, most of the academic research does not connect assortment breadth and inventory depth. Specifically, most assortment planning models ignore inventory costs (Cachon et al. 2005, Caro and Gallien 2007, Heese and Martínez-de Albéniz 2018), while most of the works on inventory management focus on the single-product case or with independent demand streams (Zipkin 2000, Axsäter 2006). There is only a minority of works that consider both: there exist models with inventory competition, but where assortment is taken as fixed (Lippman and McCardle 1997, Tsay and Agrawal 2000, Netessine and Rudi 2003) and there are assortment models that include inventory

[1]IESE Business School, University of Navarra, Av. Pearson 21, 08034 Barcelona, Spain, email: valbeniz@iese.edu.
[2]Indian School of Business, Gachibowli, Hyderabad, India, email: Sumit_Kunnumkal@isb.edu

costs, but usually with a fixed, common service level across products (van Ryzin and Mahajan 1999). These models thus do not allow retailers to prioritize certain items with higher inventory and service level, and restrict the availability of others. This has important implications: with the existing models, a retailer needs to sacrifice service level for all products when variety increases, and cannot implement policies where service level is differentiated across items.

While making assortment and inventory decisions jointly has the potential of lifting retailer profits, considering both decisions together is difficult. Indeed, there is only a handful of papers that have attempted this goal. Mahajan and van Ryzin (2001) consider a demand model, where substitution across products depends on their availability, i.e., a product is considered by the consumer only if the current inventory level is at least one. This behavior is called *stockout-based substitution*. The inventory optimization problem can be elegantly formulated, but Mahajan and van Ryzin (2001) show that the profit function is not well-behaved as a function of inventory levels and hence there is no easy way to optimize decisions if not by full enumeration of all possibilities. Because of the intrinsic difficulty, approximations have been proposed. To highlight two, Honhon et al. (2010) assume that demand has fixed proportions (i.e., the demand rate is equal to the average), until one of the products runs out, and are then able to propose an efficient procedure to optimize inventory levels; Farahat and Lee (2017) solve a relaxation of the multi-product newsvendor problem under stockout-based substitution, using an algebraic approach. Unfortunately, these existing results take a newsvendor perspective in a single period setting, i.e., where there is a single inventory decision. As a result, they have limited applicability in practical settings, because they ultimately assume that all products eventually stock out.

In this work, we try to remedy these shortcomings, by considering a stockout-based substitution problem with replenishment, so that inventory decisions directly influence maximum and average inventory levels in the store. In addition, we introduce a store capacity constraint to reflect the trade-off between breadth (variety) and depth (service level). Our ambition is to build a tractable framework that is amenable to optimization, and can possibly incorporate other types of business constraints such as maximum display quantities or nested relationships between products (Davis et al. 2013).

For this purpose, we assume that the retailer makes a one-time decision on target inventory levels for each product, and that upon a sale, items are replenished. Because demand depends on the inventory level, the resulting system is a finite-state Markov chain, where state transitions depend on the joint inventory state and in particular vary when a product is stocked out. Since the exact Markov chain steady-state probabilities cannot be determined analytically, we propose an approximated Markov chain where different products evolve independently of each other, and are affected by the average substitution patterns across products (as opposed to state-dependent

substitution). This allows us to approximate expected retailer profits accurately, using an approximation of in-stock probabilities. Specifically, the approximate service level of a given product only depends on its inventory level and a single-dimensional parameter that captures the entire assortment's attractiveness.

The resulting objective can then be optimized over all inventory choices. This optimization problem is difficult: it is a mixed integer program with a fixed-point constraint. Our main contribution in the paper is to develop algorithms that identify the optimal solution when all products have identical margins, and generate a heuristic solution otherwise. These algorithms are FPTAS schemes, that require bisection search for a one-dimensional parameter. We show that, when the heuristic solution is not optimal, it only introduces a small optimality gap, for which we offer a theoretical guarantee. We show in our numerical study that the gap is zero in 80% of the instances that we considered, and 0.06% on average if positive. Moreover, we show that joint optimization of assortment and inventory can lift profits compared to sequential approaches, especially when lead-times are long. Our approach thus advances the understanding of stockout-based substitution systems when replenishment is allowed. It provides effective algorithms for retailers that need to balance breadth vs. depth in stores with limited capacity.

The rest of the paper is organized as follows. §2 describes the relevant literature. We define the problem in §3 and characterize the steady-state probabilities in §4. §5 develops our solution techniques. We report their practical performance in a numerical study in §6. We conclude in §7. All proofs are contained in the Appendix.

2. Literature Review

Our work is related to the literature on assortment optimization. There is a large literature on assortment optimization under a variety of choice models and we refer the reader to Kök et al. (2009) for a detailed review. We limit our review to work that considers the Multinomial Logit (MNL) choice model. Talluri and van Ryzin (2004) show that the assortment optimization problem can be solved efficiently under the MNL model. Further, they show that the optimal assortment is revenue ordered and includes the top k revenue products for some k. Rusmevichientong et al. (2010) consider the assortment optimization problem under the MNL model but with a cardinality constraint that limits the number of products that can be included in the assortment. Although the optimal assortment is not necessarily revenue ordered, they show that the cardinality constrained problem can still be solved efficiently. This body of work does not consider the inventory decisions.

Stockout-based substitution models attempt to bridge this gap and connect the assortment and the inventory decisions. Mahajan and van Ryzin (2001) consider a fluid model of demand and propose a stochastic gradient algorithm that is guaranteed to converge to a stationary point of the

profit function. **?** propose an approximation of the MNL assuming static fill-rate service levels, with a spirit similar to ours but restricted to a single period without replenishment. Honhon et al. (2010) consider a deterministic approximation of the demand, while **?** study the impact of the fixed-proportion assumption and show that it only has a very minor influence on retailer profits. Farahat and Lee (2017) consider relaxations of the multi-product newsvendor problem to obtain upper bounds on the optimal profit. The above mentioned papers assume that each customer is endowed with a ranked list of preferences over the products and chooses the highest ranked product from the available ones. The MNL model can be viewed as a special case of such a choice model. **?** consider a total inventory constraint as we do, and develop a PTAS approximation when preferences are nested by price or quality, based on categorizing products as frequent vs. rare, and cheap vs. expensive. They show that an assortment with a small number of items is able to generate near-optimal profits. **?** provide an alternative approximation under general customer preference list with strong performance guarantees. Their algorithm first selects an appropriate assortment by separating frequent and rare products, and then optimizes the inventory level in a greedy newsvendor fashion. Aouad et al. (2018) propose an approximation algorithm specialized to the MNL choice model. We note that there are a number of papers which consider other choice models (for example, Gaur and Honhon 2006 use a locational choice model) and other forms of substitution (for example, Smith and Agrawal 2000 assume static substitution where the choice of a customer does not depend on the inventory levels). We refer the reader to Kök et al. (2009) and Hübner and Kuhn (2012) for more detailed reviews. As noted previously, all of the above mentioned papers take a newsvendor perspective and assume that there are no product replenishments.

Our work is also broadly related to analytical and empirical studies that assess the impact of inventory on sales through stock-out and substitution effects, see Urban (2005) for a review of classical works. More recent representative works include Kök and Fisher (2007), Musalem et al. (2010), Cachon et al. (2019), and Boada-Collado and Martínez-de Albéniz (2020). The first paper develops a methodology to measure substitution effects to optimize assortments, while the last three papers are concerned with estimating the effect of stock-outs and inventory levels on sales. In particular, Boada-Collado and Martínez-de Albéniz (2020) prescribe inventory decisions that take into account the effect of inventory depth on sales; however, their inventory model does not take into account stockout-based substitution.

3. The model

3.1 The store

Consider a retailer that is operating a store with limited capacity within a certain product range to be sold to consumers. This limitation typically appears from the space constraints that the store might have. For instance, this can be the number of facings that a supermarket might have in a certain aisle, dedicated to a product category. Or it can be the number of units that can be carried in the Women's section in a fashion store, which may include the capacity for folded garments (e.g., on a table display) plus that for hangers (e.g., on a wall rack).

Within that part of the store, the retailer must decide how many different products to offer within a potential assortment $N = \{1, \ldots, n\}$, and what the corresponding quantities of the products $\mathbf{Q} = (Q_1, \ldots, Q_n)$ should be. Each product has a profit margin and a preference weight associated with it. We let r_i denote the margin of product i and v_i denote its preference weight, which can be interpreted as a measure of its attractiveness. When product i is included in the chosen assortment then $Q_i > 0$; otherwise $Q_i = 0$. The retailer chooses the vector $\mathbf{Q} \in \mathcal{Q} \subset \mathbb{Z}_+^n$ (non-negative integers) that maximizes its expected average profits over an infinite horizon (formally defined below). In practice, the set $\mathcal{Q}$ can take different forms but in this paper we focus on situations where there is a limitation on total store inventory, i.e., $\mathcal{Q} = \{\mathbf{Q} | \sum_{i=1}^n Q_i \leq C\}$, where C denotes the store capacity. Note that one could consider other types of constraints as well, e.g., space constraints where each unit occupies b_i units of space and $\sum_{i=1}^n b_i Q_i \leq B$. While this does not affect our approximation of the steady state of the system, it crucially affects our optimization procedure, because we cannot round the fractional solutions anymore, see Kunnumkal and Martínez-de Albéniz (2019) for a lengthy discussion on this point.

3.2 The demand process

At any point in time t, there is a chance that a customer enters the store. The arrival process is modeled as a homogeneous Poisson process of rate $\lambda \geq 0$. Without loss of generality, we use $\lambda = 1$. Upon arrival, a customer will observe the number of units in the existing assortment at time t, $\mathbf{q}_t = (q_{1t}, \ldots, q_{nt})$ and choose which of the available products (i.e., those with $q_{it} > 0$) fits her tastes best. We assume that the utility provided by each product is the logarithm of the product's preference weight ($\log(v_i)$) plus a Gumbel-distributed shock, which results in the well-known Multinomial Logit (MNL) choice model. Namely, the probability of this customer picking product i is given by:

$$p_{it}(\mathbf{q}_t) = \frac{v_i 1_{q_{it}>0}}{1 + \sum_{j=1}^n v_j 1_{q_{jt}>0}} \tag{1}$$

where 1_X is a binary variable equal to one if and only if X is true. We normalize the attractiveness of the outside option to one, i.e., the probability of the customer walking away without purchasing anything is given by $p_{0t}(\mathbf{q}_t) = \frac{1}{1+\sum_{j=1}^{n} v_j 1_{q_{jt}>0}}$. This setting is the classical stockout-based substitution model of Mahajan and van Ryzin (2001) and Honhon et al. (2010). After the demand has occurred, the state of the system, i.e., the availability vector $(q_{1t}, \ldots, q_{nt})$, will be updated: if product i has been purchased q_{it} drops by one. The expected revenue from the customer is

$$R^{exact}(\mathbf{q}) = \frac{\sum_{i=1}^{n} r_i v_i 1_{q_i>0}}{1+\sum_{i=1}^{n} v_i 1_{q_i>0}}. \tag{2}$$

3.3 The replenishment process

Immediately after the sale, the store manager requests the retailer's distribution center more supply of that item. We assume that items are replenished one by one, which is a widely used inventory policy in practice, because it is the optimal policy when there are no fixed ordering costs and product demands are independent (Zipkin 2000). In other words, we employ a state-independent order-up-to policy. Note that, in our case, product demands are not independent of each other – note that when there are inter-product demand effects, the inventory policy is difficult to characterize (Beyer et al. 2001).

Each replenishment of product i is received after L_i time units after ordering, where L_i is a random variable that is realized independently for each order. That is, we allow for orders to cross (Kaplan 1970, Muharremoglu and Yang 2010). Furthermore, we assume in our analysis that L_i is exponentially distributed with rate μ_i, so as to exploit the memoryless property of the exponential distribution. Without such assumption, the evolution of the system would not only depend on the current inventory levels $(q_{1t}, \ldots, q_{nt})$, but also on all the times in which the on-order items were ordered; such case is generally untractable. Nevertheless, we complement our analysis with a simulation study with non-exponential lead times in §6.4. We find that our suggested policies, derived under exponential lead-times, continue to exhibit very good performance.

3.4 The system in steady state and the retailer's problem

The system defined above generates a dynamic stochastic process where the availability vector $\tilde{\mathbf{Q}}_t$ is a random variable, starting at $\mathbf{Q}$ at time $t = 0$ and changing at arrival and replenishment events. We denote $\mathbf{q}_t$ a sample path of this random variable. Note that because demand is generated by a Poisson process and replenishment times are exponential, the state of the system is simply defined by $\mathbf{q}_t$. The state falls within a finite set such that $q_{it} \in \{0, \ldots, Q_i\}$, in which case there are q_{it} units in stock at the store, and $Q_i - q_{it}$ units pending delivery. Because the number of states is finite, and the system has no memory, there exists a unique steady-state distribution $\tilde{\mathbf{Q}}$ following

Kolmogorov's ergodic theorem (**?**).

Specifically, for every state $\mathbf{q}$, we can define $\pi(\mathbf{q}|\mathbf{Q}) = Pr[\tilde{\mathbf{Q}} = \mathbf{q}|\mathbf{Q}]$ to be the steady-state probability. Letting $\rho_{\mathbf{q_1},\mathbf{q_2}}$ be the infinitesimal probability of transitioning from state $\mathbf{q_1}$ to state $\mathbf{q_2}$, we have that

$$\sum_{\mathbf{q}'} \pi(\mathbf{q}|\mathbf{Q})\rho_{\mathbf{q},\mathbf{q}'} = \sum_{\mathbf{q}'} \pi(\mathbf{q}'|\mathbf{Q})\rho_{\mathbf{q}',\mathbf{q}}. \tag{3}$$

The retailer's problem thus involves finding the stocking level that maximizes the expected steady-state profit:

$$R^{exact} := \max_{\mathbf{Q}\in\mathcal{Q}} \sum_{\mathbf{q}:0\leq q_i\leq Q_i \forall i} \pi(\mathbf{q}|\mathbf{Q})R^{exact}(\mathbf{q}). \tag{4}$$

Our formulation thus does not optimize the assortment dynamically over time, depending on the current state of the system, but instead optimizes in a static fashion whether an item should be included in the assortment, and, if so, what static order-up-to level should be used.

Because of the dependency of $\pi(\cdot|\mathbf{Q})$ on $\mathbf{Q}$, this problem is intractable to solve. Specifically, there are two difficulties: first, it is difficult to characterize $\pi(\cdot|\mathbf{Q})$, due to the large number of possible states in the system; second, it is even more difficult to evaluate how $\mathbf{Q}$ affects $\pi(\cdot|\mathbf{Q})$, so that it is hard to select a good strategy. Fortunately, we can characterize $\pi(\cdot|\mathbf{Q})$ exactly when there is a single product, and this is helpful to approximate $\pi(\cdot|\mathbf{Q})$ in the general multi-product case.

4. Steady-state system behavior

4.1 Single product

When there is a single product (we remove sub-index i to ease notation), there are exactly $Q+1$ states in the system: $\{0,\ldots,Q\}$. We have that $\rho_{q,q+1} = \mu(Q-q)$ because when there are q units on stock, there are $Q-q$ on order; $\rho_{q,q-1} = v/(1+v)$; and $\rho_{q,q'} = 0$ if $|q-q'| > 1$. Figure 1 depicts the states and transition probabilities.

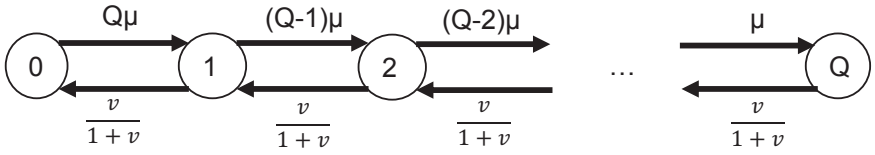

Figure 1: States and transition rates for the Markov chain with one product.

For any $q \geq 1$, Equation (3) thus becomes

$$\pi(q|Q)v/(1+v) = \pi(q-1|Q)\mu(Q-q+1)$$

which implies that

$$\pi(q|Q) = \left(\frac{\mu(1+v)}{v}\right)^q \frac{Q!}{(Q-q)!}\pi(0|Q),$$

where $\pi(0|Q)$ is the probability of the system being out of stock. Since the steady state probabilities add up to one, we have that

$$\pi(0|Q) = \frac{1}{\sum_{q=0}^{Q}\left(\frac{\mu(1+v)}{v}\right)^q \frac{Q!}{(Q-q)!}} = \frac{1}{\sum_{q=0}^{Q}\left(\frac{\mu(1+v)}{v}\right)^{Q-q} \frac{Q!}{q!}}. \tag{5}$$

Therefore the probability of the product being available

$$a(Q) = 1 - \pi(0|Q) = 1 - \frac{1}{\sum_{q=0}^{Q}\left(\frac{\mu(1+v)}{v}\right)^{Q-q} \frac{Q!}{q!}}. \tag{6}$$

Note that the in-stock probability depends on v, μ and Q; it decreases with v, and increases with μ and Q.

4.2 Multiple products

When there are multiple products ($n \geq 2$) in the assortment, the steady-state probabilities from Equation (3) cannot be characterized in closed form. While they can be computed by solving the system of linear equations, we still need to understand how they vary with the inventory decision $\mathbf{Q}$. The difficulty stems from the fact that distribution over q_i and q_j is in general correlated. Indeed, consider the case with $n = 2$ products: for $q_2 > 0$, $\rho_{(q_1,q_2)(q_1,q_2-1)} = \frac{v_2}{1+v_1 1_{q_1>0}+v_2}$: this is higher when $q_1 = 0$ compared to $q_1 > 0$. Figure 2 illustrates the different states and transitions for two products.

