European Journal of Operational Research 300 (2022) 58–72

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Production, Manufacturing, Transportation and Logistics

An inventory model with discounts for omnichannel retailers of slow moving items

Adriana F. Gabor[a,d,*], Jan-Kees van Ommeren[b], Andrei Sleptchenko[c,d]

[a] Department of Applied Mathematics, Khalifa University, Abu Dhabi, United Arab Emirates
[b] Stochastic Operations Research, Department of Applied Mathematics, University of Twente, the Netherlands
[c] Department of Industrial and System Engineering, Khalifa University, Abu Dhabi, United Arab Emirates
[d] Research Center on Digital Supply Chain and Operations Management, Khalifa University, Abu Dhabi, United Arab Emirates

ARTICLE INFO

Article history:
Received 3 August 2020
Accepted 6 July 2021
Available online 13 July 2021

Keywords:
Inventory
Supply chain management
Omnichannel
Discounts
Critical level
Lost-sales

ABSTRACT

In this paper, we study an inventory model for an omnichannel retailer, that is, a retailer that sells items both via brick-and-mortar stores and online. Online items are delivered from a warehouse, which also replenishes the stores. When the inventory in a store drops below a certain level, the retailer offers customers a discount for purchasing online. In this way, the retailer can save items for customers who need the item immediately and thus avoid lost sales. For this model, we propose an approximation method for calculating the average inventory costs for one store and one warehouse and an optimization procedure for the case of more stores. Using extensive numerical experiments, we show that the approximations are very close to the performance measured via simulation. Finally, we show that by adopting the discounts policy proposed in this paper, the retailer can reduce its total cost, on average, by 8.5% compared to the no-discounts policy.

1. Introduction

In recent years, an increasing number of retailers have deployed an omnichannel strategy, where customers can buy products in brick-and-mortar stores as well as online. For traditional stores, expanding to the online environment is a necessity for remaining competitive. At the same time, online retailers discovered the benefit of showrooms, where customers can physically see the items they are interested in before purchasing them (Bell, Gallino, & Moreno, 2018; 2020). However, from a supply chain perspective, adopting an omnichannel strategy can lead to significant complexities, as discussed by Hübner, Holzapfel, & Kuhn (2016).

Integration of the inventory for the online and offline channels is a key element for an omnichannel retailer (Bendoly, Blocher, Bretthauer, & Venkataramanan, 2007). Most of the literature on fulfillment strategies and inventory integration focuses on pure online retailers or retailers who use the stores for fast fulfillment of online orders. In this situation, assigning orders dynamically to fulfillment centers is essential (Acimovic & Graves, 2017; Jasin & Sinha, 2015; Mahar, Bretthauer, & Venkataramanan, 2009). An adequate inventory pricing scheme that predicts the fulfillment at stores can further improve the revenue (Harsha, Subramanian, & Uichanco, 2019).

In this research, we focus on an omnichannel retailer of expensive, slow-moving items. The retailer's network consists of a warehouse and a network of brick-and-mortar stores. The stores have only a few items in stock, their main role being to let customers experience the product, and thereby increase the overall demand and decrease the returns. Companies such as WarbyParker.com and Amazon have already successfully used this strategy (Bell, Gallino, & Moreno, 2020). Note that the same policy is being used by stores selling electronic appliances and furniture, which have limited space.

In this paper, we propose a mathematical model that allows studying the advantages of offering discounts to store visitors for switching to the online channel. In our model, we assume that the stores follow an $(S-1, S, IC)$ inventory policy, where, every time an item is sold, another item is ordered from the warehouse. The stores only sell items to store visitors; they do not fulfill online orders. Note that the $(S-1, S)$ policy is suitable for slow-moving items. If customers visiting the store find the inventory level below the critical level IC, they are offered to switch to the online channel in exchange for a discount. This basically means that their order will be fulfilled at a later time from the warehouse. Customers

* Corresponding author at: Department of Applied Mathematics, Khalifa University, Abu Dhabi, United Arab Emirates.
E-mail addresses: adriana.gabor@ku.ac.ae (A.F. Gabor), j.c.w.vanommeren@utwente.nl (J.-K. van Ommeren), andrei.slecptchenko@ku.ac.ae (A. Sleptchenko).

https://doi.org/10.1016/j.ejor.2021.07.017
0377-2217/

who find the inventory depleted and do not switch to the online channel are lost. The main advantage of offering this discount is to prevent future lost sales due to customers who find the inventory at stores depleted and may be tempted to buy from a competing retailer. The discount acceptance probability is assumed to be known. The warehouse has two roles. First, it is used to fulfill online orders, that is, orders that are placed directly via the online channel, as well as orders of customers who switched to the online channel after accepting the discount at the store. Second, the warehouse is used to replace the items bought at the stores. At the warehouse, we assume an (R, Q)-policy with backlogging. The arrival processes of online and store customers are assumed Poisson, and all the lead times are assumed to be deterministic. Observe that the integration of online and offline inventories in this model takes place at the warehouse. To the best of our knowledge, the proposed model has not been studied before in the literature.

