Retail Inventory Management with Stock-Out Based Dynamic Demand Substitution

Barış Tan\textsuperscript{a}, Selçuk Karabatı\textsuperscript{a}

\textsuperscript{a}College of Administrative Sciences and Economics, Koç University, Rumeli Feneri Yolu, Sariyer, 34450 Istanbul, Turkey

Abstract

We consider the inventory management problem of a product category in a retail setting with Poisson arrival processes, stock-out based dynamic demand substitution, and lost sales. The retailer uses a fixed-review period, order-up-to level system to control the inventory levels. We assume that when a customer cannot find his/her first-choice product on the shelf, he/she attempts to purchase his/her second-choice product with a known probability, which is referred to as the substitution probability. We also assume that the customer’s demand is lost when his/her second-choice product is not available either.

We first present exact computational methods to compute the expected sales, average inventory levels, and number of substitutions between all products for given demand rates, substitution probabilities, and order-up-to levels. We then develop a more tractable approach to approximately compute the same performance measures. The approximate approaches are then used in a profit maximization setting—with profit margins, inventory holding and substitution costs, and service level constraints—to find the optimal order-up-to levels.

In a computational study, we analyze the performances of the approximation approaches, and discuss the impact of profit margins, inventory holding and substitution costs, and service level constraints on the order-up-to levels and the expected profits.

Key words: Inventory Control, Substitution

1. Introduction

This paper studies an inventory management problem in a retail setting with stock-out based substitutions and multiple items in a product category.

The literature on inventory management under stock-out based substitutions studies the supplier-(or manufacturer-)controlled and customer-driven substitution schemes. In the supplier-controlled substitution scheme, in a stock-out instance, the supplier decides whether to fulfill the demand of the customer with another product. The inventory management
(and/or production planning) problem is usually studied in a “one-way substitution” setting, where a higher-graded product can be substituted for a lower-graded product. The primary objective is to minimize the sum of production, inventory holding, and, in some cases, product conversions costs. A detailed discussion of the relevant literature on supplier-controlled substitution is presented in Hsu et al. [5] and Rao et al. [16].

In this paper, inventory management under the customer-driven substitution scheme is studied. In the customer-driven substitution scheme, when the first-choice product of the customer is not available on the shelf, the customer may purchase, with a certain probability, another product in the same category in lieu of her first-choice product. Although the retailer can only indirectly affect customers’ decisions through his inventory management decisions, ignoring product substitutions in managing the inventories may result in sub-optimal performance: Mahajan and van Ryzin [10] analyze a single-period, stochastic inventory problem with substitutable products, and show that “substitution effects can have a significant impact on an assortment’s gross profits.” Ernst and Kamrad [3] study a two-product problem with customer-driven substitution in a newsvendor setting, and conclude that “using a Newsboy Model framework without regard to substitutions can be sub-optimal.”

In the single-period models, it is usually assumed that the demand realizes at the end of the period. A single-product problem can be analyzed under this assumption; however in a multi-product/customer-driven substitution setting, the dynamics of the problem is quite different because of the customer arrival process. Smith and Agrawal [17] and Mahajan and van Ryzin [10] present models, with underage and overage costs, that account for customers’ arrival order in finding the optimal order quantities. Hopp and Xu [4] approximate the dynamic substitution behavior with a fluid network model, and study inventory, price, and assortment decisions in centralized and decentralized settings.

The inventory management problem with static substitutions has been extensively studied in the literature. McGillivray and Silver [11] study a periodic review system with substitutable items having the same unit variable cost and shortage penalty, and develop an upper bound on the inventory and shortage costs savings that could be achieved when the product substitution is taken into account in choosing the order-up-to-levels. Parlar [14] generalizes
the newsvendor problem with a product that perishes in two periods, and assumes that the one-period-old and fresh products are substitutable. Parlar [14] presents an infinite horizon Markov decision model to find the optimal ordering policy. Avşar and Baykal-Gürsoy [2] analyze the competition of two retailers that offer substitutable products, and present a two-person stochastic game to characterize the Nash equilibrium. Rajaram and Tang [15] study a multi-product newsvendor problem with substitutability, and analyze the impact of demand uncertainty on order quantities and expected profits. Netessine and Rudi [13] study a single-period problem where unsatisfied demand for a product flows to other products in deterministic proportions, and present analytically tractable solutions for comparing the profits of the centralized and competitive inventory management settings. Nagarajan and Rajagopalan [12] study a two-product problem with negatively correlated demands. The substitution proportions from the first to the second and from the second to the first product are assumed to be identical. Nagarajan and Rajagopalan [12] first show that, in a single-period setting and when the substitution proportion is not very large, the optimal base-stock levels are not state-dependent. In a computational study, they also show that a heuristic based on the solution of the two-product problem performs well with multiple products and under general conditions.

A closely related research stream studies customer-driven substitution in the context of assortment planning. Kök and Fisher [8] study an assortment planning model with substitutable products, develop a procedure for estimating substitution parameters, and present a heuristic for solving the assortment planning problem. Yücel et al. [18] study assortment and inventory planning problems under customer-driven substitution in retail operations. They show that ignoring substitutability of products or shelf space limitations may result in sub-optimal assortments. Detailed reviews of the literature on assortment planning have been presented by Kök et al. [7] and Mahajan and van Ryzin [9].