We propose here an approximation of the steady-state probabilities $\hat{\pi}$ that assumes independence. Specifically, we let α_i be a parameter which captures the availability of product i in the steady state, or its service level. That is, by construction we define $\alpha_i = Pr[\tilde{Q}_i > 0|\mathbf{Q}] = E\left[1_{\tilde{Q}_i>0}|\mathbf{Q}\right]$ where expectation is taken with respect to the approximated system dynamics. Letting $\mathbf{e}_i$ denote the unit vector with a 1 in the ith component and zeros elsewhere, we approximate the transition rate $\rho_{\mathbf{q},\mathbf{q}-\mathbf{e}_i} = v_i/(1+\sum_{j=1}^{n} v_j 1_{q_j>0})$ by $\hat{\rho}_{\mathbf{q},\mathbf{q}-\mathbf{e}_i} = v_i/(1+\sum_{j=1}^{n} v_j\alpha_j)$ by replacing the indicator variables by their expected values. Alternative approximations could be considered as well, such as $\hat{\rho}_{\mathbf{q},\mathbf{q}-\mathbf{e}_i} = v_i/(1+v_i+\sum_{j\neq i} v_j\alpha_j)$; but the chosen one provides accurate approximations of the steady state with a tractable structure. We note that for positive levels of q_j, this increases the transition rate from state $\mathbf{q}$ to $\mathbf{q}-\mathbf{e}_i$ since the denominator is decreased. On the other hand,

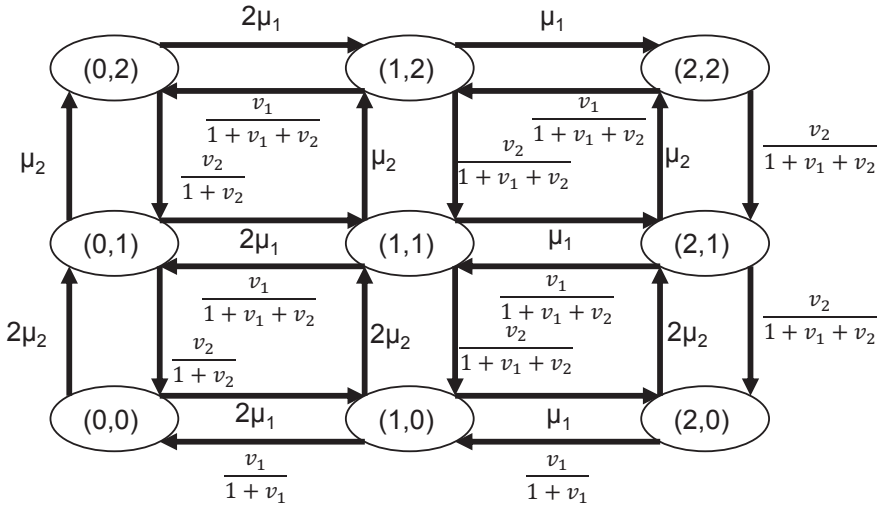

Figure 2: States and transition rates for the Markov chain with two products and $Q_1 = Q_2 = 2$.

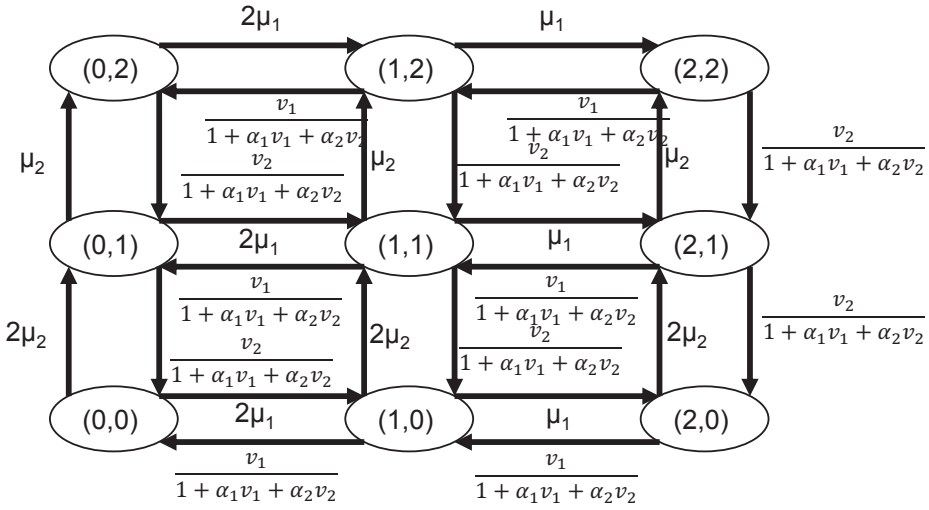

Figure 3: States and approximated transition rates for the Markov chain with two products and $Q_1 = Q_2 = 2$.

it decreases the transition rate when $q_j = 0$. Figure 3 illustrates the approximation corresponding to the Markov chain of Figure 2.

With this approximation, the rate at which the inventory of a given product gets depleted does not depend on the actual inventory levels of the remaining products, but only their average values. It can therefore be shown that the Markov chain with the approximated transition rates can be

decomposed into n single-product chains. We can then use the single-product analysis to obtain the stock-out probability of product i in the Markov chain with the approximated transition rates:

$$\hat{\pi}_i(0|Q_i) = \frac{1}{\sum_{q=0}^{Q_i} \left(\frac{\mu_i\left(1+\sum_{j=1}^n \alpha_j v_j\right)}{v_i} \right)^{Q_i - q} \frac{Q_i!}{q!}}. \tag{7}$$

The probability that product i is available, α_i, is therefore given by

$$\alpha_i = 1 - \hat{\pi}_i(0|Q_i) = 1 - \frac{1}{\sum_{q=0}^{Q_i} \left(\frac{\mu_i\left(1+\sum_{j=1}^n \alpha_j v_j\right)}{v_i} \right)^{Q_i - q} \frac{Q_i!}{q!}}. \tag{8}$$

Notice that computing the service level α_i's involves solving a system of n equations of the form (8). We can simplify the solution procedure considerably by introducing a variable $s := \sum_{j=1}^n v_j \alpha_j$, so that

$$\alpha_i = a_i(s, Q_i) := 1 - \frac{1}{\sum_{q=0}^{Q_i} \left(\frac{\mu_i(1+s)}{v_i} \right)^{Q_i - q} \frac{Q_i!}{q!}}, \tag{9}$$

where we use $a_i(s, Q_i)$ to indicate that the value α_i depends on both s and Q_i. We next characterize the properties of this function.

Lemma 1. *1. For any i, $a_i(s, Q_i)$ is concave increasing in s. Moreover, if $Q_i \geq 1$, then $a_i(s, Q_i)$ is strictly concave in s and $\lim_{s\to\infty} a_i(s, Q_i) = 1$.*
2. For any i, $a_i(s, Q_i)$ is concave increasing in Q_i and $\lim_{Q_i\to\infty} a_i(s, Q_i) = 1$.

Lemma 1 shows that $a_i(s, Q_i)$ is concave and increasing in each of its components. In particular, as the stocking level Q_i increases, $a_i(s, Q_i)$ increases, but at a diminishing rate. We can build on the results in Lemma 1 to obtain additional properties of $a_i(s, Q_i)$ that will be later useful to analyze the performance of our solution method. These are summarized in the lemma below.

Lemma 2. *1. For any i, $a_i(s, Q_i)/(1 + s)$ is decreasing in s.*
2. For any i, $a_i'(s, Q_i)/a_i(s, Q_i)$ is decreasing in Q_i, where $a_i'(s, Q) = \frac{\partial a_i(s,Q)}{\partial s}$.

Lemma 1 makes it easy to compute the product availability probabilities. Note that for a given inventory vector $\mathbf{Q}$, we can obtain $\mathbf{a}(s, \mathbf{Q}) = \Big(a_1(s, Q_1), \ldots, a_n(s, Q_n)\Big)$ by solving the one-dimensional fixed point equation

$$s = V(s, \mathbf{Q}), \tag{10}$$

where

$$V(s, \mathbf{Q}) = \sum_{i=1}^n v_i a_i(s, Q_i). \tag{11}$$

Further, Lemma 1 directly implies that Equation (10) admits a unique solution: at $s = 0$, the left hand side of Equation (10) is smaller than the right hand side since $a_i(0, Q_i) \geq 0$. On the other hand, at $s = \sum_{i=1}^{n} v_i$, the left hand side is larger than the right hand side (since $a_i(s, Q_i) \leq 1$). By Lemma 1, $V(s, \mathbf{Q}) = \sum_{i=1}^{n} v_i a_i(s, Q_i)$ is concave increasing in s. It follows that there is a unique solution, $s(\mathbf{Q})$, to Equation (10) in the interval $[0, \sum_{j=1}^{n} v_j]$; if $s \leq s(\mathbf{Q})$, then $V(s, \mathbf{Q}) \geq s$ and if $s \geq s(\mathbf{Q})$, then $V(s, \mathbf{Q}) \leq s$. We record this in the following corollary.

Corollary 1. *$V(s, \mathbf{Q}) \geq s$ if and only if $s \leq s(\mathbf{Q})$, where $s(\mathbf{Q})$ is the unique solution to Equation (10).*

We refer to $s(\mathbf{Q})$ satisfying Equation (10) as the total attractiveness of the approximated system in equilibrium, i.e., such that in-stock probabilities are all consistent with each other. Equation (10) thus internalizes the effect of availability of competing products on one's purchase probability.

Before we describe how the approximate in-stock probabilities help in making the inventory-assortment decisions, we briefly discuss how the in-stock probabilities obtained from the approximated Markov chain compare with those obtained from the original Markov chain. Figure 4 shows how the exact (Equation 6) and approximate (Equation 8) probabilities vary with Q for the single product case. The plot on the left has $\mu = 0.2$ (slow replenishment, it takes $1/\mu = 5$ periods on average to receive the product), while the plot on the right has $\mu = 1$ (quick replenishment, in one period on average). We note that while the approximation consistently underestimates the exact in-stock probabilities in the single product case, both curves qualitatively have the same shape. Moreover, the approximation becomes better as Q increases and as μ increases. At the same time, the gap between the probabilities can be quite large for small inventory quantities, and could be problematic given that, in retail, slow-moving goods may have very small inventory quantities in store, and these may represent a high fraction of the assortment. Fortunately, in the single product case, this does not affect our search for the optimal solution $Q^* = C$. Moreover, as we see next, the gap significantly reduces as more products are considered.

With two products, Figure 5 shows how the exact and approximate in-stock probabilities for the first product vary with Q_1 for different values of μ_1. We see a similar pattern as before in that the both curves have the same shape and the approximation becomes better as Q_1 increases and μ_1 increases. Finally, we compare the exact and approximate in-stock probabilities for test problems with a larger number of products. As the number of products increases, evaluating the exact steady-state probabilities analytically becomes difficult and so we use simulation to estimate them. Figure 6 shows how the exact and approximate in-stock probabilities vary with Q_1 for test problems with a larger number of products. The plot on the left compares the in-stock probabilities for a test problem with 10 products, while that on the right shows the comparison for a test problem with 25 products. Overall, we observe that the in-stock probabilities obtained from the approximated

Markov chain tend to be remarkably close to the ones of the original Markov chain even for moderate stocking levels. We further report in §6 the performance in the original Markov chain of the policies derived using the approximated Markov chain. Finally, we note that it is possible to bound the gap between the steady state probabilities obtained from the exact and approximated Markov chain using results from perturbation analysis of Markov chains (see for example, Cho and Meyer 2001 and O'Cinneide 1993). However, these bounds are not easily described in terms of the underlying problem parameters (v_i, μ_i, Q_i) and so do not provide much insight for our case. Moreover, the bounds can depend on the number of states of the Markov chain, which in our case, is exponential in the number of products. It is possible to come up with alternative, specialized bounds on the gap between the exact and the approximate in-stock probabilities. However, these bounds can be loose and so we omit them.

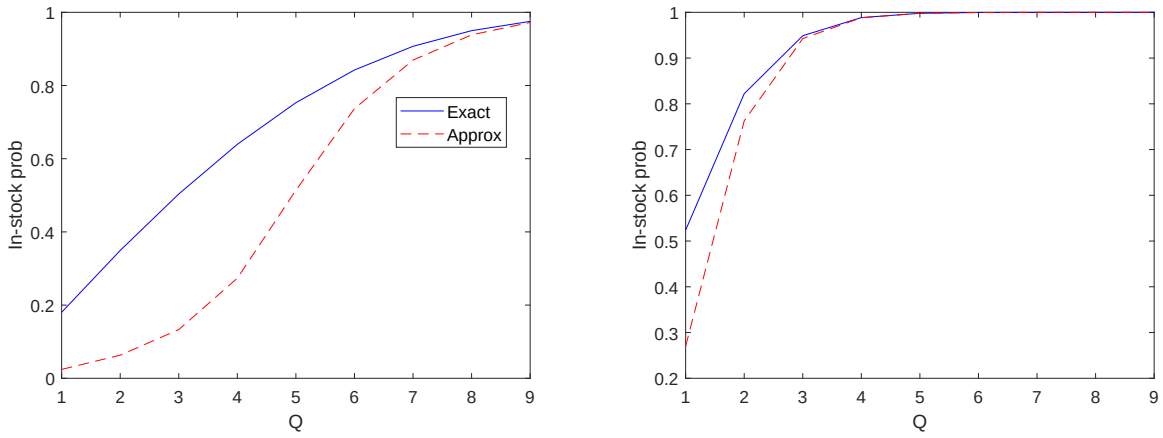

Figure 4: Exact and approximate in-stock probabilities as a function of Q for the single product case. The problem parameters for the plot on the left are $v = 10$ and $\mu = 0.2$, while those for the plot on the right are $v = 10$ and $\mu = 1$.

4.3 Approximating the revenue function

After having developed an approximation for service levels, via in-stock probabilities, we can now approximate the expected revenue function. Recall that the expected revenue corresponding to the inventory vector $\mathbf{Q}$ is $E\left[\left.\frac{\sum_{i=1}^n r_i v_i 1_{\tilde{Q}_i>0}}{1+\sum_{i=1}^n v_i 1_{\tilde{Q}_i>0}}\right| \mathbf{Q}\right]$. Replacing the indicator variables by their expectations and noting that $a_i(s(\mathbf{Q}), Q_i)$ is our approximation to $E\left[1_{\tilde{Q}_i>0}|\mathbf{Q}\right]$ – the service level measured as in-stock probability for product i– we obtain an approximated revenue function:

$$\frac{\sum_{i=1}^n r_i v_i a_i(s(\mathbf{Q}), Q_i)}{1+\sum_{i=1}^n v_i a_i(s(\mathbf{Q}), Q_i)} = \frac{\sum_{i=1}^n r_i v_i a_i\Big(s(\mathbf{Q}), Q_i\Big)}{1+s(\mathbf{Q})},$$

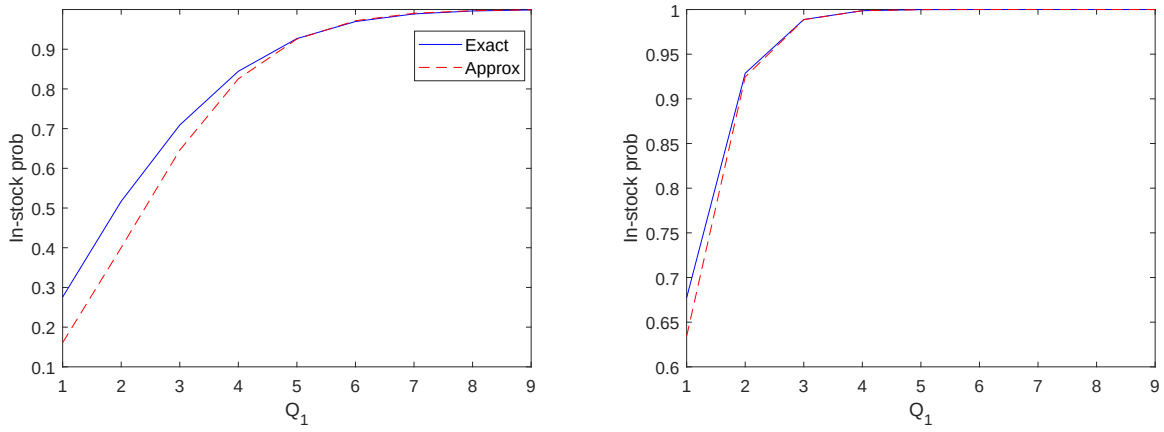

Figure 5: Exact and approximate in-stock probabilities as a function of Q_1 for a test problem with two products. The problem parameters are $v_1 = v_2 = 10$ and $Q_2 = 5$. The plot on the left has $\mu_1 = \mu_2 = 0.2$, while that on the right has $\mu_1 = \mu_2 = 1$.

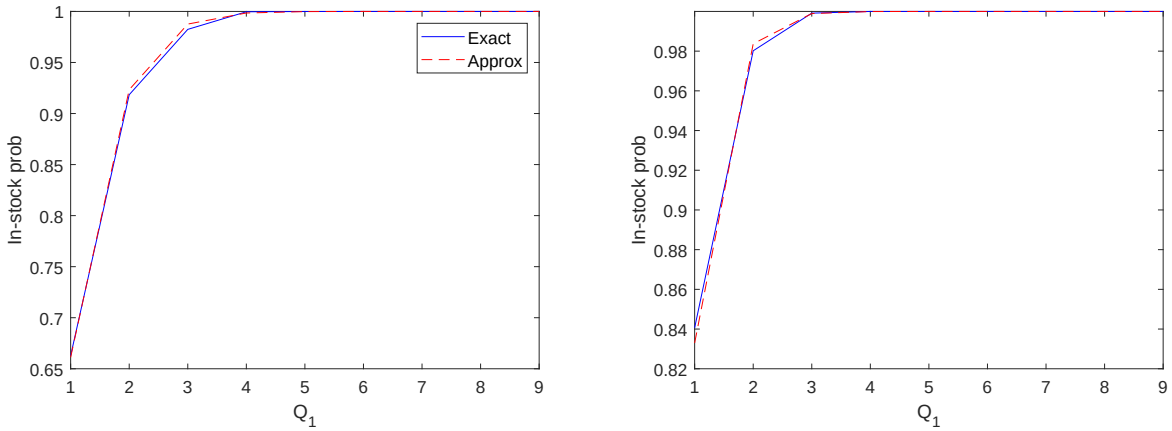

Figure 6: Exact and approximate in-stock probabilities as a function of Q_1 for test problems with $n = 10$ products (left) and $n = 25$ products (right). The problem parameters are $v_i = 10$, $\mu_i = 0.2$ and $Q_i = 5$ for all $i = 2, \ldots, n$.

where we use the fact that $s(\mathbf{Q})$ satisfies Equation (10). We note here that our approximated revenue function involves two approximations: we first replace the random variables describing the product availabilities by their expected values and then further approximate the in-stock probabilities. Notice that our approximate revenue function resembles the familiar-looking assortment planning problem, except that the inventory decision of $\mathbf{Q}$ has a direct impact on the variables $\mathbf{a}(s(\mathbf{Q}), \mathbf{Q})$ which drive the market shares of the products.