The paper is organized as follows. In Section 2, we revise the related literature and discuss the contribution of the present paper. Section 3 presents an approximation method for the long-run average costs in the special case of one store and one warehouse. In Section 4, we prove some properties of the objective function and propose a heuristic for the case of more stores based on decomposing the original problem into a series of optimization problems at the retailer level. Finally, we present extensive numerical results on the quality of the approximations and heuristic proposed and sensitivity analysis for the impact of discounts in Section 5. We conclude the paper with a discussion and managerial insights regarding the use of discounts in a multi-echelon, omnichannel setting.

2. Related literature

The model discussed in this paper is a two-echelon model with one warehouse and a set of stores. The warehouse follows an (R, Q) policy, while the stores follow a base-stock rationing policy $(S-1, S, IC)$ combined with discounts. There are two demand classes: offline demand, which is served directly from the stores if there are items on stock, and online demand, served from the warehouse. Customers who find no inventory and refuse the discount are lost. The problem studied in this paper is closely related to the omnichannel literature and three important research directions in inventory management: multi-echelon inventory models with lost sales at the lowest echelon, rationing policies, and pricing. We will revise the most related papers in all these research streams.

Improving the performance of the supply chain of omnichannel retailers is a booming area in operations management. E-fulfillment and distribution strategies together with inventory integration are key aspects for cost reduction and high customer satisfaction for these retailers. Mahar et al. (2009) have shown the importance of assigning orders to fulfillment centers in a dynamic way. Fulfillment decisions become very complex when orders might contain multiple items, having different availability at different locations. LP-based heuristics for optimizing the fulfillment of multi-item orders for online channels have been proposed by Acimovic & Graves (2015) and Jasin & Sinha (2015). The benefits of dynamic fulfillment strategies can be further enhanced if inventory is jointly optimized with fulfillment decisions, as shown by Acimovic & Graves (2017) and Govindarajan, Sinha, & Uichanco (2018). Besides location, pricing of inventory also plays an important role in omnichannel retail. Harsha et al. (2019) showed that inventory pricing can help balance inventory across the network, for example, by fulfilling more online orders from stores with low demand. Our paper differs from the literature as the online orders can only be fulfilled from the warehouse and not from the stores. Similar to Acimovic & Graves (2017), we are interested in finding the up-to-level at stores; however, we consider a continuous replenishment policy and focus of the joint replenishment of the warehouse and stores and not on the joint replenishment of the different fulfillment centers. Moreover, the demand for each channel is controlled through a discount for switching from the store to the online channel. To the best of our knowledge, the joint optimization of inventories with the possibility of channel migration has not been studied before.

The inventory system analyzed in this paper is, in essence, a two-echelon continuous review inventory system with a backlog possibility at the warehouse and lost sales at the stores. Due to the importance of these models in manufacturing and spare parts supply chains, the literature on multi-echelon systems is very vast. For extensive reviews of the multi-echelon literature, we refer to van Houtum (2006), Simchi-Levi & Zhao (2011), and de Kok et al. (2018). Here we shall comment only on papers with similar network structure (two echelons consisting of one warehouse and N stores/retailers) and related inventory policies.

Most of the multi-echelon papers on continuous review policies consider the situation where demand can be backlogged. One of the widely used approaches to calculate performance measures for multi-echelon systems is the METRIC model proposed in Sherbrooke (1968). In the METRIC approximation, the average delay at the warehouse is used to estimate costs at the stores instead of the real stochastic times. For a review of the method and articles that build upon it, we refer to Axsäter (2015). Axsäter (1990) provides an exact recursive procedure for calculating the long-run average costs in a system of one warehouse and N stores, in which all facilities follow an $(S-1, S)$ inventory policy. Axsäter (1993) and Axsäter (1998) derive the expected holding and shortage costs in a system with the same network structure and (R, Q) inventory policies at each facility. The exact probability distribution of the inventory levels in this system is studied in Axsäter (2000).

Lost sales models have received less attention in the multi-echelon literature. In one of the first papers on a multi-echelon system with lost sales, Nahmias & Smith (1994) find optimal solutions for a periodic model with zero replenishment times where some of the lost sales can be delayed. For the case of one warehouse and N stores, where all facilities follow an $(S-1, S)$ policy and excess demand at retailers is lost, Andersson & Melchiors (2001) propose an efficient queueing-based approximation that extends the METRIC-model described in Sherbrooke (1968). The authors model each store as an Erlang-loss queue, which permits the calculation of the loss probabilities and subsequently allows them to give an estimate of the arrival rate at the warehouse. An iterative heuristic is then used to calculate the base stock levels at facilities. Hill, Seifbarghy, & Smith (2007) calculate the average stock and the fraction of demand met at stores in a model where the stores follow an (R_i, Q) policy and the warehouse an $(SQ, (S-1)Q)$ policy. The model relies on the assumption that at any moment in time, each store can have at most one outstanding order at the warehouse and that the transportation time from the warehouse to store is not less than the lead time of the warehouse. Our model is similar to the one studied in Andersson & Melchiors (2001) and Hill et al. (2007); however, in this paper, the stores follow a rationing policy, and some of the customers at the stores can be served by the warehouse if they accept a discount. Moreover, in our case, the warehouse has an (R, Q) policy, which requires a different analysis of the delay at the warehouse.

Our model considers two classes of customers: the customers who visit the stores and the online customers. One of the most commonly used inventory policies when dealing with different classes of customers is the so-called rationing policy (Kleijn & Dekker, 1999). In order to improve the service level of higher priority customers, orders from lower-priority customers are backordered or rejected when inventory reaches a certain critical level.