Following up on our earlier work on the estimation of substitution probabilities (Karabati, Tan, and Öztürk [6]), the objectives of this paper are two-fold: 1) to measure the performance of a periodic review multi-product inventory system under dynamic and customer-driven demand substitution, and, building on the performance measurement approach, 2)
to determine the optimal order-up-to levels that maximize the expected profit of the system. The inventory management method we propose incorporates the effects of stock-out based dynamic substitutions, and attempts to answer the question whether it is possible to increase the total profit of a product category by setting the order-up-to levels in a way that captures the effects of substitution and profitability of the products, for example, in a way that forces customers of products with low profit margins to substitute with higher profit margin products.

This paper is organized as follows. Section 2 provides a description of the problem. In Section 3, an exact analysis of inventory system’s performance, for the 2- and multiple-product cases, is presented. Section 4 present deterministic and probabilistic approaches to approximately compute the performance measures of interest. Section 5 provides a computational analysis of the approximation approaches. The problem of finding the optimal order-up-to levels is addressed in Section 6. Section 7 concludes the paper.

2. Problem Description

We consider a retailer that stocks and sells $N$ products in a category. Demand for Product $i$ is a Poisson random variable with rate $\lambda_i$, $i = 1, ..., N$. If a customer, whose first-choice product is Product $i$, cannot find it on the shelf, she may substitute it with Product $j$ with probability $\alpha_{ij}$. The substitutions probabilities, which can be estimated with methods discussed in Anupindi et al. [1] and Karabati et al. [6], are an input of our problem. We assume that the customers make only one substitution attempt, and the demand is lost if their second-choice product is not available either. The single substitution attempt restriction is plausible when the impact of the secondary level substitutions is negligible. For example, when service levels are reasonably high, most customers find their first- or second-choice product on the shelf, eliminating the possibility of a second substitution attempt. Kök et al. [7] state that it is also possible to approximate a multiple-substitution attempt model with a single-attempt model by adjusting the parameters.

The retailer uses a fixed review period, order-up-to level system to control the inventory. The review period is equal to $T$ time units, and the order-up-to level for Product $i$ is $Q_i$,
$i = 1, \ldots, N$. The demand of Product $i$ during the review period is denoted by $D_i$, and is a Poisson random variable with rate $\lambda_i T$.

The performance measures we are interested in are the expected sales (total, direct, and through substitution) of products, the expected service and inventory levels, and system’s expected profit during a review cycle:

**Expected Sales:** The expected total sales of Product $i$ during a review cycle is denoted by $S_i, i = 1, \ldots, N$. The expected number of units of Product $i$ sold to the customers of Product $j$ during a review cycle, who substituted Product $j$ with Product $i$ due to the unavailability of Product $j$, is denoted by $S_{ji}, j, i = 1, \ldots, N; j \neq i$. Therefore, the expected number of substitution sales of Product $i$ during a review cycle is equal to $\sum_{j \neq i} S_{ji}$. The expected number of units of Product $i$ sold during a review cycle to the customers of Product $i$, i.e., direct sales of Product $i$, denoted by $S_{ii}, i = 1, \ldots, N$, is then equal to $S_i - \sum_{j \neq i} S_{ji}$.

**Service Levels:** In a multi-item retail setting with dynamic demand substitution, service levels can be measured in different ways. We define $SL_i, i = 1, \ldots, N$, as the ratio of total direct sales of Product $i$ to the total demand of Product $i$ during a review cycle:

$$SL_i = \frac{S_{ii}}{\lambda_i T}.$$  

Equation (1)

**Inventory Level:** The average inventory level for product $i$ is denoted by $T_i, i = 1, \ldots, N$.

3. Performance Evaluation: Exact Analysis

In this section, we consider the performance evaluation of the inventory system. Namely, given the demand rates, substitution structure, review period, and the order-up-to levels, we present analytical methods, for the 2- and multiple-product cases, to determine the expected sales of each product, expected number of substitutions between products, expected inventory levels, service levels achieved for each product, and service level achieved by the system.

5
3.1. The 2-Product Case

The exact analysis we present in this section is based on determining the expected duration of substitution between the two products. Namely, if $\Gamma_{ij}$ is the length of period where Product $j$ is substituted for Product $i$ during one review period, then the expected substitution number from Product $i$ to Product $j$ is $\lambda_i \alpha_{ij} E[\Gamma_{ij}]$.

Let $T_i, i = 1, 2$, be the time inventory of Product $i$ is depleted. Let us first assume that $T_1 < T_2$. If Product 1 is depleted before Product 2, $T_1$ is the sum of $Q_1$ exponentially distributed random variables with rate $\lambda_1$. Then, $T_1$ has an Erlang distribution with $Q_1$ stages and rate $\lambda_1$ for each stage:

$$P[T_1 < t | T_1 < T_2] = 1 - \sum_{j=0}^{Q_1-1} \frac{(\lambda_1 t)^j}{j!} e^{-\lambda_1 t}. \quad (2)$$

$T_2$ can now be expressed in terms of $T_1$ and $\tau_{12}$, where $\tau_{12}$ is the substitution period from Product 1 to Product 2 when $T_1 < T_2 < T$.

$$P[T_2 < t | T_1 < T_2 < T] = P[T_1 + \tau_{12} < t | T_1 < T_2 < T]. \quad (3)$$