The approximated optimization problem hence becomes

$$R^{approx} = \max_{\mathbf{Q}\in\mathcal{Q}} \frac{\sum_{i=1}^{n} r_i v_i a_i\Big(s(\mathbf{Q}), Q_i\Big)}{1+s(\mathbf{Q})}. \tag{12}$$

Note that this formulation is similar to that of Boada-Collado and Martínez-de Albéniz (2020), but with the difference that here $a_i(s(\mathbf{Q}), Q_i)$ does not only depend on Q_i, but also on the implicit attractiveness at equilibrium $s(\mathbf{Q})$. This additional dependency radically changes the nature of the optimization problem. Indeed, when Q_i increases, then not only $a_i(s(\mathbf{Q}), Q_i)$ increases (by Lemma 1), but for $j \neq i$, $a_j(s(\mathbf{Q}), Q_j)$ increases as well. To see this, note that by Equation (10), an increase in $a_i(\cdot,\cdot)$ pushes up the implicit attractiveness at equilibrium $s(\mathbf{Q})$, which in turn pushes up the value of $a_j(\cdot,\cdot)$. This means that, although the products are substitutes, one can increase the attractiveness of a product j, $v_j a_j(s(\mathbf{Q}), Q_j)$, by increasing the service level $a_i(s(\mathbf{Q}), Q_i)$ of a competing item i (by increasing its inventory level Q_i), thereby sharing the demand more evenly and increasing the availability of both. In contrast, Boada-Collado and Martínez-de Albéniz (2020) show that when this interaction is missing, then the typical properties of optimal assortments remain, e.g., revenue-ordered sets are optimal.

Finally, observe that formulating problem (12) does not require knowledge of the functional form of a_i, provided that it only depends on s, and Q_i. Most importantly, the optimization machinery developed in the remainder of the paper would still apply for any function a_i satisfying the monotonicity and concavity properties derived in Lemmas 1 and 2.

5. Breadth vs. depth optimization

We now analyze the structure of the inventory-assortment optimization problem (12) and propose solution methods. The difficulty in the optimization program is the presence of $s(\mathbf{Q})$ which is given by the solution to the fixed-point equation (12). We introduce this fixed-point condition as a constraint in the optimization problem and write (12) as

$$R^{approx} = \max_{\mathbf{Q}\in\mathcal{Q}, s\geq 0} R(s,\mathbf{Q}) \text{ s.t. } V(s,\mathbf{Q}) = s, \tag{13}$$

where $V(s,\mathbf{Q})$ is as defined in (11) and

$$R(s,\mathbf{Q}) = \frac{\sum_{i=1}^{n} r_i v_i a_i(s,\mathbf{Q})}{1+s}. \tag{14}$$

We note that it is possible to write the constraint in (13) equivalently as $V(s,\mathbf{Q}) \leq s$, since it will be satisfied as an equation at the optimal value of s. To see this, suppose that $(s^*, \mathbf{Q}^*)$ is an optimal solution to (13) with $V(s^*,\mathbf{Q}^*) < s^*$. By Corollary 1 we have that $s(\mathbf{Q}^*) < s^*$. The first part of Lemma 2 together with (14) implies that $R(s,\mathbf{Q})$ is decreasing in s. Therefore, we have $R(s(\mathbf{Q}^*),\mathbf{Q}^*) \geq R(s^*,\mathbf{Q}^*)$ and $(s(\mathbf{Q}^*),\mathbf{Q}^*)$ is also an optimal solution to (13). Moreover

the constraint is satisfied as an equation since by definition $s(\mathbf{Q}^*) = V(s(\mathbf{Q}^*), \mathbf{Q}^*)$. Writing the constraint as an inequality ensures that there is always a feasible solution to optimization problem (13) and this turns out to be useful in our numerical implementation of the solution method.

In this section, we first study the case with identical product margins, and develop a FPTAS algorithm to find the optimal solution (Lemma 3). We then turn to the much harder problem with general margins. We develop a continuous relaxation (Lemma 4) that allows us to quickly identify a value of s and a corresponding inventory profile $\mathbf{Q}$ (Lemma 5 and Theorem 1), with guaranteed performance (Theorems 2 and 3).

5.1 Equal margins

We first consider the case where $r_i = r$ for all i. In this case, we can rewrite Equation (13) as:

$$R^{approx} = \max_{s, \mathbf{Q} \in \mathcal{Q}} rs/(1+s) \text{ s.t. } V(s, \mathbf{Q}) = s.$$

Since the objective function is increasing in s, the above optimization involves finding the highest s such that $V(s, \mathbf{Q}) = s$ for some $\mathbf{Q}$. That is, to find $\bar{s} = \max_{\mathbf{Q} \in \mathcal{Q}} s(\mathbf{Q})$, where $s(\mathbf{Q})$ is the unique solution to Equation (10). Finding $\bar{s}$ directly appears difficult; however we describe below an alternative approach to compute $\bar{s}$ efficiently. For any s, define

$$V(s) = \max_{\mathbf{Q} \in \mathcal{Q}} V(s, \mathbf{Q}). \tag{15}$$

By Lemma 1, $V(s, \mathbf{Q})$ is increasing in s. It follows that $V(s)$ is increasing in s. By Corollary 1, we also have that $V(s) \geq s$ if and only if $s \leq \bar{s}$. Thus, $\bar{s}$ is the unique solution to the equation $s = V(s)$. Moreover, we can use bisection to find $\bar{s}$:

- Initialize $s_l = 0$ and $s_h = \sum_{j=1}^{n} v_j$. Then $V(s_l) \geq s_l$ and $V(s_h) \leq s_h$.
- Let $s_m = (s_l + s_h)/2$. Compute $V(s_m)$. If $V(s_m) > s_m$, update s_l to be equal to s_m; otherwise update s_h to s_m.
- Repeat the above steps until $s_h - s_l$ is sufficiently small.

Note that in each iteration of the bisection procedure, the range $s_h - s_l$ gets divided by half and still $V(s_l) \geq s_l$ and $V(s_h) \leq s_h$. Both s_l and s_h converge to $\bar{s}$ such that $V(\bar{s}) = \bar{s}$. If $V(s)$ can be computed efficiently, then this bisection procedure allows us to find the solution quickly. It turns out that this is indeed possible since $V(s)$ can be obtained by solving a linear program (LP), as shown next.

Lemma 3. *For a given s, let $\delta_i(s,q) = a_i(s,q) - a_i(s,q-1)$ for $q = 1,\ldots,C$. $V(s)$ can be computed as the following linear program:*

$$\begin{aligned} V(s) = \max_{x} \quad & \sum_{i=1}^{n}\sum_{q=1}^{C} v_i \delta_i(s,q) x_{i,q} \\ s.t. \quad & 0 \le x_{i,q} \le 1 \text{ for all } i,q, \\ & \sum_{i=1}^{n}\sum_{q=1}^{C} x_{i,q} \le C. \end{aligned} \tag{16}$$

In the above LP, the decision variable $x_{i,q}$ indicates whether we stock-up on the qth unit of product i, and so $\sum_{q=1}^{C} x_{i,q}$ gives the total number of units of product i that we stock. Note that the solution to the above LP is integer since the constraint matrix is totally unimodular. Further, since $a_i(s,q)$ is increasing and concave in q, $\delta_i(s,q) \ge 0$ and decreasing in q. As a result, $x_{i,q+1} \le x_{i,q}$ in an optimal solution to the LP. That is, we stock up on the $q+1$th unit of product i only after we have picked the qth unit. Finally note that the above LP is a just single-dimensional knapsack problem and so its solution can, in fact, be obtained by simple inspection: the optimal solution is to pick the C highest values of $\{v_i\delta_i(s,q)\}$, which requires a complexity of $O(nC)$. We note that in practice, the store capacity C is $O(n)$ so that the overall complexity is polynomial in n.

Intuitively, this procedure seeks the combination of $(Q_1,\ldots,Q_n)$ that results in the highest attractiveness $s(\mathbf{Q})$. This is done by properly adjusting inventory levels so that the contribution of the last inventory unit, $v_i\Big(a_i(s,Q_i) - a_i(s,Q_i-1)\Big)$ is balanced across products. If this was not the case, e.g., if $v_1\Big(a_1(s,Q_1) - a_1(s,Q_1-1)\Big) < v_2\Big(a_2(s,Q_2+1) - a_2(s,Q_2)\Big)$, it would be possible to remove one unit of product 1, and add one of product 2, which would increase $\sum_{i=1}^{n} v_i a_i(s,Q_i)$, and hence lead to a higher level of units sold thereby increasing profits for the retailer.

Hence, when margins are equal across products, it is possible to solve (13) to an accuracy of ϵ in polynomial time. Indeed, for $k \ge -\log(\epsilon)/\log(2)$, we have that $2^{-k} \le \epsilon$; thus, in k iterations of the bisection procedure, $s_h - s_l \le \left(\sum_{i=1}^{n} v_i\right)\epsilon$, which means that $rs_l/(1+s_l) \le R^{approx} \le rs_h/(1+s_h) \le r\left(s_l + \left(\sum_{i=1}^{n} v_i\right)\epsilon\right)/(1+s_l) \le rs_l/(1+s_l) + r\left(\sum_{i=1}^{n} v_i\right)\epsilon$. In other words, our procedure is a FPTAS for problem (13), with a complexity of $O\Big(-\log(\epsilon)nC\Big)$.

5.2 Unequal margins

When margins differ across products, then solving problem (12) becomes difficult. For one, the objective function does not simplify as it did for the equal margins case. Moreover, the optimal assortment may not have good structural properties. As the following example (adapted from Rusmevichientong et al. 2010) shows, the optimal assortment may not be revenue-ordered.

Example 1: We have four products ($n = 4$) where the margins, attractivenesses and replenishment rates are given in Table 1 below. The capacity constraint is $C = 2$. That is, we can carry at most

two units of inventory in total. Note that since we have very fast replenishments, if we stock even one unit of a product, it has an in-stock probability of nearly 100%, i.e., $a_i(s,1) \approx 1$ for all s and i. Consequently, there is very little benefit in stocking more than one unit of any product. So the problem essentially reduces to an assortment planning problem with cardinality constraints. It can be verified that the optimal solution is to stock one unit each of product 2 and product 4 (Rusmevichientong et al. 2010). Therefore, the optimal assortment is not revenue-ordered.

Product	Margin (r_i)	Attractiveness (v_i)	Replenishment rate (μ_i)
Product 1	9.5	0.2	30
Product 2	9.0	0.6	30
Product 3	7.0	0.3	30
Product 4	4.5	5.2	30

Table 1: Example where the optimal assortment is not revenue-ordered.

It is also not true that the optimal solution to problem (12) includes only those products which are in the optimal cardinality constrained assortment. To illustrate, in Example 1 if we change the replenishment rates of products 2 and 4 to 0.1 keeping everything else the same, then it can be verified that the optimal solution to problem (12) is to stock one unit each of products 1 and 3. In contrast, the optimal cardinality constrained assortment (which does not consider inventory replenishments) includes products 2 and 4.

As seen in the example above, problem (13) does not have a well-behaved structure and hence is difficult to solve. We propose a two-step approach to (i) generate the optimal inventory vector $\mathbf{Q}$ for a given s, and then (ii) find the value of s which maximizes the objective function. We begin by writing optimization problem (13) in a parametric form, depending on s:

$$Z(s) = \max_{\mathbf{Q}\in\mathcal{Q}} R(s,\mathbf{Q}) \text{ s.t. } V(s,\mathbf{Q}) = s \tag{17}$$

so that $R^{approx} = \max_{s\geq 0} Z(s)$. Note that the formulation in (17) now includes a constraint $V(s,\mathbf{Q}) = s$ in addition to the cardinality constraint $\sum_{i=1}^{n} Q_i \leq C$. This makes the integer program (17) difficult to solve for a given value of s, since it is an instance of a two-dimensional knapsack problem, known to be NP-hard (see for example Martello and Toth 1990). One could apply existing PTAS to solve this type of problem; however, the computational challenge comes from the fact that the problem needs to be solved for all possible values of s, making this approach unworkable. Furthermore, as Figure 7 shows, $Z(s)$ may not have useful structural properties and that can increase the burden associated with finding the optimal value of s.

As a result of these challenges, we propose a relaxation of (17) that provides two useful pieces of information: it generates an upper bound to R^{approx} and we can use its solution to construct

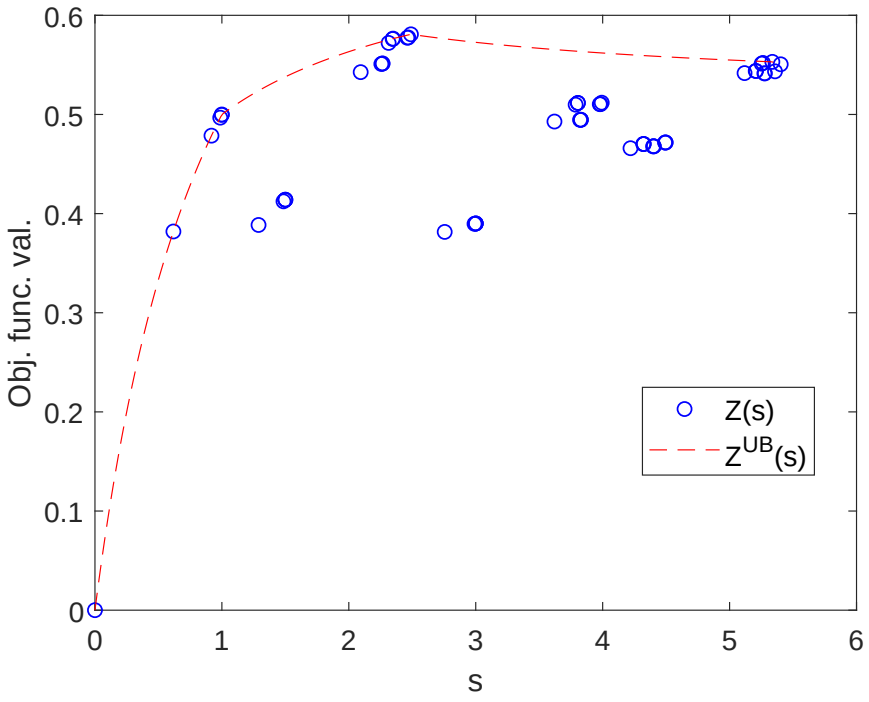

Figure 7: $Z(s)$ and $Z^{UB}(s)$ as a function of s. The problem parameters are $n = 3, C = 5, v_1 = 1, v_2 = 3, v_3 = 1.5$, $r_1 = 10, r_2 = 0.52, r_3 = 0.69$ and $\mu_1 = 1, \mu_2 = 9, \mu_3 = 4$. Note that the equality constraint in problem (17) can make it infeasible, in which case we set $Z(s) = -\infty$.

(integer) feasible solutions $\mathbf{Q}$ which have satisfactory performance. This procedure thus provides solutions that perform well, as shown in our numerical study in §6. Furthermore, the relaxation turns out to be a quasi-concave function of s, which simplifies the line-search procedure to find the optimal value of s. We begin by describing our relaxation $Z^{UB}(s)$.

Lemma 4. *For a given s, let $\delta_i(s,q) = a_i(s,q) - a_i(s,q-1) \geq 0$ for $q = 1, \ldots, C$. An upper bound of $Z(s)$ is provided by the following LP:*

$$\begin{aligned} Z^{UB}(s) = \max_{x \geq 0} \quad & \sum_{i=1}^{n} \sum_{q=1}^{C} \frac{r_i v_i \delta_i(s,q)}{1+s} x_{i,q} \\ s.t. \quad & x_{i,q} \leq 1 \text{ for all } i, q, \\ & \sum_{i=1}^{n} \sum_{q=1}^{C} x_{i,q} \leq C, \\ & \sum_{i=1}^{n} \sum_{q=1}^{C} v_i \delta_i(s,q) x_{i,q} = s. \end{aligned} \tag{18}$$

As in formulation (16), we can again interpret $x_{i,q}$ as indicating whether we stock up on the

qth unit of product i. However, now we have an additional constraint $\sum_{i=1}^{n}\sum_{q=1}^{C} v_i\delta_i(s,q)x_{i,q} = s$. Still it can be shown that there is an optimal solution to (18) that satisfies $x_{i,q+1} \leq x_{i,q}$ and we do not need to explicitly add these constraints to our formulation.

To gain insight into the structure of the solution to $Z^{UB}(s)$ we consider dualizing the last constraint in (16). In particular, we write the last constraint equivalently as $\sum_{i=1}^{n}\sum_{q=1}^{C} v_i\delta_i(s,q)x_{i,q}/(1+s) = s/(1+s)$ and associate a multiplier θ with the constraint to dualize it. Lemma 5 describes properties of the optimal dual multiplier and the optimal solution.

Lemma 5.

$$Z^{UB}(s) = \min_{\theta\in\mathbb{R}} \frac{\theta s + \hat{Z}(s,\theta)}{1+s} \tag{19}$$

where

$$\begin{aligned} \hat{Z}(s,\theta) = \max_{x\geq 0} \quad & \sum_{i=1}^{n}\sum_{q=1}^{C} (r_i-\theta)\, v_i\delta_i(s,q)x_{i,q} \\ s.t. \quad & x_{i,q} \leq 1 \text{ for all } i,q, \\ & \sum_{i=1}^{n}\sum_{q=1}^{C} x_{i,q} \leq C. \end{aligned} \tag{20}$$

Given θ, at any optimal solution x^* of (20), if $r_i < \theta$, then $x^*_{i,q} = 0$. Moreover, we have that $\hat{Z}(s,\theta)$ is a piecewise-linear convex function of θ. Furthermore, at an optimal solution θ^* in (19), two situations can arise:

1. *either there exists $\mathbf{Q} \in \mathcal{Q}$ such that $V(s,\mathbf{Q}) = s$ and*

$$Z^{UB}(s) = Z(s) = \frac{\sum_{j=1}^{n} r_j v_j a_j(s,Q_j)}{1+\sum_{j=1}^{n} v_j a_j(s,Q_j)};$$

2. *or there exists $\mathbf{Q}^l$ and $\mathbf{Q}^u$ in $\mathcal{Q}$ such that, $V(s,\mathbf{Q}^l) < s < V(s,\mathbf{Q}^u)$, $\hat{Z}(s,\theta^*) = F(s,\mathbf{Q}^l) - \theta^* V(s,\mathbf{Q}^l) = F(s,\mathbf{Q}^u) - \theta^* V(s,\mathbf{Q}^u)$ and*

$$\min\left\{\frac{F(s,\mathbf{Q}^l)}{1+V(s,\mathbf{Q}^l)}, \frac{F(s,\mathbf{Q}^u)}{1+V(s,\mathbf{Q}^u)}\right\} \leq Z^{UB}(s) \leq \max\left\{\frac{F(s,\mathbf{Q}^l)}{1+V(s,\mathbf{Q}^l)}, \frac{F(s,\mathbf{Q}^u)}{1+V(s,\mathbf{Q}^u)}\right\},$$

where $F(s,\mathbf{Q}) := \sum_{i=1}^{n} r_i v_i a_i(s,Q_i)$.