We will denote a rationing policy by the inventory policy followed by the critical level. Nahmias & Demmy (1981) derived the first approximation for the expected number of backorders and fillrates in a continuous (R, Q, K) inventory system with two demand classes modeled by Poisson processes and deterministic lead times. Orders for the low priority customers are backordered when the inventory on hand drops below K. Deshpande, Cohen, & Donohue (2003) extend the model in Nahmias & Demmy (1981) by allowing multiple outstanding orders in the pipeline. Arslan, Graves, & Roemer (2007) further generalize this model to multiple demand classes with different shortage costs or service requirements. Fadloğlu & Bulut (2010) propose an approximating embedded Markov Chain to estimate the steady-state probabilities in an $(S-1, S, K)$ system with constant lead times, two demand classes, and backlogged demand. For the same system, the exact distribution of the response times of low priority customers and an approximation of the fillrate for the high priority customers are derived in Gabor, van Vianen, Yang, & Axsäter (2018). For the case of exponential lead times, Vicil & Jackson (2016) calculate the steady-state distribution of the on-hand inventory and the number of backorders of each class. They show that under certain independence conditions, the same balance equations hold in the case of general lead times.

Ha (1997) was one of the first to analyze a critical level policy in a lost-sales environment. He analyzed a make-to-stock system with n demand classes and exponential lead times. By modeling the system as an $M/M/1/S$ queue with state-dependent service rates, he was able to show that in this setting, a base stock policy combined with a rationing policy is optimal. Melchiors, Dekker, & Kleijn (2000) calculate the expected inventory costs in an (s, Q, K) system with two demand classes and lost sales. In our model, a rationing policy is used at the stores. The main difference between our policy and the ones in the literature is that we do not reject one of the customer classes but allow customers of one class to change the delivery channel in exchange of a discount, if inventory drops below the critical level. This is equivalent to allowing a delay in delivery, with the delayed item being delivered from another echelon. Furthermore, we consider the impact of this policy in a two-echelon setting, whereas, to the best of our knowledge, most literature addresses rationing in one echelon.

The last stream of literature related to our model is the one combining pricing decisions and inventory theory. It is well known that offering economic incentives to backorder can result in significant gains (Bhargava, Sun, & Xu, 2006; Cheung, 1998; DeCroix & Arreola-Risa, 1998; Ding, Kouvelis, & Milner, 2006). The paper most related to our paper is the one of Cheung (1998), in which the impact of offering discounts for delayed service in an (R, Q) system is analyzed. Cheung (1998) proposes a policy in which, if the on-hand inventory drops below a certain level within a period T after the reorder time epoch, a discount is offered to customers in exchange for accepting a delayed service. Under the assumptions of at most one outstanding order and backlog being smaller than Q at any time, Cheung (1998) derives the expected costs of this system and, using computational experiments, concludes that discounts can lead to significant savings. DeCroix & Arreola-Risa (1998) study the benefits of offering an economic incentive to all the customers who see an inventory on hand below r in an (s, S) inventory system with uncertain supply. Lei, Jasin, & Sinha (2018) propose an LP-based heuristic for maximizing the profit of an e-commerce retailer who has to decide in each period what price to ask and from which facility to fulfill online orders. Similar to Cheung (1998) and DeCroix & Arreola-Risa (1998), in our model, a discount is offered at the retailer for accepting a delay, which in our case, is equivalent to accepting an online delivery. However, we extend these models to two echelons and study the impact of the discounts in a multi-retailer setting.

Contribution of the paper: The model analyzed in this paper allows to study whether a discount policy for switching to an online channel is beneficial in a two-echelon network, where the online orders are fulfilled from a warehouse, while physical items can be purchased in person at the stores. The results of extensive numerical experiments indicate an 8.5% cost reduction on average (with the maximum of 19.8%) compared to the case where no discounts are offered. They also indicate that the discount policy is more effective for higher holding costs at stores, a larger number of stores, and for high lost sales costs.

The analysis presented in this paper enhances the literature in several ways. First, we extend the multi-echelon lost sales models of Andersson & Melchiors (2001) and Hill et al. (2007) by considering a critical level policy at the stores, controlled by a discount and a (R, Q) policy at the warehouse. For a model with one warehouse and one store, building on the METRIC model, we propose a queueing-based approximation to find the expected delay at the warehouse. The approximation is based on a recursive procedure for which we prove convergence. Note that the approximation procedure proposed in Andersson & Melchiors (2001) for a simpler model, without critical levels, and with an $(S-1, S)$ policy at the warehouse, is shown to converge under the conjecture that the base-stock levels levels at the stores increase when the delay at the warehouse. The convergence proof for our procedure does not rely on this conjecture. For the model with more stores, we propose a novel optimization heuristic based on the discretization of the expected delay at the warehouse. This allows decomposition of the network problem in a series of independent problems per retailer and one problem for the warehouse. We show numerically that the results obtained via this procedure are close to the results obtained via simulation.