During the period $[0, T_1]$, the number of units of Product 2 sold has a Poisson distribution with rate $\lambda_2$. Therefore,

$$P[I_2(T_1) = n_2] = \frac{(\lambda_2 T_1)^{Q_2-n_2}}{(Q_2-n_2)!} e^{-\lambda_2 T_1}, \quad (4)$$

where $I_i(t)$ is the inventory level of Product $i$ at time $t$. When Product 1 is depleted but Product 2 is still available, the demand rate of Product 2 increases to $\lambda_2 + \alpha_{12} \lambda_1$. Then the distribution $\tau_{12}$ is also Erlang with rate $\lambda_2 + \alpha_{12} \lambda_1$ and $I_2(T_1)$ stages. Therefore,

$$P[\tau_{12} < t \mid T_1 < T_2 < T] = \frac{T}{\int_0^{Q_2} \sum_{n_2=0}^{Q_2-1} \left(1 - \sum_{j=0}^{n_2-1} \frac{(\lambda_2 + \alpha_{12} \lambda_1) t^j}{j!} e^{-(\lambda_2 + \alpha_{12} \lambda_1) t}\right) \frac{(\lambda_2 + \alpha_{12} \lambda_1)^{Q_1} (Q_2-n_2)!}{(Q_2-n_2)!} e^{-(\lambda_1 + \lambda_2) t_1} dt_1}{\lambda_1 + \lambda_2} \quad (5)$$

The substitution duration $\Gamma_{12}$ can be expressed as

$$\Gamma_{12} = \begin{cases} 
T - T_1 & T_1 < T < T_2 \\
\tau_{12} & T_1 < T_2 < T \\
0 & T < \min\{T_1, T_2\}
\end{cases} \quad (6)$$
Since the distributions of $T_1$, $T_2$, and $\tau_{12}$ are given in Equations (2), (3), and (5), $E[\Gamma_{12}]$ can be calculated from Equation (6). The case $T_2 < T_1$ is similar and yields $E[\Gamma_{21}]$. The performance measures of interest can now be directly computed: $S_{12} = \lambda_1 \alpha_{12} E[\Gamma_{12}]$, $S_{21} = \lambda_2 \alpha_{21} E[\Gamma_{21}]$, $S_{11} = \lambda_1 E[\min\{T_1, T\}]$ and $S_{22} = \lambda_2 E[\min\{T_2, T\}]$.

Although this approach yields the expected direct and substituted sales numbers in closed form, extending this method to more than two products is not practical due to the large number of cases that need to be considered.

3.2. The Multiple-Product Case

In this section, we present a Continuous Time Markov Chain approach to analyze the inventory management system under dynamic demand substitution.

3.2.1. Determining the State Transition Matrix

We model the system as a discrete time-discrete state space stochastic process. The state of the system in period $t$ is an $N$-tuple $I(t) = (I_1(t), ..., I_N(t))$ where $I_i(t)$ is the inventory level of Product $i$ in period $t$.

The length of each period is $\Delta$. Therefore, there are $T/\Delta$ periods in each review cycle.

We set $\Delta$ very small to ensure that the probability of having two arrivals in the same time period is very small. Therefore, we assume that only one arrival can occur in each period.

Since the arrivals are Poisson, the probability that a demand for Product $j$ arrives in period $t$ is $\lambda_j \Delta$. Similarly, the probability that a customer substitutes Product $i$ for Product $j$ due to unavailability of Product $j$ in period $t$ is $\lambda_j \Delta \alpha_{ji}$. Let the indicator variable $\delta_{I_i(t)}$ is defined to be equal to 1 when $I_i(t) > 0$, and 0 otherwise. Then the indicator variable $\delta_{I_i(t)}(1 - \delta_{I_j(t)})$ is 1 when a customer can substitute Product $j$ for Product $i$, and 0 otherwise.

Therefore, the state transitions are defined by the following equations:

$$
P[I_i(t+1) = n - 1|I_i(t) = n] = \lambda_i \Delta + \sum_{j \neq i} \lambda_j \Delta \alpha_{ji} \delta_{I_i(t)}(1 - \delta_{I_j(t)}) \quad n \geq 1, i = 1, ..., N
$$

$$
P[I_i(t+1) = 0|I_i(t) = 0] = 1 \quad i = 1, ..., N.
$$

(7)

In order to analyze the system, we first generate the state transition matrix of the system automatically. Let $P$ be the state transition probability matrix. Since the state of the
system is an $N$-tuple $\mathbf{I}(t) = (I_1(t), ..., I_N(t))$ and $0 \leq I_i(t) \leq Q_i$, $i = 1, ..., N$, there will be $|\mathbf{I}| = \prod_{i=1}^{N} (Q_i + 1)$ states in the state space. Accordingly, $\mathbf{P}$ is a $|\mathbf{I}| \times |\mathbf{I}|$ sparse matrix. Note that since there are at most $N$ transitions from each state, the number of non-zero elements will be less that $N|\mathbf{I}|$.

We start the state-space generation process at the state where all the inventory levels are at their order-up-to levels. Then we consider $N$ possible changes that correspond to the decrease of one unit in the inventory level of one of the products from this state. If this new state is not included in the state space then it is added to the state space. We then store the index of the current state, the index of the next state, and the transition rate calculated from Equation (7). We repeat this process until we reach the state where all the inventory levels are zero, and the process terminates at this state. Once the non-diagonal elements of $\mathbf{P}$ are determined in this way, the diagonal elements are set to make the row sums equal to one.

3.2.2. Performance Evaluation

Let the state probability row vector be $\mathbf{p}(t) = p(t, i_1, ..., i_N)$ given as

$$\mathbf{p}(t) = p(t, i_1, ..., i_N) = P[I_1(t) = i_1, ..., I_N(t) = i_N].$$

(8)

Since the inventory levels at $t = 0$ are equal to the order-up-to levels at the beginning of each cycle, $p(0, Q_1, ..., Q_N) = 1$. The state probability vector satisfies

$$\mathbf{p}(t + 1) = \mathbf{p}(t) \mathbf{P}.$$  

(9)

Note that each periodic review cycle starts at state $(Q_1, ..., Q_N)$ and terminates at state $(I_1(T/\Delta), ..., I_N(T/\Delta))$ with $0 \leq I_i(T/\Delta) \leq Q_i$, $i = 1, ..., N$. The state probability vector at the end of the review cycle can be determined from Equation (9) starting with $p(0, Q_1, ..., Q_N) = 1$.