Hence, for a fixed s, we are able to characterize the upper bound by solving a sequence of LPs in the form of (20). Indeed, given θ, $\hat{Z}(s,\theta)$ is obtained by picking the C largest values of $(r_i-\theta)\, v_i\delta_i(s,q)$. Using the arguments of Rusmevichientong et al. (2010), it follows that it is sufficient to inspect the values of θ such that $(r_i-\theta)\, v_i\delta_i(s,q) = (r_j-\theta)\, v_j\delta_j(s,q')$ for all product-capacity pairs (i,q) and (j,q'), which is a set of cardinality $O(n^2C^2)$. This means that we can compute $Z^{UB}(s)$ in $O(n^3C^3)$.

Furthermore, we know that the upper bound is found in two possible ways. First, it can be achieved at an integer solution, so that $Z^{UB}(s) = Z(s)$ and there is no gap in the relaxation. Otherwise, it can be obtained by combining two integer solutions $\mathbf{Q}^l$ and $\mathbf{Q}^u$. Indeed, because the constraints in (18) are all integral except for the last one ($\sum_{i=1}^{n}\sum_{q=1}^{C} v_i\delta_i(s,q)x_{i,q} = s$), any optimal solution to (18) will have at most two fractional entries (which sum up to 1). Therefore, this fractional solution can be rounded up or down to obtain feasible solutions $\mathbf{Q}^l$ and $\mathbf{Q}^u$. For these two integer solutions, we no longer guarantee that $V(s,\mathbf{Q}) = s$. But they generate two lower bounds to R^{approx}, namely $R\Big(s(\mathbf{Q}^l),\mathbf{Q}^l\Big)$ and $R\Big(s(\mathbf{Q}^u),\mathbf{Q}^u\Big)$, where $R(s,\mathbf{Q})$ is given by (14). These complement the upper bound.

Now that we have understood how to compute $Z^{UB}(s)$ for a given value of s, we next look at finding the value of s that maximizes $Z^{UB}(s)$. It turns out that $Z^{UB}(s)$ is well-behaved, as shown next.

Theorem 1. *$Z^{UB}(s)$ is quasi-concave in s.*

Theorem 1 implies that $Z^{UB}(s)$ is unimodal in s. It hence allows us to use specialized search methods such as bisection or Fibonacci search (Chong and Zak 2001) to find the optimal solution. As for the equal margin case, such methods lead to a FPTAS to maximize $Z^{UB}(s)$, with a complexity of $O\Big(-\log(\epsilon)n^3C^3\Big)$. Once we have found the maximizer, s^{UB}, of $Z^{UB}(s)$, we can identify the integer solutions $\mathbf{Q}^l$ and $\mathbf{Q}^u$ described in Lemma 5, and pick the one which gives the highest objective value. As the following example illustrates, this can lead to an integrality gap.

Example 2: We have three products ($n = 3$) where the margins, attractivenesses and replenishment rates are given in Table 2 below. The capacity constraint is $C = 1$ and so we can carry at most one unit of inventory in total. In this case, it can be verified that $Z(s)$ is maximized at $s^* = 1.2892$. The corresponding optimal inventory vector $\mathbf{Q}^* = (0,0,1)$ and $R^{approx} = 0.38848$. However, $Z^{UB}(s)$ is maximized at $s^{UB} = 1.3218$ and $Z^{UB}(s^{UB}) = 0.39377$. Therefore, we have $\max_{s\geq 0} Z(s) < \max_{s\geq 0} Z^{UB}(s)$. Furthermore, the solution at s^{UB} is made up of a combination of 0.676 of $\mathbf{Q}^l = (1,0,0)$ and 0.324 of $\mathbf{Q}^u = (0,1,0)$. It can be verified that $R(s^l,\mathbf{Q}^l) = 0.38194$ and $R(s^u,\mathbf{Q}^u) = 0.38152$, where s^l and s^u, respectively, satisfy Equation (10) for $\mathbf{Q}^l$ and $\mathbf{Q}^u$. Therefore, rounding up or rounding down the fractional solution at s^{UB} does not give us the optimal integer solution. Therefore, in this case our procedure results in a sub-optimal solution. The example is illustrated in Figure 8.

Even though our procedure does not return the optimal solution in general, it achieves remarkable performance, as shown in §6. In addition, we are able to offer worst-case performance guarantees on the integrality gap, as shown next.

Theorem 2. *Let $\gamma = \max_j\left\{\frac{v_j}{\mu_j}\right\}$ and let $\mathbf{Q}^l$ and $\mathbf{Q}^u$ be as defined in case (2) of Lemma 5. We*

Product	Margin (r_i)	Attractiveness (v_i)	Replenishment rate (μ_i)
Product 1	1.00	1	1
Product 2	0.52	3	9
Product 3	0.69	1.5	4

Table 2: Example where there is an integrality gap.

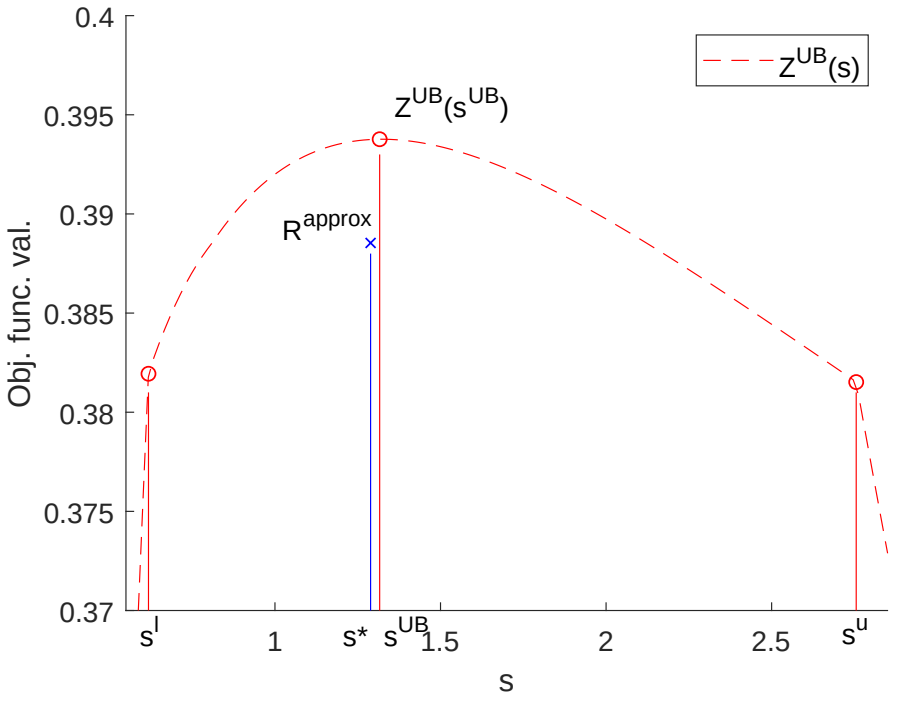

Figure 8: Example where $\max_{s\geq 0} Z^{UB}(s) > \max_{s\geq 0} Z(s) = R^{approx}$. In this example, $n = 3$, $C = 1$ and product parameters are taken from Table 2.

have

$$Z^{UB}(s) \leq \left(1 + \frac{\min\{\gamma s^u, 1 + s^u\}}{1 + s^l}\right) \max\left\{R(s^l, \mathbf{Q}^l), R(s^u, \mathbf{Q}^u)\right\},$$

where s^l *and* s^u *are, respectively, the values of* s *that satisfy Equation* (10) *for* $\mathbf{Q}^l$ *and* $\mathbf{Q}^u$.

Theorem 2 shows that the gap between the parametric integer program $\max_{s\geq 0} Z(s)$ and its continuous relaxation $\max_{s\geq 0} Z^{UB}(s)$ is small when s^u is small. Even if s^u is large, the gap can still be small if s^u/s^l is close to one. The gap is also small if $\gamma = \max_j\{v_j/\mu_j\}$ is small, which happens when the products are replenished quickly compared to the demand rate.

The bound in Theorem 2 relies on the optimality conditions associated with s^{UB}, and hence

depends on the problem parameters. It provides insight into when the integrality gap is small. We close this section by showing that the worst-case integrality gap is bounded by a factor of 2, irrespective of the problem parameters, and based on a different proof technique. Theorem 3 below shows that it is possible to obtain a bound of 2 on the integrality gap by tweaking the linear programming formulation $Z^{UB}(s)$ to only include products that satisfy $v_i a_i(s,1) \leq s$. That is, the product by itself fits into a knapsack of size s. The constant factor bound thus complements the parametric bound by showing that the gap cannot be arbitrarily large.

Theorem 3. *We have* $\max_{s\geq 0} Z^{UB}(s) \leq 2R^{approx}$.

6. Numerical Study

In this section we numerically test the performance of our solution method. We first compare the upper bounds obtained by the continuous linear programming relaxation with the integer programming formulation to understand the magnitude of the integrality gap. Then we test the revenue performance of the stocking levels generated by our solution method. Finally we test the sensitivity of the solution to the store capacity, the average lead times and the distribution of the lead times.

6.1 Performance of the continuous relaxation vs. integer optimum

We generate our test problems in the following manner: we vary $n \in \{25, 50, 100, 200\}$ and $C \in \{25, 50, 100, 200\}$. We set $\mu_i = \mu$ for all products i and vary $\mu \in \{0.05, 1\}$. Note that when μ is small, the lead times tend to be longer and the products get replenished at a slower rate. For each (n, C, μ) combination we generate thirty test problems by randomly sampling v_i from the uniform distribution on $[1/10, 10]$ and r_i from the uniform distribution on $[1, 10]$.

We compare the upper bound obtained by our continuous relaxation, $\max_{s\geq 0} Z^{UB}(s)$, with the integer solution $R^{approx} = \max_{s\geq 0} Z(s)$. By Theorem 1, $Z^{UB}(s)$ is unimodal in s. So we use a variant of the bisection method to search for the maximizer (Chong and Zak 2001) and stop the algorithm when the width of the search interval is smaller than 10^{-6}. Since $Z(s)$ does not have a similar structure, we search for the maximizer of $Z(s)$ by dividing the interval $[0, \sum_{i=1}^{n} v_i]$ into a finite number of grid points and evaluating the function at each grid point. As we increase the number of grid points we obtain more accurate solutions but at the expense of larger computation times. Based on initial setup runs, we found that having 400 grid points provided a good balance between the accuracy and the solution time.

Table 3 shows the optimality gaps and the CPU times for the test problems with $\mu = 0.05$. The first column describes the problem characteristics using (n, C, μ). The second column gives

the average percentage difference between the upper bound obtained by the continuous relaxation and the integer solution. The third column gives the maximum percentage difference whereas the fourth column shows the percentage of test problems where the upper bound is within 0.1% of the integer solution. The last two columns give the average computation times for the integer program and its linear programming relaxation. All of our computational experiments are carried out on a Core i7 desktop with 3.4GHz CPU and 16-GB RAM. We use GUROBI 8.0.1 to solve the linear and integer programs.

We observe that the optimality gap is generally within a fraction of a percentage, and the performance tends to be particularly good when the capacity C is large relative to the number of products n. Furthermore, the continuous relaxation provides noticeable savings in computation time for problems with a large number of products and large capacities.

Table 4 shows the optimality gaps and the CPU times for the test problems with $\mu = 1$. Similar to before, we see that the continuous relaxation obtains solutions that are near optimal. Interestingly we observe that for a fixed n, the optimality gaps do not change with the capacity C. It turns out that for $\mu = 1$, the products get replenished at a relatively fast rate and so the approximate in-stock probabilities ($a_i(s, Q)$) tend to one very quickly. That is, the in-stock probabilities are close to 1 even for low stocking levels and the marginal benefit from stocking additional units (when C increases) is minimal. Therefore, the objective function does not change very much when C increases and we are allowed to stock more units.

6.2 Revenue performance for small instances

We next test the revenue performance of our solution method. We first consider small instances where it is feasible to carry out complete enumeration and evaluate the revenues exactly. This allows us to accurately benchmark the quality of our solution method. We generate our test problems in the same manner as before except that we vary $n \in \{2, 3, 4\}$ and $C \in \{10, 20, 30\}$. For a given test problem, we obtain the steady state probabilities associated with an inventory vector $\mathbf{Q}$ by solving Equation (3) and then compute the corresponding expected revenue. We repeat this process for all $\mathbf{Q} \in \mathcal{Q}$ to obtain the optimal expected revenue. We compare this with the revenue obtained by our solution method: Letting $s^{UB} := \text{argmax}_{s \geq 0} Z^{UB}(s)$, recall that by Lemma 5, the solution to $Z^{UB}(s^{UB})$ can have at most two fractional values, which can be rounded up and down to yield integer stocking levels. We compute the expected revenues (for the original Markov chain) generated by the two integer stocking levels and take the higher of the two as the expected revenue obtained by our solution method.

In order to assess the value of jointly making the assortment and inventory decisions, we also benchmark the performance of our solution method with a heuristic that makes the assortment

Problem	% opt. gap			CPU secs.	
(n, C, μ)	Avg.	Max.	% optimal	IP	LP
(25, 25, 0.05)	0.13	0.34	57	4.37	0.30
(25, 50, 0.05)	0.05	0.22	83	8.27	0.38
(25, 100, 0.05)	0.05	0.22	83	10.86	0.54
(25, 200, 0.05)	0.05	0.22	83	15.27	0.89
(50, 25, 0.05)	0.09	0.32	70	5.27	0.40
(50, 50, 0.05)	0.04	0.13	87	9.98	0.57
(50, 100, 0.05)	0.04	0.13	87	16.22	0.96
(50, 200, 0.05)	0.04	0.13	87	25.19	1.59
(100, 25, 0.05)	0.08	0.20	67	7.42	0.66
(100, 50, 0.05)	0.02	0.13	97	14.37	0.99
(100, 100, 0.05)	0.02	0.14	97	23.80	1.62
(100, 200, 0.05)	0.02	0.14	97	45.65	3.08
(200, 25, 0.05)	0.05	0.11	97	12.76	1.23
(200, 50, 0.05)	0.02	0.06	100	37.49	2.92
(200, 200, 0.05)	0.01	0.04	100	50.59	3.64
(200, 200, 0.05)	0.01	0.04	100	87.08	7.62

Table 3: Optimality gaps and CPU times for the test problems with $\mu = 0.05$.

Problem	% opt. gap			CPU secs.	
(n, C, μ)	Avg.	Max.	% optimal	IP	LP
(25, 25, 1)	0.13	0.54	57	5.07	0.32
(25, 50, 1)	0.13	0.54	57	5.97	0.41
(25, 100, 1)	0.13	0.54	57	7.69	0.58
(25, 200, 1)	0.13	0.54	57	12.51	0.98
(50, 25, 1)	0.06	0.28	77	5.22	0.38
(50, 50, 1)	0.06	0.28	77	8.11	0.55
(50, 100, 1)	0.06	0.28	77	11.71	0.89
(50, 200, 1)	0.06	0.28	77	20.74	1.58
(100, 25, 1)	0.06	0.29	83	6.59	0.57
(100, 50, 1)	0.06	0.29	83	11.77	0.94
(100, 100, 1)	0.06	0.29	83	21.65	1.63
(100, 200, 1)	0.06	0.29	83	41.91	3.08
(200, 25, 1)	0.02	0.11	97	9.81	1.00
(200, 50, 1)	0.02	0.11	97	19.64	1.74
(200, 200, 1)	0.02	0.11	97	46.96	3.63
(200, 200, 1)	0.02	0.11	97	89.41	7.77

Table 4: Optimality gaps and CPU times for the test problems with $\mu = 1$.

and inventory decisions in a sequential manner. The sequential heuristic works as follows: We first solve problem

$$R^{assort} = \max_{\mathbf{x}\in\{0,1\}^n} \frac{\sum_{i=1}^n r_i v_i x_i}{1+\sum_{i=1}^n v_i x_i}, \tag{21}$$

to decide which products to include in the assortment. We then make the inventory allocation decisions in the following fashion, inspired by the D'Hondt method (**?**) : Letting $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$ denote the optimal solution to (21), we can interpret $d_i = \frac{1}{\mu_i} \frac{v_i x_i^*}{1+\sum_{j=1}^n v_j x_j^*}$ as the expected demand for product i over its lead time. Accordingly, the fraction of capacity C allocated to product i should be $\frac{d_i}{\sum_{j=1}^n d_j} C$. However this can be a fractional number. In order to get a feasible solution, we allocate inventory in the following manner: we initialize $q_i = 0$ for all products i. We allocate the first unit of inventory to the product which attains the highest value of $d_i/(1+q_i)$, update the allocations q_i and repeat this process until we have allocated all the C units.

Table 5 reports the revenue gaps. The second and third columns, respectively, report the average and the maximum percent difference (over the 30 randomly generated instances) between the optimal expected revenue and that obtained by our solution method. The last two columns give the corresponding numbers for the heuristic method which makes the assortment and inventory decisions sequentially. We note that a gap of zero simply indicates that the number is less than 0.01%. We notice that the sequential allocation heuristic can perform especially poorly in instances where the store capacity is small and the lead times are long. On the other hand, our solution method obtains revenues that are near optimal and the revenue gap is typically a fraction of a percent. The performance of our solution method is better for $\mu = 1$ compared to $\mu = 0.05$. Recall that as μ increases, the products get replenished at a fast rate and we rarely have stockouts. Consequently, we expect both the approximate and exact in-stock probabilities to be close to 1 and to each other (see the plots on the right side of Figures 4 and 5). Therefore, we expect our approximation to perform better for relatively larger values of μ.

6.3 Revenue performance for larger instances

We next test the revenue performance of our solution method on test problems with a larger number of products and larger values of capacity. We consider $n \in \{50, 100, 200\}$, $C \in \{50, 100, 200\}$ and $\mu \in \{0.05, 1\}$. For each (n, C, μ) combination we generate ten test problems by randomly sampling v_i from the uniform distribution on $[1/10, 10]$ and r_i from the uniform distribution on $[1, 10]$. For test problems of this size, it becomes impractical to evaluate the steady state probabilities by solving Equation (3) and use that to compute the expected revenue associated with a given inventory vector. Therefore, we use simulation instead and generate a stream of 10,000 customer arrivals and take the average revenue obtained from the last 5,000 customers. We repeat this exercise 10 times and take the average to estimate the expected revenue associated with an inventory vector.