3. Problem formulation for one warehouse and one store

In this section, we consider a simplified model of an omnichannel retailer who has only one brick-and-mortar store which is replenished from one warehouse. The retailer sells expensive items and has limited storage space; hence she prefers to keep only a few items in store (showroom). The store follows an $(S-1, S, IC)$ inventory policy, meaning that when an item is sold, it is replenished with another item ordered from the warehouse. If the inventory on hand at the store is less or equal to IC, customers are offered a discount d for a delayed delivery from the warehouse. A customer accepts this offer with probability $p_a(d)$. We assume that $p_a(\cdot)$ is increasing in the discount. If the customer rejects the discount, the retailer will give an item from the stock, if available. Customers who find no inventory at the store and do not order online are lost. Through the discount, the retailer tries to reduce the number of lost customers at the store. We assume that the extra costs incurred for switching to the online channel are covered by the discounts. Customers can also buy the same item directly online without visiting the shop. In this case, the purchased item is sent to them from the warehouse, and no discount is offered.

The warehouse satisfies three types of orders: orders placed by online customers, orders originating from store visitors who accepted the discount, and replenishment orders from the store. The warehouse follows an (R, Q) policy, where orders that find the inventory at the warehouse depleted are backlogged. Fig. 1 summarizes the main features of the supply network.

We assume that requests at the store follow a Poisson process with rate λ_s, while online requests at the warehouse follow a Poisson process with rate λ_o. The lead times at the store and warehouses are denoted by L_s and L_w, and assumed constant. The store and the warehouse incur a holding cost per time unit and per item in stock of h_s and h_w, respectively. Each lost customer has an associated cost of l_s. Finally, there is a cost of $c_t(w)$ for transport-

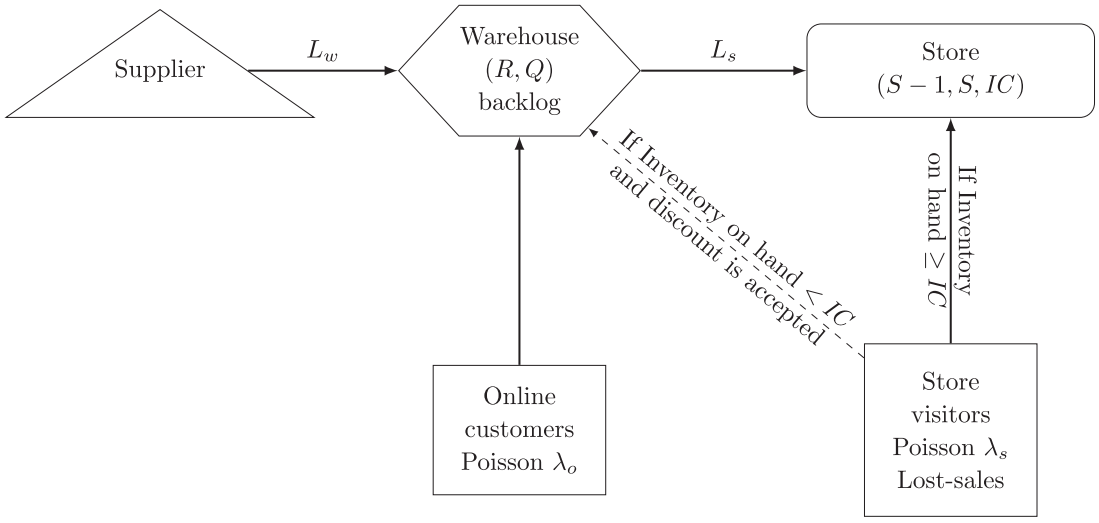

Fig. 1. Network for an omnichannel retailer with one store.

ing an item between supplier and warehouse and a cost of $c_t(w, s)$ for transporting an item between warehouse and store. These costs are incurred for all the items sold in store, as well as for the orders that switched from the store to the online channel. In our model, we assume that the extra delivery costs between stores and home due to a customer switching to the online channel are already incorporated in the offered discount. The costs for delivering orders that were originally placed online are not included in the model, as we assume that all online orders are eventually fulfilled, and hence these costs do not affect the optimization.

The goal is to find the vector $\mathbf{u} = (R, S, IC, d)$ that minimizes the long-run average costs for the retailer. The order quantity Q at the warehouse is assumed given, as it is usually decided based on the annual demand, holding costs, and transportation costs between the warehouse and its supplier.

3.1. Approximating model for one store

We approximate the $(S-1, S, IC)$ inventory system at the store by an $M(n)/G/S/S$ system as follows. The arrival process at this queue is formed by all the customers who, upon arrival, see an inventory higher than IC and by the customers who, upon arrival, see an inventory below IC and refuse the discount. Every time a customer leaves with an item, a service is started in the queueing model. Hence, having k items in stock at the store is equivalent to having $S-k$ busy servers in the queueing system. Customers who see the inventory depleted at the store, or equivalently, all the S servers in the queueing system busy, are lost. Thus, the arrival process at the $M(n)/G/S/S$ queue is a Poisson process with rate $\lambda(j, d)$ depending on the number of busy servers j as follows:

$$\lambda(j,d) = \begin{cases} \lambda_s & \text{when } j < S - IC \\ \lambda_s(1 - p_a(d)) & \text{when } S - IC \leq j \leq S. \end{cases} \tag{1}$$