Once the probability vector at the end of the review cycle is determined, a number of performance measures can be determined directly:

$Expected Sales$: The expected number of Product $i$ sold in each cycle can be determined
from the state probabilities as

\[ S_i = \sum_{n=0}^{Q_i} (Q_i - n)P[I_i(T/\Delta) = n]. \tag{10} \]

Therefore the expected number of unsold items of Product \( i \) is \( Q_i - S_i \).

**Expected Number of Substitutions:** The expected number of substitution sales from customers of Product \( i \) to Product \( j \) is

\[ S_{ij} = \sum_{t=0}^{T/\Delta} P[I_i(t) = 0, I_j(t) > 0] \lambda_i \Delta \alpha_{ij}. \tag{11} \]

**Service Level** The expected direct sales of Product \( i \) to customers of Product \( i \) can be determined from Equation (10) and Equation (11) as \( S_{ii} = S_i - \sum_{j \neq i} S_{ji} \). Therefore the service level is \( SL_i = \frac{S_{ii}}{\lambda_i T} \).

**Inventory level** The expected inventory level can be determined as

\[ I_i = \Delta T \sum_{t=0}^{T/\Delta} \sum_{n=0}^{Q_i} nP[I_i(T/\Delta) = n]. \tag{12} \]

4. Performance Evaluation: Approximate Analysis

With the exact analyses we have presented in Section 3, the performance measures can be determined computationally, and, particularly in the multiple-product case, they cannot be expressed in closed form. In this section, we present approximation methods that will result in expressions that are more tractable for optimization.

4.1. Deterministic Approximation of the Multiple-Product Case

We first present an approximate analysis of the inventory system with stock-out based substitutions by ignoring the stochastic nature of the customer arrival and choice processes. When the stochastic nature of the problem is ignored, the substitutions quantities are poorly estimated, and, therefore, we will specifically focus on the computation of the average inventory levels.

In a multiple-product setting with given order-up-to levels and simultaneous replenishments, the inventory system does not experience any stock-out instances in the period that
starts with the replenishments and ends with the depletion of one of the product inventories or completion of the review period. Let $l_1$ be the index of the product which has the smallest $\frac{Q_i}{\lambda_i}$ ratio: $l_1 = \arg\min_i \frac{Q_i}{\lambda_i}$. If $\frac{Q_{l_1}}{\lambda_{l_1}} > T$, then, under the assumption that the customer arrival rates are constant, the inventory system does not experience any stock-outs, and, therefore, substitutions. Let $T_{l_1}$ be equal to $\min\{T, \frac{Q_{l_1}}{\lambda_{l_1}}\}$. If $\frac{Q_{l_1}}{\lambda_{l_1}} > T$, then, in the period that follows $T_{l_1}$, the effective arrival rates of other products increase due to the substitutions the customers of Product $l_1$ may make. If we assume that the customers make substitutions with fixed proportions that are equal to the substitution probabilities, in the period that follows $T_{l_1}$, the effective arrival rate of Product $i, i \neq l_1$, increases by $\lambda_{l_1}\alpha_{l_1,i}$. Let now $l_2$ be the product index with

$$l_2 = \arg\min_{i: i \neq l_1} \frac{Q_i - \lambda_i T_{l_1}}{\lambda_i + \lambda_{l_1} \alpha_{l_1,i}}.$$ 

We note that $Q_i - \lambda_i T_{l_1}$ in the above relationship corresponds to the inventory level of Product $i, i \neq l_1$, at time $T_{l_1}$. Similarly, let $T_{l_2}$ be equal to $\min\{T, T_{l_1} + \frac{Q_{l_2} - \lambda_{l_2} T_{l_1}}{\lambda_{l_2} + \lambda_{l_1} \alpha_{l_1,l_2}}\}$. If $T_{l_2} < T$, then, in the period that follows $T_{l_2}$, the effective arrival rate of Product $i, i \neq l_1, l_2$ becomes $\lambda_i + \sum_{k=1}^2 \lambda_k \alpha_{k,i}$.

In Figure 1, we present a graphical representation of a system with 3 products, and with $l_1 = 1$ and $l_2 = 3$. In the example presented in Figure 1, Product 2 is not depleted within the review period, and completes the period with positive stock. The above outlined approximation scheme can be generalized by defining $l_n$ and $T_{l_n}, n = 1, ..., N$ as follows:

$$l_n = \arg\min_{i: i \neq l_1, ..., l_{n-1}} \frac{Q_i - \sum_{j=1}^{n-1} (T_{l_j} - T_{l_{j-1}})(\lambda_i + \sum_{k=1}^{j-1} \lambda_k \alpha_{k,i})}{\lambda_i + \sum_{j=1}^{n-1} \lambda_j \alpha_{j,i}},$$

and

$$T_{l_n} = \min\{T, \frac{Q_{l_n} - \sum_{j=1}^{n-1} (T_{l_j} - T_{l_{j-1}})(\lambda_{l_n} + \sum_{k=1}^{j-1} \lambda_k \alpha_{k,l_n})}{\lambda_{l_n} + \sum_{j=1}^{n-1} \lambda_j \alpha_{j,l_n}}\}.$$ 

Once the $l_n$ and $T_{l_n}, n = 1, ..., N$ values are computed, the average inventory levels can be computed in a straightforward manner by considering beginning and ending inventories of each product in periods defined by the $T_{l_n}, n = 1, ..., N$ values. As illustrated in Figure 1, the substitutions between the products can be computed comparing the direct and effective demand rates of each product in periods formed by the $T_{l_n}, n = 1, ..., N$ values. However,
as we have noted earlier, we will use the deterministic approximation only to estimate the average inventory levels.