Problem	% opt. gap - Joint		% opt. gap - Seq	
(n, C, μ)	Avg.	Max.	Avg.	Max.
(2, 10, 0.05)	0.04	1.04	3.85	39.86
(2, 20, 0.05)	0.05	0.96	0.99	9.72
(2, 30, 0.05)	0.01	0.15	0.38	4.18
(3, 10, 0.05)	0.00	0.10	2.80	31.98
(3, 20, 0.05)	0.12	2.16	0.73	9.16
(3, 30, 0.05)	0.00	0.01	0.28	4.79
(4, 10, 0.05)	0.00	0.01	0.81	6.06
(4, 20, 0.05)	0.05	0.68	0.18	1.90
(4, 30, 0.05)	0.00	0.04	0.04	0.24

Problem	% opt. gap - Joint		% opt. gap - Seq	
(n, C, μ)	Avg.	Max.	Avg.	Max.
(2, 10, 1)	0.00	0.01	0.72	14.78
(2, 20, 1)	0.00	0.01	0.18	4.39
(2, 30, 1)	0.00	0.01	0.15	4.39
(3, 10, 1)	0.00	0.01	0.50	8.09
(3, 20, 1)	0.00	0.01	0.33	5.78
(3, 30, 1)	0.00	0.01	0.14	3.83
(4, 10, 1)	0.00	0.01	0.10	1.06
(4, 20, 1)	0.00	0.01	0.06	1.06
(4, 30, 1)	0.00	0.01	0.00	0.04

Table 5: Revenue gaps for the test problems with small n and C.

As it also becomes impractical to enumerate over all feasible stocking levels, we restrict our attention to inventory vectors that are in a small neighborhood of our solution. Given a stocking level $\hat{\mathbf{Q}}$, we define its neighborhood to be vectors $\mathbf{Q}$ such that $Q_i = \max\{\hat{Q}_i - 1, 0\}, Q_j = \min\{\hat{Q}_j + 1, C\}$ for some i, j and $Q_k = \hat{Q}_k$ for $k \neq i, j$. Therefore, the neighborhood of a vector $\hat{\mathbf{Q}}$ consists of all vectors $\mathbf{Q}$ that differ from $\hat{\mathbf{Q}}$ in at most two components. Given the integer stocking levels obtained by rounding the solution to $Z^{UB}(\hat{s})$, we randomly sample 250 inventory vectors from the neighborhood of these solutions and estimate the corresponding expected revenues.

Table 6 compares the revenue performance of our solution method and the heuristic that makes the assortment and inventory decisions sequentially. The second and third columns, respectively, give the average and maximum gaps between the expected revenue obtained by the best neighboring solution and our solution, while the last two columns show the same metrics for the sequential heuristic. The gaps are within a fraction of a percent for both methods, although our solution method has a slight advantage for the more capacity constrained cases.

6.4 Sensitivity analysis

In this section, we vary some of the key input parameters including the store capacity and the lead times to further illustrate how our approach can be used to inform the assortment and inventory decisions. We also test the sensitivity of our solution method to the assumption that the product lead times are exponentially distributed.

Varying the store capacity: One possible application of our algorithm is to find the optimal inventory distribution for different levels of store capacity C. This piece of information can be quite useful for retailers that run different store sizes. Specifically, our approach can prescribe how store size should change breadth and depth. Most retailers group different stores in clusters so that

Problem	% opt. gap - Joint		% opt. gap - Seq	
(n, C, μ)	Avg.	Max.	Avg.	Max.
(50, 50, 0.05)	0.03	0.13	0.10	0.29
(50, 100, 0.05)	0.04	0.12	0.05	0.15
(50, 200, 0.05)	0.04	0.14	0.04	0.16
(100, 50, 0.05)	0.02	0.07	0.08	0.17
(100, 100, 0.05)	0.02	0.07	0.02	0.08
(100, 200, 0.05)	0.02	0.07	0.02	0.08
(200, 50, 0.05)	0.02	0.06	0.07	0.21
(200, 100, 0.05)	0.02	0.04	0.03	0.07
(200, 200, 0.05)	0.02	0.04	0.02	0.04

Problem	% opt. gap - Joint		% opt. gap - Seq	
(n, C, μ)	Avg.	Max.	Avg.	Max.
(50, 50, 1)	0.01	0.12	0.05	0.25
(50, 100, 1)	0.03	0.22	0.03	0.25
(50, 200, 1)	0.02	0.18	0.02	0.21
(100, 50, 1)	0.01	0.10	0.03	0.15
(100, 100, 1)	0.02	0.21	0.02	0.22
(100, 200, 1)	0.02	0.24	0.02	0.25
(200, 50, 1)	0.01	0.07	0.03	0.14
(200, 100, 1)	0.01	0.06	0.02	0.09
(200, 200, 1)	0.02	0.09	0.02	0.09

Table 6: Revenue gaps for the test problems with large n and C.

the assortment is the same in each cluster (and inventory is adjusted to fit in the store). With our method, it is no longer necessary to use the same assortment across different stores: since our algorithm is quite efficient, we can have a different assortment for each store. To illustrate this, we show in Figure 9 how optimal solutions vary with capacity, in a three-product case (larger n examples yield similar insights).

We can see from the figure that, as C increases, the retailer first introduces product 1 which is the most profitable product; then introduces 2 and 3. When $C > 4$, the retailer finds it beneficial to keep increasing product 3: while this is the least profitable product, it also has the highest demand ($v_3 > \max\{v_1, v_2\}$) and the slowest replenishment time ($\mu_3 < \min\{\mu_1, \mu_2\}$), which implies that any additional unit of inventory of 3 increases the service level more than in other products.

Note that in the figure, the optimal solution displays a nested structure and the stocking level of a product is non-decreasing in C. Hence when going from C to $C+1$ (in this example) one only needs to identify which of the three products gets an additional unit of inventory. However, this is not necessarily true in general, as also shown in Rusmevichientong et al. (2010) (for the cardinality constrained assortment optimization problem). If we go back to Example 1, for $C = 2$, the optimal solution is to set $Q_1 = 0, Q_2 = 1, Q_3 = 0$ and $Q_4 = 1$. However, when $C = 3$, it can be verified that the optimal solution is to set $Q_1 = Q_2 = Q_3 = 1$ and $Q_4 = 0$. Therefore, the inventory level of a product is not necessarily monotone in the capacity C. This further illustrates the utility of our solution method since simple rules of thumb are unlikely to work in general.

Varying the average lead time: We next vary the mean lead time to understand its effect on the trade-off between assortment breadth and inventory depth. This may be useful, for example, to a retailer to assess the impact of a more responsive supply chain. To illustrate, we show in Figure 10 how the optimal solutions vary with the mean lead time. In this example, we have $n = 10$ products,

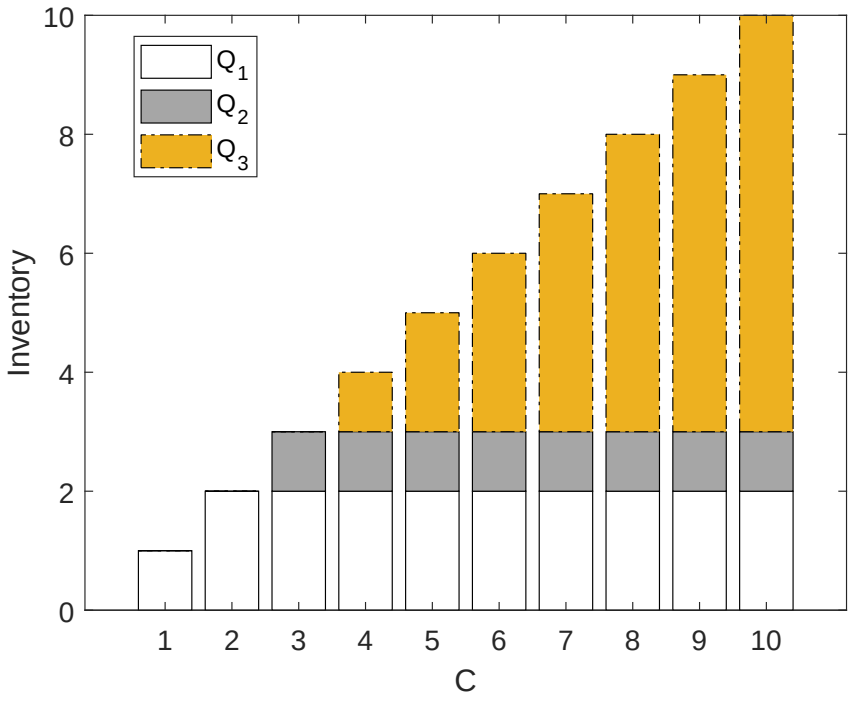

Figure 9: Optimal solution as C varies. In this example, $n = 3$; $r_1 = 0.80, r_2 = 0.55, r_3 = 0.52$; $v_1 = 0.54, v_2 = 0.29, v_3 = 0.87$; $\mu_1 = 0.40, \mu_2 = 0.67, \mu_3 = 0.17$.

store capacity $C = 20$ and the same average lead time for all the products ($\mu_i = \mu$ for all i). From the figure we see that for low values of μ (long lead times), the store stocks only the highest margin product (product 1). As μ increases, the marginal benefit from stocking additional units of product 1 decreases and the store begins to offer a more varied assortment: it starts stocking up on product 2 followed by products 4, 5 and 3. Note that the increased breadth is accompanied by a decrease in the inventory levels of the previously stocked products. This is because shorter lead-times "save" space in the store and allow the retailer to offer more products without sacrificing the sales of the existing items. Notice also that the order in which products are added to the assortment is not necessarily according to their margins: product 3 has a higher margin than products 4 and 5, but gets added to the assortment only later. Finally when $\mu = 0.2$, product 6 also gets added to the assortment and now the assortment coincides with the optimal MNL assortment, identified in problem (21).

Deterministic lead times: Finally, we relax the assumption that the lead times are exponentially distributed and consider a replenishment process where the product lead times are deterministic and $1/\mu_i$ denotes the number of time units needed to receive a replenishment order of product i.

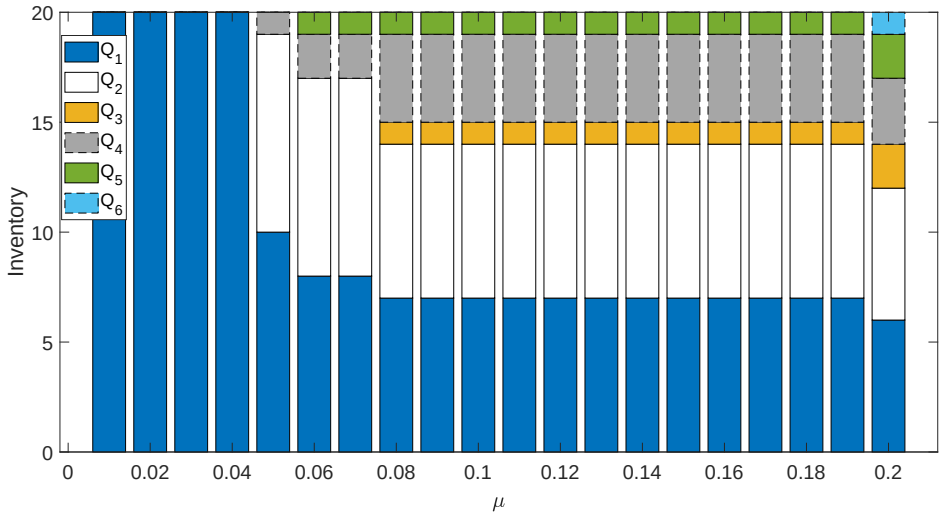

Figure 10: Optimal solution as μ varies. In this example, $n = 10$; $C = 20$; $r_1 = 8.91, r_2 = 7.17, r_3 = 7.11, r_4 = 7.10, r_5 = 7.08, r_6 = 7.03, r_7 = 2.84, r_8 = 2.78, r_9 = 2.26, r_{10} = 1.25$; $v_1 = 3.09, v_2 = 7.23, v_3 = 1.94, v_4 = 4.22, v_5 = 3.52, v_6 = 1.01, v_7 = 0.10, v_8 = 5.43, v_9 = 4.03, v_{10} = 1.55$.

Now the model described in §3 is misspecified in that there is a discrepancy between the actual replenishment process (deterministic lead time) and that assumed by the model (exponential lead time). Still, we use our model to obtain a candidate solution by using $1/\mu_i$ as the mean value of the exponential distribution. We then test the revenue performance of this candidate solution in the actual system. With deterministic lead times, it becomes difficult to evaluate the expected revenue analytically even for small instances. We therefore resort to simulation and estimate the expected revenues following the procedure described in §6.3. Table 7 reports the revenue performance for the small instances where we benchmark the performance of our solution method against the optimal solution (as described in §6.2). Table 8 reports the performance of our solution method for the large instances where we benchmark the performance of our solution method against the best neighboring solution (as described in §6.3). It is encouraging that the gaps are still within a fraction of a percent on average and our solution method continues to perform well.

7. Conclusions and Further Research

In this paper, we have develop a tractable decision model that allows retailers to coordinate assortment breadth and inventory depth decisions. Our framework is grounded on an accurate approximation of the steady-state probabilities of the inventory vector under stockout-based substitution, which allows us to formulate the decision problem in a static, compact fashion. This formulation

Problem	% opt. gap		Problem	% opt. gap	
(n, C, μ)	Avg.	Max.	(n, C, μ)	Avg.	Max.
(2, 10, 0.05)	0.06	1.10	(2, 10, 1)	0.00	0.01
(2, 20, 0.05)	0.06	1.01	(2, 20, 1)	0.00	0.01
(2, 30, 0.05)	0.01	0.14	(2, 30, 1)	0.00	0.01
(3, 10, 0.05)	0.00	0.00	(3, 10, 1)	0.00	0.01
(3, 20, 0.05)	0.16	1.09	(3, 20, 1)	0.00	0.01
(3, 30, 0.05)	0.01	0.05	(3, 30, 1)	0.00	0.01
(4, 10, 0.05)	0.00	0.00	(4, 10, 1)	0.00	0.01
(4, 20, 0.05)	0.26	2.19	(4, 20, 1)	0.00	0.01
(4, 30, 0.05)	0.00	0.01	(4, 30, 1)	0.00	0.01

Table 7: Revenue gaps for the test problems with deterministic lead times, for small n and C.

Problem	% opt. gap		Problem	% opt. gap	
(n, C, μ)	Avg.	Max.	(n, C, μ)	Avg.	Max.
(50, 50, 0.05)	0.08	0.29	(50, 50, 1)	0.06	0.16
(50, 100, 0.05)	0.09	0.29	(50, 100, 1)	0.05	0.18
(50, 200, 0.05)	0.01	0.05	(50, 200, 1)	0.00	0.04
(100, 50, 0.05)	0.09	0.29	(100, 50, 1)	0.06	0.21
(100, 100, 0.05)	0.06	0.15	(100, 100, 1)	0.07	0.24
(100, 200, 0.05)	0.01	0.04	(100, 200, 1)	0.05	0.11
(200, 50, 0.05)	0.03	0.13	(200, 50, 1)	0.05	0.17
(200, 100, 0.05)	0.05	0.24	(200, 100, 1)	0.05	0.09
(200, 200, 0.05)	0.10	0.27	(200, 200, 1)	0.03	0.07

Table 8: Revenue gaps for the test problems with deterministic lead times, for large n and C.

includes an auxiliary variable s which represents the average attractiveness of the assortment, and involves a constraint that is a fixed-point equation on s. We are able to solve this problem to optimality when all items have the same margin, and provide an effective heuristic in the general case, for which we offer performance guarantees. In the general case, the algorithm has a competitive running time, and returns a solution within a factor $1+\epsilon$ of the optimum, in $O\Big(n^3C^3\log(1/\epsilon)\Big)$, while the complexity is $O\Big(nC\log(1/\epsilon)\Big)$ when margins are all equal. Finally, we provide a numerical study where our approach is tested, and show that on random instances our approach finds the optimal solution in more than 80% of the cases, and is within 0.1% of the optimal profit.

Our research opens a number of follow-up questions for study. First, the main challenge under stockout-based substitution is to accurately model the probabilities of being in a given inventory state. We develop an approximation of the dynamics of the system that decouples a n-dimensional Markov chain into n independent one-dimensional Markov chains. This idea can be used in other settings where state dynamics are complex and couple different systems, like queuing systems where the demand rate of a given queue depends on the length of other queues, or markets where supply and demand in a given place depend on prices of nearby markets (i.e., spatial competition models). Second, our static formulation of the assortment problem in (12) includes a fixed-point equation as a constraint. We can expect this type of constraint to appear in situations where one of the demand parameters is endogenous to the retailer's decision, such as rational expectation models (Su and Zhang 2008) or network effects (Du et al. 2016, Wang and Wang 2016). Our approach of introducing an auxiliary variable s and solving a sequence of easier parametric problems is a promising approach for difficult assortment problems, used in Rusmevichientong et al. (2010) or Kunnumkal and Martínez-de Albéniz (2019) for instance. Finally, our model relies on the MNL specification for choice. A challenging yet valuable question would be to extend our breadth-depth decision model to situations where demand follows a more general demand process, such as the Markov chain choice (Blanchet et al. 2016, Feldman and Topaloglu 2017).

References

Aouad, A., R. Levi, and D. Segev. 2018. Greedy-Like Algorithms for Dynamic Assortment Planning under Multinomial Logit Preferences. *Operations Research* 66 (5): 1321–1345.

Axsäter, S. 2006. *Inventory control*, Volume 90. Springer Verlag, New York.

Beyer, D., S. P. Sethi, and R. Sridhar. 2001. Stochastic multiproduct inventory models with limited storage. *Journal of Optimization Theory and Applications* 111 (3): 553–588.

Blanchet, J., G. Gallego, and V. Goyal. 2016. A markov chain approximation to choice modeling. *Operations Research* 64 (4): 886–905.

Boada-Collado, P., and V. Martínez-de Albéniz. 2020. Estimating and optimizing the impact of inventory on consumer choices in a fashion retail setting. *Manufacturing & Service Operations Management* 22 (3): 582–597.