The service times are assumed to be independent and equal to $L_s + T$, where T is a random variable representing the delay at the warehouse. In reality, the return times of store orders are dependent on the inventory on hand at the warehouse. Following an idea similar to the METRIC model, the service rate μ in the $M(n)/G/S/S$ queue is given by

$$\frac{1}{\mu} = L_s + E(T). \tag{2}$$

An approximation of $E(T)$ will be presented in Section 3.2. Assume for a moment that μ is known. Let $P_k(\mathbf{u})$ be the steady-state probability that the store has k orders outstanding with the warehouse when the discount is d, or, equivalently, that the inventory at the store is equal to $S-k$. By Theorem 3 in Brumelle (1978), in an $M(n)/G/S/S$ system,

$$P_k(\mathbf{u}, E(T)) = \frac{\frac{\pi_k}{\lambda(k,d)}}{\sum_{j=0}^{S} \frac{\pi_j}{\lambda(j,d)}} \tag{3}$$

where

$$\pi_k = \pi_0 \prod_{j=1}^{k} \frac{\lambda(j,d)}{j\mu}$$

and $\pi_0 + \ldots + \pi_k = 1$.

Replacing the rate $\lambda(j, d)$ by (1) we obtain:

$$\pi_k = \begin{cases} \pi_0 \frac{\lambda_s^k}{k!\mu^k}, & \text{for } k < S - IC \\ \pi_0 \frac{\lambda_s^k}{k!\mu^k}(1 - p_a(d))^{k-S+IC+1}, & \text{for } S - IC \leq k \leq S \end{cases}$$

and

$$\pi_0 = \left[\sum_{k=0}^{S-IC} \frac{\lambda_s^k}{k!\mu^k} + \sum_{k=S-IC}^{S} \frac{\lambda_s^k}{k!\mu^k}(1 - p_a(d))^{k-S+IC+1} \right]^{-1}.$$

Note that the steady-state probabilities depend on the distribution of $L_s + T$ only through the mean. The next section explains the calculation of $E(T)$.

3.2. Expected fulfillment time of a store order

The warehouse uses an (R, Q) inventory model with backorders, that is, every time the inventory position IP_w drops to R, an order of Q is ordered at the supplier. The lead time from the supplier to the warehouse is assumed deterministic and equal to L_w.

Recall that the orders arriving at the warehouse are of two types: orders placed directly online, which arrive according to a Poisson process at rate λ_o, and orders originating from the customers who visit the store, except the ones who found the stock depleted and refused a delayed delivery. Assuming that the orders originating from the store also follow a Poisson process, with rate $\lambda_s[1 - P_S(\mathbf{u})(1 - p_a(d))]$, the arrival process at the warehouse is Poisson with rate:

$$\lambda_w(\mathbf{u}, E(T)) = \lambda_o + \lambda_s[1 - P_S(\mathbf{u}, E(T))(1 - p_a(d))]. \tag{4}$$

Let $IL_w(t)$ be the inventory on hand at the warehouse at time t. At any time $t \geq L_w$,

$$IL_w(t) = IP_w(t - L_w) - D_w(t - L_w, t). \tag{5}$$

where $D_w(a, b)$ denotes the demand at the warehouse during the time interval (a, b).

Tag an order arriving at the warehouse from the store at an arbitrary time t. Assume that at time t, $IL_w(t) = -k$ with $k \geq 1$, and $IP_w(t - L_w) = l$. Based on (5), we conclude that $D_w(t - L_w, t) = k + l$. These $l + k$ arrivals generated $1 + \lfloor \frac{R+k}{Q} \rfloor$ orders at the supplier: the first order is generated after the arrival of the first $l - R$ orders, when $IP_w = R$, after which the rest of $R + k$ orders arrive at the warehouse. The tagged order is fulfilled right after the $N_o(k) = \lfloor \frac{k}{Q} \rfloor + 1$ -th order is delivered from the supplier. This order is placed right after the $l - R + (N_o(k) - 1)Q$-th order after time $t - L_w$ arrives at the warehouse. Since orders arrive at the warehouse according to a Poisson process, the expected time between two arrivals in the interval $[t - L_w, t)$ is $\frac{L_w}{k+l+1}$ (see Tijms, 2003). Hence, the tagged store order is delivered after $t - L_w + (l - R + (N_o(k) - 1)Q)\frac{L_w}{k+l+1} + L_w - t = (l - R + (N_o(k) - 1)Q)\frac{L_w}{k+l+1}$.

By the law of total expectation, (5) and the fact that in an (R, Q) system, the inventory position is uniformly distributed on $[R + 1, R + Q]$ we obtain

$$\begin{aligned}
&E(T) = \\
&= \frac{1}{Q}\sum_{k=0}^{\infty}\sum_{l=R+1}^{R+Q} E(T|IP_w(t - L_w) = l, IL_w(t^-) = -k)P(IP_w(t - L_w) \\
&= l, IL_w(t^-) = -k) \\
&= \frac{1}{Q}\frac{1}{Q}\sum_{k=0}^{\infty}\sum_{l=R+1}^{R+Q} E(T|IP_w(t - L_w) = l, IL_w(t^-) = -k)P(D_w(t - L_w, t) = l + k) \\
&= \frac{1}{Q}\sum_{k=0}^{\infty}\sum_{l=R+1}^{R+Q} (l - R + (N_0(k) - 1)Q)\frac{L_w}{k + l + 1}\mathbf{po}(\lambda_w(\mathbf{u}, E(T))L_w, k + l),
\end{aligned} \tag{6}$$

where $\mathbf{po}(\lambda L, k) = \frac{(\lambda L)^k}{k!}e^{-\lambda L}$.