4.2. Stochastic Approximation of the Multiple-Product Case

We will use the exact analysis of the 2-product case, presented in Section 3.1 to develop an approximation method for the multiple-product case.

We assume that the retailer sets the order-up-to levels in such a way that $P[D_i \geq Q_i]$ is small, and at most two products are depleted before the review period. This assumption implies that the holding cost is much lower than the cost of losing sales.

Let $T_i < T_j$ and $T_k > T$, $k \neq i, j$. The distribution of $T_i$ is Erlang with rate $\lambda_i$ and $Q_i$ stages (see Section 3.1). This random variable can be approximated with a normal random variable with mean

$$\mu_i = E[T_i] = \frac{Q_i}{\lambda_i}, \quad (13)$$

and variance

$$\sigma_i^2 = Var[T_i] = \frac{Q_i}{\lambda_i^2}. \quad (14)$$
Note that this approximation follows the central limit theorem and quite accurate for large values of $Q_i$. Similarly, when $T_i < T_j$, assuming constant demand and substitution rates, $T_j$ can be approximated with a normal distribution with mean

$$
\mu_j = E[T_j] = \frac{Q_j - \left(\frac{Q_i}{\lambda_i}\right)\lambda_j}{\lambda_j + \alpha_{ij}\lambda_i} + \frac{Q_i}{\lambda_i},
$$

(15)

and variance

$$
\sigma_j^2 = Var[T_j] = \frac{Q_j - \left(\frac{Q_i}{\lambda_i}\right)\lambda_j}{(\lambda_j + \alpha_{ij}\lambda_i)^2} + \frac{Q_i}{\lambda_i} \frac{1}{\lambda_j},
$$

(16)

We assume that $T_i$ and $T_j$ are independent. We know from Section 3.1 that $T_i$ and $T_j$ are not independent. However, when $T_i < T < T_j$, the effect of this approximation on $E[(T - T_i)^+]$ will not be significant. Similarly, when $T_i < T_j < T$, the effect on $E[(T_j - T_i)^+]$ will not be significant.

Let $\Gamma_{ij}$ be the length of the time Product $i$ is substituted with Product $j$.

$$
\Gamma_{ij} = \begin{cases} 
T - T_i & T_i < T < T_j \\
T_j - T_i & T_i < T_j < T \\
0 & T < \min\{T_i, T_j\}
\end{cases}
$$

(17)

Under the above stated assumptions and following Equation (17),

$$
E[\Gamma_{ij}] \approx E[(T - T_i)^+, T_i < T < T_j] + E[(T_j - T_i)^+, T_i < T_j < T].
$$

Furthermore, since it is assumed that $T_i$ and $T_j$ are independent,

$$
E[\Gamma_{ij}] \approx E[(T - T_i)^+]P[T < T_j] + E[(T_j - T_i)^+]P[T_i < T]P[T_i < T_j < T].
$$

Note that with the normal approximation of $T_i$ and $T_j$, we can determine $E[(T - T_i)^+]$, $E[(T_j - T_i)^+]$ and $P[T_i < T]$ directly. Let $\eta(z)$ be the expected number of units short of a standard normal random variable:

$$
\eta(z) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} (x - z)e^{-\frac{1}{2}x^2}dx = \phi(z) - z\Phi(z)
$$

(18)

where $\phi(z)$ and $\Phi(z)$ are the density function and cumulative distribution function of the standard normal given as $\phi(z) = e^{-\frac{1}{2}z^2}$ and $\Phi(z) = \int_{-\infty}^{z} \phi(z)dz$. We can then write the
expected values as

$$E[(T - T_i)^+] = T - \mu_i + \sigma_i \eta \left( \frac{T - \mu_i}{\sigma_i} \right),$$ \hspace{1cm} (19)

$$E[(T - T_i)^+] = T - \mu_i + \sigma_i \left( \phi \left( \frac{T - \mu_i}{\sigma_i} \right) - \frac{T - \mu_i}{\sigma_i} \Phi \left( \frac{T - \mu_i}{\sigma_i} \right) \right),$$ \hspace{1cm} (20)

and

$$E[(T_i - T_j)^+ | T_i < T_j < T] = \sigma_{ji} \left( \phi \left( \frac{\mu_{ji}}{\sigma_{ji}} \right) - \frac{\mu_{ji}}{\sigma_{ji}} \Phi \left( \frac{\mu_{ji}}{\sigma_{ji}} \right) + \phi \left( \frac{T + \mu_{ji}}{\sigma_{ji}} \right) + \frac{T + \mu_{ji}}{\sigma_{ji}} \Phi \left( \frac{T + \mu_{ji}}{\sigma_{ji}} \right) \right) \right),$$ \hspace{1cm} (21)

where $T_j - T_i$ is approximately normal with mean

$$\mu_{ji} = E[T_j - T_i] = E[T_j] - E[T_i],$$ \hspace{1cm} (22)

and variance

$$\sigma_{ji}^2 = Var[T_j - T_i] = Var[T_j] + Var[T_i],$$ \hspace{1cm} (23)

where the first term is determined in closed form following Equation (21) and the second term is also given in closed form in Equation (20). $E[(T - T_j)^+]$ and $E[(T_j - T_i)^+ | T_i < T_j < T]$ can be expressed in a similar fashion. The expected time Product $i$ is substituted with Product $j$ $\Gamma_{ij}$ can be written as

$$\Gamma_{ij} = E[(T_j - T_i)^+ | T_i < T_j < T] \left( 1 - \Phi \left( \frac{T - \mu_i}{\sigma_i} \right) \right) \left( 1 - \Phi \left( \frac{T - \mu_j}{\sigma_j} \right) \right) \right),$$ \hspace{1cm} (24)