Boyd, S., and L. Vandenberghe. 2004. *Convex optimization*. Cambridge University Press, Cambridge, UK.

Cachon, G. P., S. Gallino, and M. Olivares. 2019. Does adding inventory increase sales? Evidence of a scarcity effect in US automobile dealerships. *Management Science* 65 (4): 1469–1485.

Cachon, G. P., C. Terwiesch, and Y. Xu. 2005. Retail assortment planning in the presence of consumer search. *Manufacturing & Service Operations Management* 7 (4): 330–346.

Caro, F., and J. Gallien. 2007. Dynamic assortment with demand learning for seasonal consumer goods. *Management Science* 53 (2): 276–292.

Cho, G. E., and C. D. Meyer. 2001. Comparison of perturbation bounds for the stationary distribution of a Markov chain. *Linear Algebra and its Applications* 335:137–150.

Chong, E., and S. Zak. 2001. *An introduction to optimization*. John Wiley & Sons, New York NY.

Davis, J. M., G. Gallego, and H. Topaloglu. 2013. Assortment planning under the multinomial logit model with totally unimodular constraint structures. Working paper, Cornell University.

Du, C., W. L. Cooper, and Z. Wang. 2016. Optimal pricing for a multinomial logit choice model with network effects. *Operations Research* 64 (2): 441–455.

Farahat, A., and J. Lee. 2017. The Multiproduct Newsvendor Problem with Customer Choice. *Operations Research* 66 (1): 123–136.

Feldman, J. B., and H. Topaloglu. 2017. Revenue management under the markov chain choice model. *Operations Research* 65 (5): 1322–1342.

Gaur, V., and D. Honhon. 2006. Assortment Planning and Inventory Decisions under a Locational Choice Model. *Management Science* 52 (10): 1528–1543.

Glasserman, P. 1994. Perturbation Analysis of Production Networks. In *Stochastic Modeling and Analysis of Manufacturing Systems, Chapter 6*, ed. D. D. Yao, 233–280. Springer, New York.

Heese, H. S., and V. Martínez-de Albéniz. 2018. Effects of assortment breadth announcements on manufacturer competition. *Manufacturing & Service Operations Management* 20 (2): 302–316.

Honhon, D., V. Gaur, and S. Seshadri. 2010. Assortment planning and inventory decisions under stockout-based substitution. *Operations Research* 58 (5): 1364–1379.

Hübner, A. H., and H. Kuhn. 2012. Retail category management: State-of-the-art review of quantitative research and software applications in assortment and shelf space management. *Omega* 40 (2): 199–209.

Kaplan, R. 1970. A dynamic inventory model with stochastic lead times. *Management Science* 16 (7): 491–507.

Kök, A. G., and M. L. Fisher. 2007. Demand estimation and assortment optimization under substitution: Methodology and application. *Operations Research* 55 (6): 1001–1021.

Kök, A. G., M. L. Fisher, and R. Vaidyanathan. 2009. Assortment planning: Review of literature and industry practice. In *Retail supply chain management*, 99–153. Springer.

Kunnumkal, S., and V. Martínez-de Albéniz. 2019. Tractable Approximations for Assortment Planning with Product Costs. *Operations Research* 67 (2): 436–452.

Lippman, S. A., and K. F. McCardle. 1997. The competitive newsboy. *Operations Research* 45 (1): 54–65.

Mahajan, S., and G. van Ryzin. 2001. Stocking retail assortments under dynamic consumer substitution. *Operations Research* 49 (3): 334–351.

Martello, S., and P. Toth. 1990. *Knapsack problems: Algorithms and computer implementations*. John Wiley & Sons, England.

Muharremoglu, A., and N. Yang. 2010. Inventory management with an exogenous supply process. *Operations Research* 58 (1): 111–129.

Musalem, A., M. Olivares, E. Bradlow, C. Terwiesch, and D. Corsten. 2010. Structural Estimation of the Effect of Out-of-Stocks. *Management Science* 56 (7): 1180–1197.

Netessine, S., and N. Rudi. 2003. Centralized and competitive inventory models with demand substitution. *Operations Research* 51 (2): 329–335.

O'Cinneide, C. A. 1993. Entrywise perturbation theory and error analysis for Markov chains. *Numerische Mathematik* 65:109–120.

Rusmevichientong, P., Z.-J. M. Shen, and D. B. Shmoys. 2010. Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations Research* 58 (6): 1666–1680.

Smith, S. A., and N. Agrawal. 2000. Management of multi-item retail inventory systems with demand substitution. *Operations Research* 48 (1): 50–64.

Su, X., and F. Zhang. 2008. Strategic Customer Behavior, Commitment, and Supply Chain Performance. *Management Science* 54 (10): 1759–1773.

Talluri, K., and G. van Ryzin. 2004. Revenue Management under a general discrete choice model of consumer behavior. *Management Science* 50 (1): 15–33.

Tsay, A. A., and N. Agrawal. 2000. Channel dynamics under price and service competition. *Manufacturing & Service Operations Management* 2 (4): 372–391.

Urban, T. L. 2005. Inventory models with inventory-level-dependent demand: A comprehensive review and unifying theory. *European Journal of Operational Research* 162 (3): 792–804.

van Ryzin, G., and S. Mahajan. 1999. On the relationship between inventory costs and variety benefits in retail assortments. *Management Science* 45 (11): 1496–1509.

Wang, R., and Z. Wang. 2016. Consumer choice models with endogenous network effects. *Management Science* 63 (11): 3944–3960.

Zipkin, P. 2000. *Foundations of inventory management*. Irwin/McGraw-Hill, Boston MA.

Proofs

Proof of Lemma 1

Proof. To simplify notation, we omit the product subscript i and write $a_i(s, Q_i)$ as $a(s, Q)$. We let $h(x, Q) = \sum_{q=0}^{Q} x^{Q-q} \frac{Q!}{q!}$ for $x \geq 0$, so that $a(s, Q) = 1 - 1/h\Big(\mu(1+s)/v, Q\Big)$. It can be verified that

$$h(x, Q) = 1 + Qxh(x, Q-1). \tag{22}$$

The first part of the lemma shows the dependency of $a(s, Q)$ on Q for a fixed s, or equivalently the dependency of $h(x, Q)$ on x for a fixed Q. For ease of exposition, we let $h_Q(x), h'_Q(x)$, and $h''_Q(x)$, respectively, denote $h(x, Q)$, $\frac{dh(x,Q)}{dx}$, and $\frac{d^2h(x,Q)}{dx^2}$. Note that $h'_Q(x) = \sum_{q=0}^{Q-1}(Q-q)x^{Q-q-1}\frac{Q!}{q!} \geq 0$. Therefore, $h_Q(x)$ is an increasing function of x. We have that $h_Q\Big(\mu(1+s)/v\Big)$ is an increasing function of s and thus $a(s, Q)$ is an increasing function of s. Furthermore, for $Q \geq 1$, as $s \to \infty$, $h_Q\Big(\mu(1+s)/v\Big)$ tends to infinity and so $\lim_{s\to\infty} a(s, Q) = 1$.

To complete the proof of the first part of the lemma, we show below that $\frac{1}{h_Q(x)}$ is a convex function of x which implies that $a(s, Q)$ is a concave function of s. The proof is by induction over Q. The statement holds for $Q = 0$ since $h_0(x) = 1$. Assume the statement holds for $Q-1$. That is, assume $\frac{1}{h_{Q-1}(x)}$ is convex in x. Then, we have $\frac{d}{dx}(\frac{1}{h_{Q-1}(x)})$ is an increasing function of x. Therefore, $\frac{h'_{Q-1}(x)}{[h_{Q-1}(x)]^2}$ is decreasing in x and we have

$$\frac{h'_{Q-1}(x)}{[h_{Q-1}(x)]^2} \leq Q - 1 \leq Q \tag{23}$$

where the first inequality uses $h'_{Q-1}(0) = Q-1$ and $h_{Q-1}(0) = 1$. Convexity of $\frac{1}{h_{Q-1}(x)}$ also implies that

$$0 \leq \frac{d^2}{dx^2}\left(\frac{1}{h_{Q-1}(x)}\right) = \frac{-h_{Q-1}(x)h''_{Q-1}(x) + 2[h'_{Q-1}(x)]^2}{[h_{Q-1}(x)]^3},$$

and so

$$-h_{Q-1}(x)h''_{Q-1}(x) + 2[h'_{Q-1}(x)]^2 \geq 0. \tag{24}$$

We now use (23) and (24) to show that $-h_Q(x)h''_Q(x) + 2[h'_Q(x)]^2 \geq 0$. This implies that

$\frac{d^2}{dx^2}\left(\frac{1}{h_Q(x)}\right) \geq 0$, which completes the induction step. Using (22), we have

$$\begin{aligned}
-h_Q(x)h''_Q(x) + 2[h'_Q(x)]^2 &= -[1+Qxh_{Q-1}(x)][2Qh'_{Q-1}(x) + Qxh''_{Q-1}(x)] + 2[Qh_{Q-1}(x) + Qxh'_{Q-1}(x)]^2 \\
&= (Qx)^2\left[-h_{Q-1}(x)h''_{Q-1}(x) + 2[h'_{Q-1}(x)]^2\right] \\
&\quad +Qx[2Qh_{Q-1}(x)h'_{Q-1}(x) - h''_{Q-1}(x)] - 2Qh'_{Q-1}(x) + 2Q^2[h_{Q-1}(x)]^2 \\
&\geq Qx\left[2Qh_{Q-1}(x)h'_{Q-1}(x) - 2\frac{[h'_{Q-1}(x)]^2}{h_{Q-1}(x)}\right] - 2Qh'_{Q-1}(x) + 2Q^2[h_{Q-1}(x)]^2 \\
&= 2Qxh_{Q-1}(x)h'_{Q-1}(x)\left[Q - \frac{h'_{Q-1}(x)}{[h_{Q-1}(x)]^2}\right] + 2Q[h_{Q-1}(x)]^2\left[Q - \frac{h'_{Q-1}(x)}{[h_{Q-1}(x)]^2}\right] \\
&\geq 0
\end{aligned}$$

where the first inequality uses (24) and the last inequality uses (23) and the fact that $h'_{Q-1}(x) \geq 0$. Finally, we note that $\frac{1}{h_1(x)} = \frac{1}{1+x}$ is strictly convex. Therefore, the above arguments can be repeated for $Q \geq 1$ by replacing the weak inequalities by strict inequalities to show that $\frac{1}{h_Q(x)}$ is strictly convex for $Q \geq 1$. This completes the proof of the first statement of the lemma.

Now we prove the second statement which shows the dependency of $a(s,Q)$ on Q for a fixed s. Letting $b_x(Q) = 1/h(x,Q)$, we show below that $b_x(Q)$ is decreasing in Q, $\lim_{Q\to\infty} b_x(Q) = 0$ and that $b_x(Q)$ is a convex function of Q. Since $a(s,Q) = 1 - b_{\mu(1+s)/v}(Q)$, this immediately proves the second statement of the lemma.

Using (22), we have that

$$b_x(Q) = \frac{1}{h(x,Q)} = \frac{1}{1+Qxh(x,Q-1)} = \frac{1}{1+Qx/b_x(Q-1)} = \frac{b_x(Q-1)}{b_x(Q-1)+Qx} \tag{25}$$

for $Q \geq 1$, with $b_x(0) = 1$. We can easily see that $b_x(Q) \leq b_x(Q-1)$, which we prove by induction. Indeed, $b_x(1) \leq b_x(0)$, and when $b_x(Q-1) \leq b_x(Q-2)$, since $\frac{b_x(Q-1)}{b_x(Q-1)+Qx} \leq \frac{b_x(Q-1)}{b_x(Q-1)+(Q-1)x} \leq \frac{b_x(Q-2)}{b_x(Q-2)+(Q-1)x}$, it follows that $b_x(Q) \leq b_x(Q-1)$. Hence $b_x(Q)$ is decreasing in Q. Since $0 \leq b_x(Q-1) \leq 1$, it also follows from (25) that $\lim_{Q\to\infty} b_x(Q) = 0$.

We complete the proof by showing that $b_x(Q+1) - b_x(Q) \geq b_x(Q) - b_x(Q-1)$ which implies that $b_x(Q)$ is convex in Q. From (25) we have that $b_x(Q-1) = \frac{Qxb_x(Q)}{1-b_x(Q)}$. Therefore, the increasing first differences property is equivalent to $b_x(Q)$ satisfying the inequality

$$1 - \frac{1}{b_x(Q)+(Q+1)x} \leq \frac{Qx}{1-b_x(Q)} - 1. \tag{26}$$

Letting $LHS(y,Q,x) = 1 - \frac{1}{y+(Q+1)x}$ and $RHS(y,Q,x) = \frac{Qx}{1-y} - 1$, it can be verified (using Mathematica, for example) that $LHS(y,Q,x) \leq RHS(y,Q,x)$ provided

$$y \leq y^l(Q,x) = \frac{1}{4}\left(-(3Q+2)x - \sqrt{x\left((Q+2)^2x - 2Q + 4\right) + 1} + 3\right)$$

or

$$y \geq y^u(Q,x) = \frac{1}{4}\left(-(3Q+2)x + \sqrt{x\left((Q+2)^2x - 2Q + 4\right) + 1} + 3\right).$$

We next show that $b_x(Q) \geq y^u(Q,x)$, which implies that $b_x(Q)$ satisfies (26) and thus has increasing first differences. We prove this by induction. The statement holds for $Q = 0$ since $b_x(0) = 1 = y^u(0,x)$. Assuming it holds for $Q-1$, we have $b_x(Q) = \frac{b_x(Q-1)}{b_x(Q-1)+Qx} \geq \frac{y^u(Q-1,x)}{y^u(Q-1,x)+Qx}$, where the inequality follows from the induction assumption. It can further be shown (again using Mathematica) that $\frac{y^u(Q-1,x)}{y^u(Q-1,x)+Qx} \geq y^u(Q,x)$, thus completing the induction step. ∎

Proof of Lemma 2

To simplify notation, we omit the product subscript i and write $a_i(s,Q_i)$ as $a(s,Q)$. We also write $h(x,Q)$ as $h_Q(x)$.

To prove the first statement, we first argue that $\frac{1}{h_{Q-1}(x)} + Qx$ is increasing in x. To see this, by Lemma 1 we have that $1/h_{Q-1}(x)$ is convex in x. Therefore $\frac{d}{dx}(\frac{1}{h_{Q-1}(x)}) \geq \frac{d}{dx}(\frac{1}{h_{Q-1}(x)})|_{x=0} = -(Q-1)$, where we use $h_{Q-1}(0) = 1$ and $h'_{Q-1}(0) = Q-1$. We have that

$$\frac{d}{dx}\left(\frac{1}{h_{Q-1}(x)} + Qx\right) \geq -(Q-1) + Q \geq 0$$

and so $\frac{1}{h_{Q-1}(x)} + Qx$ is increasing in x. This together with (22) implies that $h_{Q-1}(x)/h_Q(x) = \frac{1}{1/h_{Q-1}(x)+Qx}$ is decreasing in x. We now prove that for a given Q and $s' < s$, $a(s,Q)/a(s',Q) \leq (1+s)/(1+s')$. We have

$$\frac{a(s,Q)}{a(s',Q)} = \frac{\frac{h_Q(\mu(1+s)/v)-1}{h_Q(\mu(1+s)/v)}}{\frac{h_Q(\mu(1+s')/v)-1}{h_Q(\mu(1+s')/v)}} = \frac{1+s}{1+s'}\frac{\frac{h_{Q-1}(\mu(1+s)/v)}{h_Q(\mu(1+s)/v)}}{\frac{h_{Q-1}(\mu(1+s')/v)}{h_Q(\mu(1+s')/v)}} \leq \frac{1+s}{1+s'},$$

where the last equality follows from (22) and the inequality holds since $h_{Q-1}(x)/h_Q(x)$ is decreasing in x.

To show the second statement, we note that $h'_Q(x) = \sum_{q=0}^{Q-1}(Q-q)x^{Q-q-1}\frac{Q!}{q!} = [h_Q(x)-1]/x$. Since $a(s,Q) = 1 - 1/h_Q(\mu(1+s)/v, Q)$

$$a'(s,Q) = \frac{\mu}{v}\frac{h'_Q(\mu(1+s)/v)}{[h_Q(\mu(1+s)/v)]^2} = \frac{[h_Q(\mu(1+s)/v)-1]}{(1+s)[h_Q(\mu(1+s)/v)]^2} = \frac{a(s,Q)}{(1+s)h_Q(\mu(1+s)/v)}.$$

Therefore $a'(s,Q)/a(s,Q) = \frac{1}{(1+s)h_Q(\mu(1+s)/v)}$. From Lemma 1, we have that $1/h_Q(x)$ is decreasing in Q. It follows that $a'(s,Q)/a(s,Q)$ is decreasing in Q. ∎

Proof of Lemma 3

Let Q^* be an optimal solution to (15) (we omit the dependency on s to simplify the notation). By letting $x_{i,q} = 1_{q \leq Q_i^*}$, it is immediate that $x = \{x_{i,q} | \forall i, q\}$ is a feasible solution to LP

(16). Moreover, the objective function value is $\sum_{i=1}^{n}\sum_{q=1}^{C} v_i\delta_i(s,q)x_{i,q} = \sum_{i=1}^{n} v_i \sum_{q=1}^{Q_i^*} \delta_i(s,q) = \sum_{i=1}^{n} v_i \sum_{q=1}^{Q_i^*}[a_i(s,q) - a_i(s,q-1)] = \sum_{i=1}^{n} v_i a_i(s,Q_i^*) = V(s)$.

Consider now an optimal solution $x^* = \{x_{i,q}^* | \forall i,q\}$ to (16). Because the constraint matrix is totally unimodular, the optimal solution is integer. Furthermore, since $\delta_i(s,q)$ is decreasing in q, we have $x_{i,q}^* \geq x_{i,q+1}^*$. Indeed, if $0 = x_{i,q}^* < x_{i,q+1}^* = 1$, we could simply interchange the values of $x_{i,q}^*$ and $x_{i,q+1}^*$, the resulting solution would still feasible, but its objective function value would increase by $\delta_i(s,q) - \delta_i(s,q+1) \geq 0$, thereby leading to a contradition. Therefore, we have $x_{i,q}^* \geq x_{i,q+1}^*$. Letting $Q_i = \max\{q | x_{i,q}^* = 1\}$, we have that $(Q_1, \ldots, Q_n)$ is a feasible solution to problem (15) with the same objective function value. It follows that problem (15) is equivalent to LP (16).