Observe that the expression of $E(T)$ derived above depends on $\lambda_w(\mathbf{u})$, which in turn depends on $E(T)$. In order to calculate $E(T)$ for a fixed $\mathbf{u}$, we propose the following recursive procedure:

Next, we show that this procedure converges.

Lemma 1. *The function $f : \mathbf{R}_+ \mapsto R_+$ given by*

$$f(\lambda) = \frac{1}{Q}\sum_{k=0}^{Q-1}\sum_{n=0}^{\infty}\sum_{\ell=R+1}^{R+Q}\frac{\ell - R + nQ}{\ell + nQ + k + 1}\mathbf{po}(\lambda L_w, \ell + nQ + k)$$

is increasing.

Proof. We rewrite Eq. (6) to

$$\begin{aligned}
f(\lambda) &= \frac{1}{Q}\sum_{k=0}^{Q-1}\sum_{n=0}^{\infty}\sum_{\ell=R+1}^{R+Q}\frac{\ell - R + nQ}{\ell + nQ + k + 1}\mathbf{po}(\lambda L_w, \ell + nQ + k) \\
&= \frac{1}{Q}\sum_{k=0}^{Q-1}\sum_{m=R+1}^{\infty}\frac{m - R}{m + k + 1}\mathbf{po}(\lambda L_w, m + k).
\end{aligned}$$

Now we note that the function $f_k(m) = (m - R)/(k + m + 1)$ is increasing in m, so we may write

$$f_k(m) := \frac{m - R}{k + m + 1} = f_k(m - 1) + \Delta_{km},$$

where $\Delta_{km} > 0$. Continuing, we find

$$f_k(m) = \sum_{\tilde{m}=R+1}^{m}\Delta_{k\tilde{m}}$$

and

$$\begin{aligned}
f(\lambda) &= \frac{1}{Q}\sum_{k=0}^{Q-1}\sum_{m=R+1}^{\infty}\sum_{\tilde{m}=R+1}^{m}\Delta_{k\tilde{m}}\mathbf{Po}(\lambda L_w, m + k) \\
&= \frac{1}{Q}\sum_{k=0}^{Q-1}\sum_{\tilde{m}=R+1}^{\infty}\Delta_{k\tilde{m}}\sum_{m=\tilde{m}}^{\infty}\mathbf{po}(\lambda L_w, m + k).
\end{aligned}$$

Observe that $\sum_{m=\tilde{m}}^{\infty}\mathbf{po}(\lambda L_w, m + k) = Po(\lambda L_w, \tilde{m} + k)$ is increasing in λ, as the family of Poisson distributions is stochastically increasing (see Shaked & Shanthikumar, 2007, Example 8.A.2). As $\Delta_{km} \geq 0$, it follows that $f(\lambda)$ is increasing in λ. □

Theorem 1. *The sequence $(E_{2k+1}(T))_{k\geq 0}$ constructed in Algorithm 1 converges.*

Algorithm 1

1: Step 1: Set $k = 0$, $E_{-1} = E_{-2} = -1$, $E_0(T) = 0$.
2: **while** $|E_k(T) - E_{k-2}(T)| > \epsilon$ **do**
3: Step 2 Calculate μ, $P_S(\mathbf{u}, E_k(T))$, $\lambda_w(\mathbf{u}, E_k(T))$.
4: Step 3: Calculate $E_{k+1}(T)$ based on (6)
5: Step 4: Set $k = k + 1$
6: **end while**

Proof. Fix $\mathbf{u}$. We show that $(E_{2k+1}(T))_{k\geq 0}$ is monotonic and bounded. Assume that $E_1(T) \leq E_3(T)$. We show by induction that if $E_{2k-1}(T) \leq E_{2k+1}(T)$ for some $k > 1$, then $E_{2k+1}(T) \leq E_{2k+3}(T)$.

As $P_S(\mathbf{u}, E(T))$ is increasing in $\frac{1}{\mu}$, it follows that $P_S(\mathbf{u}, E(T))$ is increasing in $E(T)$. Relation (4) implies that $\lambda_w(\mathbf{u}, E(T))$ is decreasing in $P_S(\mathbf{u}, E(T))$, hence decreasing in $E(T)$. Thus, $\lambda_w(\mathbf{u}, E_{2k+1}(T)) \leq \lambda_w(\mathbf{u}, E_{2k-1}(T))$. As the parameters of the inventory policy at the warehouse are fixed, an increase in arrival rate corresponds to an increase in the average delay by Lemma 1. Hence $E_{2k+1}(T) \leq E_{2k}(T)$. By the monotonicity of λ_w, we find that $E_{2k+1}(T) \leq E_{2k+3}(T)$.

Similarly, one can show that if $E_1(T) \geq E_3(T)$, then $(E_{2k+1})_{k\geq 1}$ is decreasing. As $E(T)$ is bounded between 0 and the delay corresponding to $\lambda = \lambda_o + \lambda_s$, the sequence $(E_{2k+1})_{k\geq 1}$ is convergent. □

3.3. Total costs

The total costs comprise the costs at the store, the costs due to discounts, the costs at the warehouse, and the transportation costs.