As a result, we can determine $S_{ij}$ $i \neq j$ as

$$S_{ij} = E[\Gamma_{ij}] \lambda_i \alpha_{ij},$$ \hspace{1cm} (25)

and $S_{ii}$ as

$$S_{ii} = E[\min\{Q_i, \lambda_i T\}].$$ \hspace{1cm} (26)

Following the normal approximation, $S_{ii}$ can be evaluated as

$$S_{ii} = \lambda_i T - \sqrt{\lambda_i T} \left[ \phi \left( \frac{Q_i - \lambda_i T}{\sqrt{\lambda_i T}} \right) - \phi \left( \frac{Q_i - \lambda_i T}{\sqrt{\lambda_i T}} \right) \right] \left( 1 - \Phi \left( \frac{Q_i - \lambda_i T}{\sqrt{\lambda_i T}} \right) \right).$$ \hspace{1cm} (27)
Finally, if this approximation yields $S_{ij}$ values such that $\sum_j S_{ij} \geq Q_i$, we normalize the values such that

$$\hat{S}_{ij} = S_{ij} \frac{Q_i}{\sum_j S_{ij}}.$$ 

5. Computational Study: Accuracy of the Approximation Approaches

In this section, we present a computational analysis of the approximation quality of the approaches developed in Section 4. In the computational analysis, we consider a set of randomly created 4-product problems. As noted in Karabati, Tan, and Öztürk [6], when the observed number of substitutions is not statically significant, estimation of substitution probabilities is a very challenging task. To deal with this issue, a certain number of products with similar characteristics can be lumped together for analysis purposes. For example, modeling the first three products with the largest market shares explicitly, and lumping the other products with smaller market shares into a single product yields a model with 4 products. The performance of a model of this size can be evaluated quite accurately by using the approximation method presented in this study and also can be optimized effectively.

The products’ order-up-to levels are set to satisfy a randomly selected fill rate without taking the substitution effect into account. Three different service level ranges, [60%, 99%], [70%, 99%], [80%, 99%], are used in problem generation, and the target service level of each product is randomly selected using a uniform distribution between the lower and upper limits of the ranges. The demand rates of the Product 1 and 2 (3 and 4) are generated using a uniform distribution in the [15, 25] ([5, 15]) range. The customer choice model is assumed to be Market-Share Based (see Smith and Agrawal [17], Netessine and Rudi [13], and Kök and Fisher [8]) where the substitute product is chosen according to the substitution probability matrix $\alpha_{ij} = \theta \frac{\lambda_j}{\sum_{i \in N(i)} \lambda_i}, i, j = 1, 2, \ldots, N, \text{ and } i \neq j$. The review time is taken as 20 times units, and the substitution probability, $\theta$, is taken as 60%. For each service level range, 120 problems are randomly generated and simulated for 10 independent replications with 50 review periods in each replication.

In Table 1, we report the deterministic approximation’s performance for products’ aver-
average inventory levels, and stochastic approximation’s performance for products’ total direct sales, and total sales. The average and maximum approximation errors are reported, over 120 problems in each row of Table 1, relative to the average performance observed over 10 replications of the simulation model in each problem instance.

We note that, in order to capture the probabilistic nature of substitutions, all performance measures, with the exception of inventory levels, are estimated with the probabilistic approximation. The figures presented in Table 1 indicate that the inventory levels, products’ total direct and total sales can be closely estimated with approximation approaches described earlier.

In Figure 2, we report the average performance of the probabilistic approach in estimating the number of substitutions between products. We first note that the performance of the approximation approach is dependent on the rate of the realized substitutions: when the number of substitutions per review period is low, the observations are not statistically significant, and approximations are relatively poor. In light of this observation, we report the approximation performance for 3 different minimum levels of substituted demand. For example, when the minimum level is taken as 1%, and when the estimated number of substitutions from Product $i$ to Product $j$, i.e., $S_{ij}$, is less than $\lambda_i \times T \times 0.01$, substitutions from Product $i$ to Product $j$ is considered to be statistically insignificant, and approximation error with $S_{ij}$ is not included in the reported average approximation errors. In Figure 2, we observe that the approximation quality increases when minimum level of statistical significance and the variability in service levels increase. As we discuss in the next section, where a model to find the optimal order-up-to levels under substitution is presented, we need better substitution approximations in cases where the service level of a product is deliberately set low to channel some of its demand to other products through substitutions. In these instances, because the number of substitutions is high, the approximation quality of the probabilistic approach will be high too.
Table 1: Estimation Errors.