■

Proof of Lemma 4

Proof. This is an upper bound because any feasible solution of (17) is also feasible in (18). ■

Proof of Lemma 5

Proof. Letting $\theta/(1+s)$ denote the dual multiplier of the constraint $\sum_{i=1}^{n}\sum_{q=1}^{C} v_i\delta_i(s,q)x_{i,q} = s$, we have that for any θ, $(\theta s + \hat{Z}(s,\theta))/(1+s) \geq Z^{UB}(s)$. Since this is an LP, the strong duality theorem applies and hence $\min_{\theta\in\mathbb{R}} \dfrac{\theta s + \hat{Z}(s,\theta)}{1+s} = Z^{UB}(s)$.

Notice that at an optimal solution of (20), we necessarily have that if $r_i < \theta$, then $x_{i,q} = 0$, because otherwise we can reduce $x_{i,q}$ to zero (which maintains feasibility) and increase the objective.

Finally, $\hat{Z}(s,\theta)$ is the maximum of linear functions of θ, hence convex in θ.

Before we prove the last two statements of the lemma, we note that by making the transformation $Q_i = \sum_{q=1}^{C} x_{i,q}$, we can write LP (20) equivalently as

$$\hat{Z}(s,\theta) = \max_{\mathbf{Q}\in\mathcal{Q}} \sum_{i=1}^{n} (r_i - \theta) v_i a_i(s,Q_i) = \max_{\mathbf{Q}\in\mathcal{Q}} F(s,\mathbf{Q}) - \theta V(s,\mathbf{Q}), \tag{27}$$

where $F(s,\mathbf{Q}) = \sum_{i=1}^{n} r_i v_i a_i(s,Q_i)$ and $V(s,\mathbf{Q})$ is as defined in (11). Further note that $\frac{\theta s + \hat{Z}(s,\theta)}{1+s}$ is a convex function of θ and its subgradient at θ is given by $\frac{s - V(s,\mathbf{Q}^*(\theta))}{1+s}$ where $\mathbf{Q}^*(\theta)$ is an optimal solution to (27).

Now if θ^* maximizes $\frac{\theta s + \hat{Z}(s,\theta)}{1+s}$, then 0 is an element of the subgradient at θ. If the subgradient is unique (that is we have an unique optimal solution $\mathbf{Q}^* = \mathbf{Q}^*(\theta^*)$), then it follows that $\frac{s - V(s,\mathbf{Q}^*)}{1+s} = 0$, which implies that $s = V(s,\mathbf{Q}^*)$. Therefore, $\mathbf{Q}^*$ is feasible to (17). We have $Z(s) \geq R(s,\mathbf{Q}^*) = R(s,\mathbf{Q}^*) + \theta^*\left(\frac{s - V(s,\mathbf{Q}^*)}{1+s}\right) = \frac{\theta^* s + F(s,\mathbf{Q}^*) - \theta^* V(s,\mathbf{Q}^*)}{1+s} = Z^{UB}(s)$. Since $Z^{UB}(s)$ is an upper bound on $Z(s)$, we have $Z(s) = Z^{UB}(s)$ and this completes the proof of the first part.

Next, consider the case that the subgradient at θ^* is not unique. Let $\mathbf{Q}^l$ and $\mathbf{Q}^u$ be the optimal solutions which define the extreme rays of the subdifferential. Assume without loss of generality that $V(s,\mathbf{Q}^l) < V(s,\mathbf{Q}^u)$. Since 0 is an element of the subgradient, we have that

$$\lambda\left(\frac{s-V(s,\mathbf{Q}^l)}{1+s}\right)+(1-\lambda)\left(\frac{s-V(s,\mathbf{Q}^u)}{1+s}\right)=0$$

for some $0 \leq \lambda \leq 1$. This implies that

$$s = \lambda V(s,\mathbf{Q}^l)+(1-\lambda)V(s,\mathbf{Q}^u), \tag{28}$$

and so $V(s,\mathbf{Q}^l) < s < V(s,\mathbf{Q}^u)$. Further, since $\mathbf{Q}^l$ and $\mathbf{Q}^u$ are optimal solutions to (27) evaluated at θ^*, we have

$$\hat{Z}(s,\theta^*) = F(s,\mathbf{Q}^l)-\theta^* V(s,\mathbf{Q}^l) = F(s,\mathbf{Q}^u)-\theta^* V(s,\mathbf{Q}^u). \tag{29}$$

Finally, note that

$$\begin{aligned}
Z^{UB}(s) &= \frac{\theta^* s+\hat{Z}(s,\theta^*)}{1+s} \\
&= \frac{\theta^* s+\lambda[F(s,\mathbf{Q}^l)-\theta^* V(s,\mathbf{Q}^l)]+(1-\lambda)[F(s,\mathbf{Q}^u)-\theta^* V(s,\mathbf{Q}^u)]}{1+s} \\
&= \frac{\lambda F(s,\mathbf{Q}^l)+(1-\lambda)F(s,\mathbf{Q}^u)}{1+\lambda V(s,\mathbf{Q}^l)+(1-\lambda)V(s,\mathbf{Q}^u)} \\
&= \frac{\lambda F(s,\mathbf{Q}^l)+(1-\lambda)F(s,\mathbf{Q}^u)}{\lambda[1+V(s,\mathbf{Q}^l)]+(1-\lambda)[1+V(s,\mathbf{Q}^u)]}
\end{aligned}$$

where the second equality uses (29) and the third one uses (28). Therefore, we have that

$$\min\left\{\frac{F(s,\mathbf{Q}^l)}{1+V(s,\mathbf{Q}^l)},\frac{F(s,\mathbf{Q}^u)}{1+V(s,\mathbf{Q}^u)}\right\} \leq Z^{UB}(s) \leq \max\left\{\frac{F(s,\mathbf{Q}^l)}{1+V(s,\mathbf{Q}^l)},\frac{F(s,\mathbf{Q}^u)}{1+V(s,\mathbf{Q}^u)}\right\},$$

completing the proof. ■

Proof of Theorem 1

We first give a heuristic argument that illustrates the key ideas. The argument is heuristic since it assumes differentiability of $\hat{Z}(s,\theta)$. Note that $\hat{Z}(s,\theta)$ is a piecewise-linear convex function of θ, but it is not necessarily differentiable. Still, the heuristic argument provides insight and we later make it rigorous by working with a perturbed version of $\hat{Z}(s,\theta)$ that is differentiable.

For $s > 0$, letting $\theta^*(s)$ denote the optimal solution to $\min_{\theta\in\mathbb{R}} \frac{\theta s+\hat{Z}(s,\theta)}{1+s}$, we can use the envelope theorem to obtain

$$\frac{dZ^{UB}(s)}{ds} = \frac{\theta^*(s)+\frac{\partial \hat{Z}(s,\theta^*(s))}{\partial s}-Z^{UB}(s)}{1+s}. \tag{30}$$

Consider a critical point of $Z^{UB}(s)$, i.e., s such that $dZ^{UB}(s)/ds = 0$. At that point,

$$\frac{d^2Z^{UB}(s)}{ds^2} = \frac{\frac{\partial\theta^*(s)}{\partial s}\left(1 + \frac{\partial^2\hat{Z}(s,\theta^*(s))}{\partial s\partial\theta}\right) + \frac{\partial^2\hat{Z}(s,\theta^*(s))}{\partial s^2}}{1+s}. \tag{31}$$

Noting that at the optimal solution $\theta^*(s)$, $s + \partial\hat{Z}(s,\theta^*(s))/\partial\theta = 0$ we have that

$$\frac{\partial^2\hat{Z}(s,\theta^*(s))}{\partial\theta\partial s} = -1.$$

Therefore, if $\frac{\partial^2\hat{Z}(s,\theta^*(s))}{\partial s^2} < 0$, then we have that $\frac{\partial^2 Z^{UB}(s)}{\partial s^2} < 0$, which implies that $Z^{UB}(s)$ is quasi-concave (Boyd and Vandenberghe 2004). Intuitively, if the optimal solution to LP (20) has a product i with $Q_i^* > 0$ and $r_i - \theta^* > 0$ (from Lemma 5), then from optimality we should have $\frac{\partial^2 a_i}{\partial s^2}(s, Q_i^*) < 0$ (from Lemma 1) and hence $\frac{\partial^2\hat{Z}(s,\theta^*(s))}{\partial s^2} < 0$.

Now we make the above arguments rigorous: we apply a small random perturbation to $\hat{Z}(s,\theta)$ so that it becomes a smooth function. We also show that the perturbed function satisfies the conditions stated in the last line of the previous paragraph. We begin by noting that from (27), we can write $\hat{Z}(s,\theta)$ equivalently as $\hat{Z}(s,\theta) = \max_{\mathbf{Q}\in\mathcal{Q}}\sum_{i=1}^n (r_i - \theta)v_i a_i(s,Q_i)$. We work with the following perturbed version of $\hat{Z}(s,\theta)$:

$$\hat{Z}^P(s,\theta) = E\left[\max_{\mathbf{Q}\in\mathcal{Q}}\sum_{i=1}^n (r_i - \theta + \omega_i)v_i a_i(s,Q_i)\right], \tag{32}$$

where ω_i are independent, uniformly distributed random variables on the interval $[0,\epsilon]$ for some $\epsilon > 0$. Note that for small ϵ, $\hat{Z}^P(s,\theta)$ is going to be a good approximation to $\hat{Z}(s,\theta)$. Moreover, by Lemma 6.3.1 in Glasserman (1994) $\hat{Z}^P(s,\theta)$ is differentiable in its arguments and we can interchange the order of the expectation and the derivative. We use these observations to show below that the perturbed version of $Z^{UB}(s)$, $Z^{UB-P}(s) = \min_{\theta\in\mathbb{R}}\frac{\theta s + \hat{Z}^P(s,\theta)}{1+s}$ is quasi-concave in s.

Let $r_{max} = \max_i\{r_i\}$ and let $\theta^*(s)$ denote the minimizer of $\frac{\theta s + \hat{Z}^P(s,\theta)}{1+s}$. Also, for $\hat{\mathbf{Q}} \in \mathcal{Q}$, we let $\Omega(\hat{\mathbf{Q}}) = \{\omega : \sum_{i=1}^n (r_i - \theta + \omega_i)v_i a_i(s,\hat{Q}_i) \geq \sum_{i=1}^n (r_i - \theta + \omega_i)v_i a_i(s,Q_i)\,\forall \mathbf{Q} \in \mathcal{Q},\, \omega_i \in [0,\epsilon]\,\forall i\}$. That is, $\Omega(\hat{\mathbf{Q}})$ is the set of values for ω such that the inventory vector $\hat{\mathbf{Q}}$ attains that maximum in the expression on the right-hand side of Equation (32). We begin with some preliminary results. Lemma 6 below establishes an upper bound on $\theta^*(s)$.

Lemma 6. *We have that $\theta^*(s) < r_{max} + \epsilon$ for $s > 0$.*

Proof. We have

$$\frac{\partial\hat{Z}^P(s,\theta)}{\partial\theta} = E\left[\sum_{\mathbf{Q}\in\mathcal{Q}} 1_{\omega\in\Omega(\mathbf{Q})}\left(-\sum_{i=1}^n v_i a_i(s,Q_i)\right)\right] = \sum_{\mathbf{Q}\in\mathcal{Q}}\left(-\sum_{i=1}^n v_i a_i(s,Q_i)\right)P(\Omega(\mathbf{Q})),$$

where the first equality follows from interchanging the order of the derivative and the expectation. Now, if $\theta \geq r_{max} + \epsilon$, since $\omega_i \leq \epsilon$ and $a_i(s, Q_i) \geq 0$, it follows that $\sum_{i=1}^{n}(r_i - \theta + \omega_i)v_i a_i(s, Q_i) \leq 0$ for all $\mathbf{Q} \in \mathcal{Q}$. Since $a_i(s, 0) = 0$, it follows that $P(\Omega(\mathbf{0})) = 1$ and we have $\frac{\partial \hat{Z}^P(s,\theta)}{\partial \theta} = 0$ for $\theta \geq r_{max} + \epsilon$. Therefore

$$\frac{\partial(\theta s + \hat{Z}^P(s, \theta))}{\partial \theta} = s + \frac{\partial \hat{Z}^P(s, \theta)}{\partial \theta} = s > 0 \tag{33}$$

for $\theta \geq r_{max} + \epsilon$. Since $\frac{\theta s + \hat{Z}^P(s,\theta)}{1+s}$ is a convex function of θ, it follows that $\theta^*(s) < r_{max} + \epsilon$. ∎

The following lemma roughly states that if an inventory vector maximizes the right-hand side of Equation (32) then every product that is stocked ($Q_j > 0$) has a positive "net" contribution ($r_j - \theta + \omega_j$). It can be viewed as the analog of the optimality condition stated in Lemma 5.

Lemma 7. *For $\mathbf{Q} \in \mathcal{Q}$, we have $r_j - \theta + \omega_j \geq 0$ for all $\omega \in \Omega(\mathbf{Q})$, for all products j with $Q_j > 0$.*

Proof. Let $Q_j > 0$ and consider the inventory vector $\hat{\mathbf{Q}}$ where $\hat{Q}_i = Q_i$ for $i \neq j$ and $\hat{Q}_j = 0$. Note that we have $\hat{\mathbf{Q}} \in \mathcal{Q}$. For all $\omega \in \Omega(\mathbf{Q})$, we have that $\sum_{i=1}^{n}(r_i - \theta + \omega_i)v_i a_i(s, Q_i) \geq \sum_{i=1}^{n}(r_i - \theta + \omega_i)v_i a_i(s, \hat{Q}_i)$. Using $Q_i = \hat{Q}_i$ for $i \neq j$ and $a_j(s, \hat{Q}_j) = 0$, this reduces to $(r_j - \theta + \omega_j)v_j a_j(s, Q_j) \geq 0$. Since $Q_j > 0$, $a_j(s, Q_j) > 0$ and therefore $r_j - \theta + \omega_j \geq 0$. ∎

The next lemma shows that if $\theta < r_{max} + \epsilon$, then there exists a non-zero inventory vector $\hat{\mathbf{Q}}$ with $P(\Omega(\hat{\mathbf{Q}})) > 0$.

Lemma 8. *If $\theta < r_{max} + \epsilon$, then there exists $\hat{\mathbf{Q}} \subset \mathcal{Q}$, $\hat{\mathbf{Q}} \neq \mathbf{0}$ with $P(\Omega(\hat{\mathbf{Q}})) > 0$.*

Proof. We first note that $P(\Omega(\mathbf{0})) < 1$. To see this, consider an inventory vector $\bar{\mathbf{Q}}$ with $\bar{Q}_i = 0$ if $r_i < r_{max}$ and $\bar{Q}_i = 1$ if $r_i = r_{max}$. We have $\sum_{i=1}^{n}(r_i - \theta + \omega_i)v_i a_i(s, \bar{Q}_i) = \sum_{i:r_i = r_{max}}(r_{max} - \theta + \omega_i)v_i a_i(s, \bar{Q}_i) > 0 = \sum_{i=1}^{n}(r_i - \theta + \omega_i)v_i a_i(s, 0)$ provided $\omega_i > \theta - r_{max}$ for all i with $r_i = r_{max}$. Since, $\theta - r_{max} < \epsilon$, this happens with a positive probability and for such ω, the inventory vector $\mathbf{0}$ does not attain the maximum on the right-hand side of Equation (32). Therefore $P(\Omega(\mathbf{0})) < 1$. Since $P([0, \epsilon]^n) = 1$ and $\mathcal{Q}$ has finitely many elements, it follows that there exists a non-zero inventory vector $\hat{\mathbf{Q}}$ with $P(\Omega(\hat{\mathbf{Q}})) > 0$. ∎

Finally, we have the following lemma.

Lemma 9. *For $\theta < r_{max} + \epsilon$, $\frac{\partial^2 \hat{Z}^P(s,\theta)}{\partial s^2} < 0$.*

Proof. We have

$$\frac{\partial^2 \hat{Z}^P(s, \theta)}{\partial s^2} = \sum_{\mathbf{Q} \in \mathcal{Q}} \left[\sum_{i=1}^{n} v_i a_i''(s, Q_i) E\left\{(r_i - \theta + \omega_i) 1_{\omega \in \Omega(\mathbf{Q})}\right\} \right] \tag{34}$$

where ${a_i}''(s,Q_i) = \frac{\partial^2 a_i(s,Q_i)}{\partial s^2}$. By the first part of Lemma 1, we have that ${a_i}''(s,Q_i) \leq 0$. On the other hand, by Lemma 7, $E\left\{(r_i - \theta + \omega_i)1_{\omega\in\Omega(\mathbf{Q})}\right\} \geq 0$. Therefore, each term on the right-hand side of Equation (34) is less than or equal to zero. We now argue that there is one term that is strictly negative completing the proof.

Let $\hat{\mathbf{Q}}$ be as in Lemma 8 with $\hat{Q}_i > 0$. By the first part of Lemma 1, we have that ${a_i}''(s,Q_i) < 0$. By Lemma 7, we have that $r_i - \theta + \omega_i \geq 0$ for all $\omega \in \Omega(\hat{\mathbf{Q}})$. Since $P(\Omega(\hat{\mathbf{Q}})) > 0$, there exists ϵ_i^l $(\epsilon > \epsilon_i^l > \theta - r_i)$ such that the set $H(\hat{\mathbf{Q}}) = \{\omega \in \Omega(\hat{\mathbf{Q}}) : \omega_i > \epsilon_i^l\}$ has $P(H(\hat{\mathbf{Q}})) > 0$. We have

$$E\left\{(r_i - \theta + \omega_i)1_{\omega\in\Omega(\hat{\mathbf{Q}})}\right\} \geq E\left\{(r_i - \theta + \omega_i)1_{\omega\in H(\hat{\mathbf{Q}})}\right\} > 0$$

where the last inequality holds since $r_i - \theta + \omega_i > 0$ for all $\omega \in H(\hat{\mathbf{Q}})$ and $P(H(\hat{\mathbf{Q}})) > 0$. It follows that $v_i {a_i}''(s,\hat{Q}_i) E\left\{(r_i - \theta + \omega_i)1_{\omega\in\Omega(\hat{\mathbf{Q}})}\right\} < 0$. ■

We can now apply the heuristic arguments given earlier to $\hat{Z}^P(s,\theta)$. In particular, $\hat{Z}^P(s,\theta)$ is differentiable and the partial derivatives $\frac{\partial^2 \hat{Z}^P(s,\theta)}{\partial s^2}$ and $\frac{\partial^2 \hat{Z}^P(s,\theta)}{\partial \theta \partial s}$ are well defined. Finally lemmas 6 and 9 together imply that $\frac{\partial^2 \hat{Z}^P(s,\theta^*(s))}{\partial s^2} < 0$. This implies that the perturbed function $Z^{UB-P}(s)$ is quasi-concave.