Since the store incurs holding costs of h_s per time unit an item is in stock and the probability of having k items in stock equals $P_{S-k}(\mathbf{u})$, the expected holding cost at the store is equal to $\sum_{k=0}^{S} P_{S-k}(S, IC, d)kh_r$.

As the probability that a customer is lost equals $(1 - p_a(d))P_S(\mathbf{u})$, the expected cost due to lost customers equals $l_s\lambda_s(1 - p_a(d))P_S(\mathbf{u})$.

Recall that the retailer offers a discount d to all customers who arrive when the inventory on hand is below IC. The probability that the inventory at the store is less or equal to the critical level is $q(\mathbf{u}) = \sum_{k=S-IC}^{S} P_k(\mathbf{u})$. Hence, the total expected costs incurred due to discounts are $d\lambda_s p_a(d)q(\mathbf{u})$.

The warehouse incurs holding costs h_w per unit hold and backorder costs b for each item that is backordered (from the store or from online customers). The backorder costs at the warehouse are given by $bE(IL_w^-)$, where $E(IL_w^-) = \lambda_w(\mathbf{u})E(T)$ by Little's law. Based on (5)

$$E(IL_w^+) = E(IL_w^-) + R + \frac{Q}{2} - \lambda_w(\mathbf{u}, E(T))L_w. \tag{7}$$

Finally, there are transportation costs for every echelon and from the warehouse to the brick-and-mortar stores. As all the customers who initially ordered online are assumed to be served, the

transportation costs of their items do not influence the optimal solution of the problem. Therefore we will not further consider these costs in the optimization. Recall that the delivery costs for items ordered by customers who switched from the offline to the online channel are assumed included in the discounts.

The transportation costs per time unit between supplier and warehouse are $c_t(w)\lambda_w$ and between warehouse and store are $c_t(w,s)\lambda_s[1-P_S(\mathbf{u})(1-p_a(d)]$.

Summarizing all the costs described above, the total costs (TC) incurred in the system are equal to:

$$TC = h_s \sum_{k=0}^{S} P_{S-k}(\mathbf{u})k + l_s\lambda_s P_S(\mathbf{u})(1-p_a(d)) + h_w E(IL_w^+) + bE(IL_w^-) + d\lambda_s p_a(d)q(\mathbf{u}) + c_t(w,s)\lambda_w(\mathbf{u}).$$

Remark 1. The optimal value of S and R that achieve minimal total costs for one warehouse and one retailer can now be obtained via direct enumeration. For each value of S and R, the recursive procedure in Algorithm 1 for calculating the expected delay at the warehouse is guaranteed to converge. This also holds for the optimal levels S^* and R^*. In Section 5.1, we will show that the expected delay at the warehouse converges to the expected delay at the warehouse obtained by simulation. Note that the convergence of the sequence of the expected delays at the warehouse constructed by the recursive procedure proposed in Andersson & Melchiors (2001) relies on the conjecture that the up-to-level at the retailer is increasing in the delay at the warehouse. Our proof does not rely on this conjecture.

4. Network of more stores

In this section, we generalize the results in the previous section to a network of N stores and a warehouse. The demand at each store i is independent of the demand at other stores and follows a Poisson process with rate λ_i. As in the previous section, each store i adopts an (S_i-1, S_i, IC_i) inventory policy with discount d, and the warehouse adopts an (R,Q) policy. We assume the discount is the same for all stores. Let $(\mathbf{S},\mathbf{IC}) = (S_i, IC_i)_{i=1,\ldots,n}$. The goal is to find the vector $\mathbf{u} = (R, Q, \mathbf{S}, \mathbf{IC}, d)$ that minimize the total costs at the warehouse and the stores.

As in the previous section, the (S_i-1, S_i, IC_i) system at each store i is approximated by an $M(n)/G/S/S$ system, with arrival rate defined as in (1). The flow of orders from store i to the warehouse is approximated by a Poisson process with rate $\tilde{\lambda}_i = \lambda_i(1-P_S(\mathbf{u}_i))(1-p_a(d))$. Thus, the orders (online and offline) are assumed to arrive at the warehouse according to a Poisson process with rate

$$\lambda_w(\mathbf{S},\mathbf{IC}) = \lambda_o + \sum_{i=1}^{N} \tilde{\lambda}_i. \tag{8}$$

The service rate in the $M(n)/G/S/S$ is defined as in the previous section, the main difference being that the arrival rate at the warehouse is $\lambda_w(\mathbf{S},\mathbf{IC})$.

The total costs that need to be minimized are given by the costs of all stores, the costs at the warehouse, and the total transportation costs.

4.1. Optimization procedure

Before presenting the optimization procedure in detail, we discuss a few monotonicity results regarding (R,Q) inventory models with a backlog and $(S-1,S)$ inventory models with lost sales.

The first statement in the following Lemma is a special case of Theorem 1 in Song, Zhang, Hou, & Wang (2010), while the second is a special case of Corollary 1 (f) in Federgruen & Wang (2013).

Lemma 2. *Consider an (R,Q) inventory model with Poisson arrivals and backlogging, with the order quantity Q fixed.*

(a) *For two constant lead times L_1 and L_2 such that $L_1 \le L_2$, the corresponding optimal reorder levels $R^*(L_1)$ and $R^*(L_2)$ satisfy $R^*(L_1) \le R^*(L_2)$.*
(b) *The optimal R^* is increasing in the Poisson arrival rate.*

Lemma 2 implies that the maximum value of R^* in the network with one warehouse and N stores is attained for $\lambda = \lambda_o + \sum_{i=1}^{N}\lambda_i$ and lead time L_w.