<table>
<thead>
<tr>
<th>Service Level Distribution</th>
<th>Inventory</th>
<th>Total Sales</th>
<th>Direct Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>Maximum</td>
<td>Average</td>
</tr>
<tr>
<td>[60%,99%]</td>
<td>0.587%</td>
<td>2.287%</td>
<td>0.005%</td>
</tr>
<tr>
<td>[70%,99%]</td>
<td>0.510%</td>
<td>1.798%</td>
<td>0.010%</td>
</tr>
<tr>
<td>[80%,99%]</td>
<td>0.422%</td>
<td>1.295%</td>
<td>0.071%</td>
</tr>
</tbody>
</table>

Figure 2: Substitution Estimation Errors.
6. Optimal Order-up-to Levels

In this section, we present an approximate solution procedure for the problem of finding the optimal order-up-to levels under stock-out based substitutions, and a computational analysis of the optimization problem over a large set of problems.

6.1. Model

Let the per unit profit of Product $i, i = 1, 2, \ldots, N$, be $\pi_i = p_i - c_i$, where $p_i$ ($c_i$) is the retail price (cost) of product $i$. Also let the inventory carrying cost rate per unit per review time be $h\%$ of a product’s cost, and $s_i, i = 1, 2, \ldots, N$, be the per unit substitution cost when a customer of Product $i$ substitutes Product $i$ with another product. We note that $s_i$ may be used to capture long term effect of substitutions, such as changes in the repeat visits, especially when customers are forced to substitute by the inventory management policy of the retailer. The expected total profit $\Pi$ obtained per unit time can be expressed as

$$\Pi = \frac{1}{T} \sum_{i=1}^{N} (\pi_i S_i - T_i c_i h - s_i \sum_{j \neq i} S_{ij}).$$

(28)

One possible objective of the inventory control policy can be the maximization of expected total profit while maintaining a minimum service level for the direct customers for each product:

$$\begin{align*}
(OOLPS) \quad & \underset{Q_1, \ldots, Q_N}{\text{Max}} \quad \Pi \\
& \quad S_{ii} \geq T\lambda_i \gamma_i, \quad i = 1, 2, \ldots, N,
\end{align*}$$

(29)

where $0 \leq \gamma_i < 1$ is the minimum service level for Product $i$.

Since the approximations of direct sales, and total sales are non-linear, and approximations of expected inventory levels involve integer variables that determine the depletion order of the products, $(OOLPS)$ is still a difficult optimization problem. To approximately solve the problem, we resort to a genetic algorithm where the expected inventory levels are estimated with deterministic approximation (Section 4.1), and expected direct and substitution sales are estimated with the stochastic approximation (Section 4.2).
6.2. Computational Analysis

In this section, we report our computational experience with the optimization problem outlined in the previous section. We first discuss two problem instances, then analyze the results over a larger set of problems.

6.2.1. Case 1

We consider a 4-product problem with $T = 20$, $h = 1.64\%$, and the following parameters:

<table>
<thead>
<tr>
<th>Product</th>
<th>$p_i$</th>
<th>$\pi_i$</th>
<th>$s_i$</th>
<th>$\lambda_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.00</td>
<td>0.60</td>
<td>0.06</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>6.00</td>
<td>0.60</td>
<td>0.06</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>8.00</td>
<td>1.20</td>
<td>0.12</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>10.00</td>
<td>2.00</td>
<td>0.20</td>
<td>6</td>
</tr>
</tbody>
</table>

The substitution probabilities are as follows:

<table>
<thead>
<tr>
<th>$\alpha_{ij}$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>–</td>
<td>0.25</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>2</td>
<td>0.25</td>
<td>–</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>3</td>
<td>0.1</td>
<td>0.1</td>
<td>–</td>
<td>0.3</td>
</tr>
<tr>
<td>4</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>–</td>
</tr>
</tbody>
</table>

When we do not take the effects substitution into account, for a fill-rate service level of 99\%, the order-up-to levels are set as $Q = (251, 251, 170, 130)$. The total profit of the system, when simulated with substitutions, is computed as $\Pi = 670.98$.

The optimization model yields, under a 40\% minimum service level constraint, a solution with $Q = (236, 243, 171, 136)$, and a total profit of $\Pi = 672.90$. The realized total substitutions from Products 1, 2, 3, and 4 are 3.25, 1.41, 0.52, and 0.11, respectively. Although the cost of substitution is low relative to the differences between profit margins of the products, there is no forced substitution in the optimal solution. This is mainly due to the substitution probabilities of the problem, because when decrease the order-up-to levels of Product
1 and/or 2 to increase substitution from these products to more profitable products, i.e., Products 3 and 4, a substantial portion of their demand is lost.

When we increase $\alpha_{1,3}$ and $\alpha_{2,3}$ to 0.3, the optimal solution becomes $Q = (97, 276, 207, 139)$, with a total profit of $\Pi = 680.00$. Because of its low order-up-to level, the direct sales of Product 1 realizes as 40%, the minimum level required by the direct service level constraint. The realized total substitutions from Products 1, 2, 3, and 4 are now 89.35, 1.55, 0.82, and 0.29, respectively. We note that, in the optimal solution, forced substitution is observed in only one of the less profitable products. The substitution probabilities between Products 1 and 2 ($\alpha_{1,2} = \alpha_{2,1} = 0.25$) are significant, and lowering the order-up-to levels of Products 1 and 2 simultaneously results in lost sales when customers of Products 1 and 2 attempt to substitute their preferred product with another one.

When we increase $\alpha_{1,3}$ and $\alpha_{2,3}$ to 0.5, a major portion of customers of both Product 1 and Product 2 are forced to substitute, because, in the optimal solution, the order-up-to levels are set as $Q = (98, 99, 302, 149)$. The total profit of the system is now $\Pi = 715.60$. Although we lost 25% of demands of Products 1 and 2, we recover lost sales with the increased profits when 28% of customers of Products 1 and 2 substitute with Product 3.