It can be verified that $Z^{UB}(s) \leq Z^{UB-P}(s) \leq Z^{UB}(s) + \epsilon \sum_{i=1}^n v_i$ for all $s \geq 0$. Since $Z^{UB-P}(s)$ is quasi-concave, this implies that $Z^{UB}(\phi s_1 + (1-\phi)s_2) \geq \min\{Z^{UB}(s_1), Z^{UB}(s_2)\} - \epsilon \sum_{i=1}^n v_i$ for $0 \leq \phi \leq 1$ and $s_1, s_2 \geq 0$. Taking the limit as ϵ tends to 0, we have that $Z^{UB}(s)$ is quasi-concave.
■

Proof of Theorem 2

Proof. We begin with some observations. First note that $s^l < s < s^u$. To see this, from Lemma 5 we have $V(s,\mathbf{Q}^l) < s < V(s,\mathbf{Q}^u)$. Corollary 1 then implies that $s < s(\mathbf{Q}^u) = s^u$ and $s > s(\mathbf{Q}^l) = s^l$.

Second, we note that $Z^{UB}(s)$ can be written as a convex combination of $R(s,\mathbf{Q}^l)$ and $R(s,\mathbf{Q}^u)$. To see this, we have from Lemma 5 that

$$Z^{UB}(s) = \frac{\theta^* s + \hat{Z}(s,\theta^*)}{1+s} = \frac{\theta^* s + F(s,\mathbf{Q}^l) - \theta^* V(s,\mathbf{Q}^l)}{1+s} \geq \frac{F(s,\mathbf{Q}^l)}{1+s} = R(s,\mathbf{Q}^l),$$

where $F(s,\mathbf{Q}) = \sum_{i=1}^n r_i v_i a_i(s,Q_i)$, and the inequality holds since $s > V(s,\mathbf{Q}^l)$ (Lemma 5). In a similar manner, it can be shown that $Z^{UB}(s) \leq R(s,\mathbf{Q}^u)$. Therefore

$$Z^{UB}(s) = \lambda R(s,\mathbf{Q}^l) + (1-\lambda) R(s,\mathbf{Q}^u) \tag{35}$$

for some $\lambda \in (0,1)$.

Third, $F(s,\mathbf{Q})$ is concave in s since $a_i(s,Q_i)$ is concave in s (Lemma 1). Therefore, $F(s,\mathbf{Q}^l) \leq F(s^l,\mathbf{Q}^l) + F'(s^l,\mathbf{Q}^l)[s - s^l]$, where $F'(s,\mathbf{Q})$ denotes the derivative of $F(s,\mathbf{Q})$ with respect to s.

Similarly, $F(s,\mathbf{Q}^u) \leq F(s^u,\mathbf{Q}^u) - F'(s^u,\mathbf{Q}^u)[s^u - s]$. This implies

$$\frac{R(s,\mathbf{Q}^l)}{R(s^l,\mathbf{Q}^l)} = \frac{1+s^l}{1+s}\frac{F(s,\mathbf{Q}^l)}{F(s^l,\mathbf{Q}^l)} \leq \frac{1+s^l}{1+s}\left(1 + \frac{F'(s^l,\mathbf{Q}^l)}{F(s^l,\mathbf{Q}^l)}[s - s^l]\right) \tag{36}$$

and

$$\frac{R(s,\mathbf{Q}^u)}{R(s^u,\mathbf{Q}^u)} = \frac{1+s^u}{1+s}\frac{F(s,\mathbf{Q}^u)}{F(s^u,\mathbf{Q}^u)} \leq \frac{1+s^u}{1+s}\left(1 - \frac{F'(s^u,\mathbf{Q}^u)}{F(s^u,\mathbf{Q}^u)}[s^u - s]\right). \tag{37}$$

Fourth, by the second statement of Lemma 2 $a_i'(s,Q)/a_i(s,Q)$ is decreasing in Q, so that $a_i'(s,Q)/a_i(s,Q) \leq a_i'(s,1)/a_i(s,1)$. Since $a_i(s,1) = 1 - \frac{1}{1+(1+s)\mu_j/v_j} = \frac{(1+s)\mu_j/v_j}{1+(1+s)\mu_j/v_j}$, and $a_i'(s,1) = \frac{\mu_j/v_j}{[1+(1+s)\mu_j/v_j]^2}$, we have that $a_i'(s,Q)/a_i(s,Q) \leq \frac{1}{(1+s)[1+(1+s)\mu_j/v_j]} \leq \frac{v_j}{\mu_j(1+s)}$. This implies that

$$\frac{F'(s,\mathbf{Q})}{F(s,\mathbf{Q})} = \frac{\sum_{i=1}^n r_i v_i a_i'(s,Q_i)}{\sum_{i=1}^n r_i v_i a_i(s,Q_i)} \leq \frac{\gamma}{1+s} \tag{38}$$

where $\gamma = \max_j \{v_j/\mu_j\}$.

We are now ready to prove the theorem. We first show that $Z^{UB}(s) \leq \left(1 + \frac{\gamma s^u}{1+s^l}\right)\max\left\{R(s^l,\mathbf{Q}^l), R(s^u,\mathbf{Q}^u)\right\}$. We have

$$\begin{aligned}
\frac{Z^{UB}(s)}{\max\{R(s^l,\mathbf{Q}^l), R(s^u,\mathbf{Q}^u)\}} &\leq \lambda\frac{R(s,\mathbf{Q}^l)}{R(s^l,\mathbf{Q}^l)} + (1-\lambda)\frac{R(s,\mathbf{Q}^u)}{R(s^u,\mathbf{Q}^u)} \\
&\leq \lambda\frac{1+s^l}{1+s}\left\{1 + \frac{F'(s^l,\mathbf{Q}^l)}{F(s_l,\mathbf{Q}^l)}[s - s^l]\right\} + (1-\lambda)\frac{1+s^u}{1+s}\left\{1 - \frac{F'(s^u,\mathbf{Q}^u)}{F(s^u,\mathbf{Q}^u)}[s^u - s]\right\} \\
&\leq \lambda\frac{1+s^l}{1+s}\left\{1 + \frac{F'(s^l,\mathbf{Q}^l)}{F(s_l,\mathbf{Q}^l)}[s - s^l]\right\} + (1-\lambda)\frac{1+s^u}{1+s} \\
&= \frac{1+\lambda s^l + (1-\lambda)s^u}{1+s} + \lambda\frac{1+s^l}{1+s}\frac{F'(s^l,\mathbf{Q}^l)}{F(s^l,\mathbf{Q}^l)}(s - s^l) \\
&\leq \frac{1+\lambda s^l + (1-\lambda)s^u}{1+s} + \lambda\frac{\gamma(s-s^l)}{1+s} \\
&= 1 + \frac{\lambda s^l + (1-\lambda)s^u - s}{1+s} + \frac{\lambda\gamma(s-s^l)}{1+s},
\end{aligned} \tag{39}$$

where the first inequality uses (35), while the second inequality uses (36) and (37). The third inequality uses the facts that $F'(s,\mathbf{Q}) \geq 0$ (since $a_i(s,Q)$ is increasing in s by Lemma 1) and $s^u \geq s$ so that $\frac{F'(s^u,\mathbf{Q}^u)}{F(s^u,\mathbf{Q}^u)}[s^u - s] \geq 0$. The last inequality follows from (38). Now

$$\frac{\lambda s^l + (1-\lambda)s^u - s + \lambda\gamma(s-s^l)}{1+s} \leq \frac{s^u + (\gamma-1)[s-s^l]}{1+s} \leq \frac{s^u + (\gamma-1)s^u}{1+s} \leq \frac{\gamma s^u}{1+s^l}$$

where the first inequality uses $\lambda \leq 1$, the second inequality uses $s < s^u$ and the last inequality uses $s > s^l$. Using this in (39) gives the desired result.

We next show that $Z^{UB}(s) \leq \left(1 + \frac{1+s^u}{1+s^l}\right) \max\left\{R(s^l, \mathbf{Q}^l), R(s^u, \mathbf{Q}^u)\right\}$. We have

$$\begin{aligned}
\frac{Z^{UB}(s)}{\max\left\{R(s^l, \mathbf{Q}^l), R(s^u, \mathbf{Q}^u)\right\}} &\leq \lambda \frac{R(s, \mathbf{Q}^l)}{R(s^l, \mathbf{Q}^l)} + (1-\lambda) \frac{R(s, \mathbf{Q}^u)}{R(s^u, \mathbf{Q}^u)} \\
&\leq \lambda \frac{R(s, \mathbf{Q}^l)}{R(s^l, \mathbf{Q}^l)} + (1-\lambda) \frac{1+s^u}{1+s} \left\{1 - \frac{F'(s^u, \mathbf{Q}^u)}{F(s^u, \mathbf{Q}^u)}[s^u - s]\right\} \\
&\leq \lambda \frac{R(s, \mathbf{Q}^l)}{R(s^l, \mathbf{Q}^l)} + (1-\lambda) \frac{1+s^u}{1+s} \\
&\leq 1 + \frac{1+s^u}{1+s} \\
&\leq 1 + \frac{1+s^u}{1+s^l}
\end{aligned}$$

where the first inequality uses (35), the second inequality uses (37), and the third inequality holds since $\frac{F'(s^u, \mathbf{Q}^u)}{F(s^u, \mathbf{Q}^u)}[s^u - s] \geq 0$. The fourth inequality holds since $R(s, \mathbf{Q})$ is a decreasing function of s (using (14) and the first part of Lemma 2 which states that $a(s, Q)/(1+s)$ is decreasing in s). This together with $s > s^l$ implies that $R(s^l, \mathbf{Q}^l) \geq R(s, \mathbf{Q}^l)$. The last inequality uses $s > s^l$. Putting the two bounds together proves the theorem. ∎

Proof of Theorem 3

Proof. We begin by noting that in the integer program $Z(s)$ we can restrict attention to products i such that $v_i a_i(s, 1) \leq s$. For if $v_i a_i(s, 1) > s$, then any inventory vector $\mathbf{Q}$ with $Q_i > 0$ violates the constraint $V(s, \mathbf{Q}) = s$ since $V(s, \mathbf{Q}) \geq v_i a_i(s, Q_i) \geq v_i a_i(s, 1) > s$. We use this observation to tighten the linear program $Z^{UB}(s)$ to also only include products that satisfy $v_i a_i(s, 1) \leq s$. To simplify the exposition, we also include a dummy product 0 with $r_0 = 0, v_0 = 1$, and $\mu_0 = 0$ so that $a_0(s, Q) = 0$ for all Q. Letting $\mathcal{I}_s = \{i | v_i a_i(s, 1) \leq s\}$, the revised formulation of LP (18) is

$$\begin{aligned}
Z^{UB}(s) = \max_{x \geq 0} \quad & \sum_{i \in \mathcal{I}_s} \sum_{q=1}^{C} \frac{r_i v_i \delta_i(s, q)}{1+s} x_{i,q} \\
\text{s.t.} \quad & x_{i,q} \leq 1 \text{ for all } i \in \mathcal{I}_s, q, \\
& \sum_{i \in \mathcal{I}_s} \sum_{q=1}^{C} x_{i,q} \leq C, \\
& \sum_{i \in \mathcal{I}_s} \sum_{q=1}^{C} v_i \delta_i(s, q) x_{i,q} = s.
\end{aligned} \tag{40}$$

Next, note that an optimal solution to LP (40) has at most two fractional variables. In the rest of the proof, we focus on the case where there are exactly two fractional variables in the optimal solution; the remaining cases can be handled in a similar manner.

We first argue that there is an optimal solution where the variables assuming fractional values correspond to different products. Suppose x is an optimal solution to LP (40) with $x_{i,q}$ and

$x_{i,l}$ being the variables that assume fractional values, with $q < l$. Since there are exactly two variables assuming fractional values, the cardinality constraint must be tight so that $x_{i,q} + x_{i,l} = 1$. We show below that this solution can be transformed into another optimal solution where the variables assuming fractional values correspond to distinct products. Let q_0 be the smallest index such that $x_{0,q_0} = 0$. Consider the solution $\hat{x}$ which is identical to x except that $\hat{x}_{i,q} = x_{i,q} + x_{i,l}\delta_i(s,l)/\delta_i(s,q)$, $\hat{x}_{i,l} = 0$ and $\hat{x}_{0,q_0} = 1 - \hat{x}_{i,q}$. Since $q < l$, $\delta_i(s,l) \leq \delta_i(s,q)$. Therefore $\hat{x}_{i,q} \leq x_{i,q} + x_{i,l}\delta_i(s,q)/\delta_i(s,q) \leq x_{i,q} + x_{i,l} = 1$ and $\hat{x}_{0,q_0} \geq 0$. Further it can be verified that $\hat{x}$ satisfies the remaining constraints and attains the same objective function value as x, and so is optimal. However, $\hat{x}$ is an optimal solution where the variables assuming fractional values correspond to distinct products (i and 0).

So let x be an optimal solution to LP (40) with $x_{j,k}$ and $x_{j',k'}$ being the variables assuming fractional values. Since the cardinality constraint is tight, we have $x_{j,k} + x_{j',k'} = 1$. Let $Q_i = \sum_{q=0}^{C} \hat{x}_{i,q}$ and note that Q_i is integer except for $i = j$ and $i = j'$. We have $Q_j = k - 1 + x_{j,k}$ and $Q_{j'} = k' - 1 + x_{j',k'} = k' - x_{j,k}$. Rounding $x_{j,k}$ down we get an integer solution $\mathbf{Q}^l \in \mathcal{Q}$ with $Q^l_i = Q_i$ for $i \neq j, j'$, $Q^l_j = k - 1$ and $Q^l_{j'} = k'$. Rounding $x_{j,k}$ up we get a integer solution $\mathbf{Q}^u \in \mathcal{Q}$ with $Q^u_i = Q_i$ for $i \neq j, j'$, $Q^u_j = k$ and $Q^u_{j'} = k' - 1$. It can be verified that the objective function of LP (40) can be written as

$$Z^{UB}(s) = \lambda R(s, \mathbf{Q}^l) + (1 - \lambda) R(s, \mathbf{Q}^u) \tag{41}$$

and the last constraint of the LP can be written as

$$s = \lambda V(s, \mathbf{Q}^l) + (1 - \lambda) V(s, \mathbf{Q}^u),$$

where $\lambda = x_{j',k'}$.

Without loss of generality assume that $V(s, \mathbf{Q}^l) \leq V(s, \mathbf{Q}^u)$, which implies that $V(s, \mathbf{Q}^l) \leq s \leq V(s, \mathbf{Q}^u)$. By Corollary 1 we have $s(\mathbf{Q}^l) = s^l \leq s \leq s^u = s(\mathbf{Q}^u)$. On the other hand, the first part of Lemma 2 together with (14) implies that $R(s, \mathbf{Q})$ is decreasing in s. Therefore, we have

$$R(s, \mathbf{Q}^l) \leq R(s^l, \mathbf{Q}^l) \leq Z(s^l) \leq R^{approx}, \tag{42}$$

where the first inequality holds since $s \geq s^l$ and the second inequality holds since $V(s^l, \mathbf{Q}^l) = s^l$ and so $\mathbf{Q}^l$ is a feasible integer solution to $Z(s^l)$ in problem (17).

Next, we consider $R(s, \mathbf{Q}^u)$. We have

$$\begin{aligned} R(s,\mathbf{Q}^u) &= \frac{\sum_{i=1}^n r_i v_i a_i(s, Q_i^u)}{1+s} \\ &= \frac{\sum_{i\neq j,j'} r_i v_i a_i(s,Q_i^l) + r_j v_j a_j(s, Q_j^l+1) + r_{j'} v_{j'} a_{j'}(s, Q_{j'}^l - 1)}{1+s} \\ &\leq \frac{\sum_{i\neq j,j'} r_i v_i a_i(s,Q_i^l) + r_j v_j a_j(s, Q_j^l) + r_{j'} v_{j'} a_{j'}(s, Q_{j'}^l) + r_j v_j [a_j(s, Q_j^l+1) - a_j(s,Q_j^l)]}{1+s} \\ &\leq R(s,\mathbf{Q}^l) + \frac{r_j v_j a_j(s,1)}{1+s} \end{aligned} \tag{43}$$

where the second equality uses that $\mathbf{Q}^u$ and $\mathbf{Q}^l$ are identical in all components except j and j', the first inequality uses the fact that $a_{j'}(s,Q)$ is increasing in Q (Lemma 1) and the last inequality uses the fact that $a_j(s,Q)$ is concave in Q and so the difference $a_j(s,Q+1) - a_j(s,Q)$ is decreasing in Q (Lemma 1).

Let $\hat{\mathbf{Q}}$ denote the inventory vector where we stock only 1 unit of product j. That is, we have $\hat{Q}_i = 0$ for all $i \neq j$ and $\hat{Q}_j = 1$. Now note that since product j is included in LP (40), $j \in \mathcal{I}_s$. Therefore, we have $V(s, \hat{\mathbf{Q}}) = v_j a_j(s,1) \leq s$. By Corollary 1, $\hat{s} = s(\hat{\mathbf{Q}}) \leq s$. Therefore,

$$\frac{r_j v_j a_j(s,1)}{1+s} = R(s,\hat{\mathbf{Q}}) \leq R(\hat{s},\hat{\mathbf{Q}}) \leq Z(\hat{s}) \leq R^{approx} \tag{44}$$

where the first inequality holds since $\hat{s} \leq s$ and $R(s,\hat{\mathbf{Q}})$ is a decreasing function of s. The second inequality holds since $\hat{\mathbf{Q}}$ is a feasible integer solution to $Z(\hat{s})$: $\hat{\mathbf{Q}} \in \mathcal{Q}$ and $\hat{s} = V(\hat{s}, \hat{\mathbf{Q}})$.

Using (44) in (43), we have that $R(s,\mathbf{Q}^u) \leq R(s,\mathbf{Q}^l) + R^{approx}$. Plugging this into (41), we have $Z^{UB}(s) \leq R(s,\mathbf{Q}^l) + (1-\lambda)R^{approx} \leq 2R^{approx}$ where the last inequality uses (42). ■