6.2.2. Case 2

We consider a 3-product problem with $T = 20$, $h = 5\%$, and the following parameters:

<table>
<thead>
<tr>
<th>Product</th>
<th>$p_i$</th>
<th>$\pi_i$</th>
<th>$s_i$</th>
<th>$\lambda_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.00</td>
<td>0.40</td>
<td>0.04</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>4.20</td>
<td>0.42</td>
<td>0.04</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>4.30</td>
<td>0.43</td>
<td>0.04</td>
<td>20</td>
</tr>
</tbody>
</table>

The substitution probabilities are as follows:

<table>
<thead>
<tr>
<th>$\alpha_{ij}$</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>–</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>2</td>
<td>0.50</td>
<td>–</td>
<td>0.50</td>
</tr>
<tr>
<td>3</td>
<td>0.00</td>
<td>0.00</td>
<td>–</td>
</tr>
</tbody>
</table>
In this example, although substitution costs are higher than the profit margin differences, we obtain an optimal solution with $Q = (352, 422, 423)$, and $\Pi = 367.77$. Approximately 10% of Product 1’s demand is substituted with Products 2 and 3, in equal proportions. When we do not take the effects substitution into account, for a fill-rate service level of 99%, the order-up-to levels are set as $Q = (410, 410, 410)$, and total profit of the system, when simulated with substitutions, is computed as $\Pi = 366.46$. Because the holding cost rate is high in this particular problem, the first solution creates a partial “pooling effect” for demands of Products 1 and 2, and 1 and 3 by channeling 10% of Product 1’s demand through substitutions. This in turn decreases the inventory costs, and results in a slightly better total profit. We note that the optimal solution of $Q = (352, 422, 423)$ starts the period with a total of 1198 units of inventory, and the 99% fill-rate solution’s initial total inventory is equal to 1230 units.

6.2.3. Randomly Generated Problems

In this section we consider a larger set of randomly created problems to study the impact of substitutions on system’s profit performance.

We consider 4-product problems with identical costs, four demand scenarios ((10,10,10,10), (15,15,5,5), (12,12,8,8), and (20,10,5,5)), four levels of substitution costs (0%, 5%, 10%, and 25% of products’ profit margins), and nine market share dependent profit margin scenarios where $\pi_i = \left( A - B \frac{\lambda_i}{\sum_j \lambda_j} \right)$ with $(A, B) \in \{(0.2, 0.3), (0.2, 0.2), (0.2, 0.1), (0.1, 0.15), (0.1, 0.1), (0.1, 0.05), (0.05, 0.075), (0.05, 0.05), (0.05, 0.025)\}$. We note that, according to above expression, profit margins are negatively correlated with market shares.

The problem generation scheme results in 144 test problems with profit margins that are negatively correlated with market shares, and identical product costs. The customer choice model is again assumed to be Market-Share Based (see Section 5).

For comparison purposes, we create a benchmark solution for every problem instance. We first solve a given problem instance optimally, as outlined in Section 6.1, and then compute the value of the initial inventory by multiplying the optimal order-up-to levels by the corresponding product costs. We then find the fill-rate service level that would
result in the same inventory value when the order-up-to level of each product is determined, independently and by ignoring the effects of substitution, with this particular service level. We then report the profit performance of the optimization approach relative to the profit obtained in the benchmark solution.

In Figure 3, we analyze the relationship between the profit improvement and substitution probability $\theta$. For every value of $\theta$ in set (60%, 70%, 80%, 90%, 100%), we solve the 144 test problems with $T = 20, h = 1.37\%$, and minimum direct service level of 40%, and report the average profit improvement over the benchmark solutions. The results presented in Figure 3 indicate that, by accounting for substitutions, the performance of the inventory system can be substantially improved. As expected, the higher the substitution probability, the larger the profit improvement.

In Figure 4, we analyze the relationship between the profit improvement and minimum direct service level requirement. For every value of in set (40%, 50%, 60%, 70%, 80%), we solve the 144 test problems with $T = 20, h = 1.37\%$, and $\theta = 60\%$, and report the average profit improvement over the benchmark solutions. The results presented in Figure 4 show that minimum direct service level requirement can have a significant impact on the profit improvement, and it is difficult to achieve substantial improvements when the inventory
Figure 4: Profit Improvement and Minimum Direct Sales Requirement: Negatively Correlated Profit Margin and Market Share.

Figure 5: Profit Improvement and Substitution Probability: Positively Correlated Profit Margin and Market Share.
system operates under a high level of minimum direct service level requirement.

Finally in Figure 5, we present the analysis of Figure 3 when profit margins and market shares are positively correlated, i.e., $\pi_i = \left( A + B \sum \lambda_j \right)$. The results presented in Figure 5 indicate that when profit margins and market shares are positively correlated, it is difficult to substantially improve the performance of the inventory system by accounting for substitutions.

7. Conclusions

In this paper, we consider the inventory management problem of a product category with stock-out based dynamic demand substitutions and lost sales. Although the retailer can only indirectly affect customers’ decisions through his inventory management decisions, as discussed in the literature, ignoring product substitutions in managing the inventories may result in sub-optimal performance. We present an approximate approach to find the order-up-to levels in a profit maximization setting with profit margins, inventory holding and substitution costs, and service level constraints. Through a computational study, we show that, by explicitly accounting for substitutions, the performance of the inventory system can be significantly improved in certain problem instances.

Acknowledgments

This research has been sponsored by TÜBİTAK grant 106K175.

References