Next consider the $(S-1,S)$ inventory model at a store i, for a fixed delay at the warehouse $\hat{T}$ and arrival rate given by (1). Remark that $\hat{T}$ is the main factor through which the warehouse impacts the stores. If the delay at the warehouse was known, the stores can optimize their system without knowing the explicit value of R.

The expected costs per time unit at store i are given by

$$\begin{aligned} C_i(\hat{T},S_i,IC_i,d) &= h[S-(1-P_S(\hat{T},S_i,IC_i,d)(1-p_a(d)))\lambda_i(\hat{T}+L_i)] + \\ &\quad + P_S(\hat{T},S_i,IC_i,d)\lambda_i l_i \\ &= hS_i + P_S(\hat{T},S_i,IC,d)\lambda_i(l_i + h(\hat{T}+L_i)(1-p_a(d))) \\ &\quad - h\lambda_i(\hat{T}+L_i)(1-p_a(d)) \end{aligned} \tag{9}$$

Lemma 3. *The cost function at store i has the following properties:*

(a) *Let $d_{\min}$ be a discount such that $p(d_{\min})=0$ and $d_{\max}$ a discount for which $p_a(d_{\max})=1$. Then, $C_i(\hat{T},S,IC,d_{\min})$ and $C_i(\hat{T},S,IC,d_{\max})$ are convex in S.*
(b) *For d fixed, $C_i(\hat{T},S,S,d)$ is convex in S.*
(c) *$C_i(\hat{T},S,IC,d)$ is decreasing in d.*

Proof.

(a) As $p_a(d_{\min})=0$, the $(S-1,S,IC)$ system at the store reduces to an $(S-1,S)$ inventory system. The corresponding queueing model at store i is an $M/G/S/S$ with constant rate λ_i. The convexity in S of $C_i(\hat{T},S,IC,d_{\min})$ follows from the convexity in the number of servers of the loss probability in an $M/G/c/c$ queueing model (Messerli, 1972).
As $p_a(d_{\max})=1$, all customers accept the discount and are redirected to the warehouse. Hence, for $S>IC$, the $(S-1,S,IC)$ inventory system at the store is equivalent to an $(S-IC-1,S-IC)$ inventory system. The corresponding queueing model is an $M/D/S-IC/S-IC$ with constant rate λ_r. The convexity in S of $C(\hat{T},S,IC,d_{\max})$ follows from the convexity in the number of servers of the Erlang formula. If $S=IC$ and $p_a(d_{\max})=1$, all customers are directed to the warehouse, hence the store's costs are equal to the holding costs, which are linear, thus convex.
(b) When $IC=S$, the actual arrival rate at store i is $\lambda_i p_a(d)$. The inventory system can be modeled as an $M/G/S/S$ queue, for which the loss probability is known to be convex in S.
(c) From (9), it follows that it suffices to show that for given μ, $P_S(\hat{T},S,IC,d)$ is decreasing in d. By simple algebraic manipulations of (3), one can show that

$$P_S(\hat{T},S,IC,d) = \frac{\frac{\lambda^S}{S!\mu^S}}{\sum_{k=0}^{S-IC-1}(1-p_a(d))^{-IC}\frac{\lambda^k}{k!\mu^k} + \sum_{k=S-IC}^{S}\frac{\lambda^k}{k!\mu^k}(1-p_a(d))^{k-S}},$$

where any sum from a to b with $b<a$ is considered 0. It follows readily that $P_S(\hat{T},S,IC,d)$ is decreasing in d. □

Observe that Lemma 3 a) and c) implies that for $d \in [d_{\min}, d_{\max}]$, such that $p_a(d_{\min})=0$ and $p_a(d_{\max})=1$, the cost function of store i is between two convex functions in S, $C_i(\hat{T},S,IC,d_{\min})$ and $C_i(\hat{T},S,IC,d_{\max})$. We were not able to prove the quasi-convexity of $C_i(\hat{T},S,IC,d)$ in S, although we observed

quasi-convexity in all our experiments. In our optimization procedure, we will assume that a stronger statement than Lemma a) and c) holds, namely that the optimal base stock levels at stores are decreasing in d.

The difficulty of optimizing the entire network relies on the fact that the up-to-level at each store influences the other stores through the impact they have on the delay at the warehouse. In order to facilitate an independent optimization at each store, we propose to discretize the possible delays $E(T)$ at the warehouse. Note that $E(T) \in [0, L_w]$.

For each value of the discount d, we proceed as follows. For a chosen value $\eta \in \mathbb{N}$, we discretize $[0, L_w]$ with a step $a = \lfloor \frac{L_w}{\eta} \rfloor$. Let $\mathcal{T} = \{ka | k \in \mathbb{N}, 0 \leq k \leq \eta\}$. For each $\hat{T} \in \mathcal{T}$, we solve a separate inventory optimization for each store to find the optimal values (S_i^*, IC_i^*). The optimal parameters of the inventory policy at each store are found by enumeration, as described in Algorithm 2.

Algorithm 2 Optimization procedure for store i.

1: Let d be a specific discount value
2: Let $\hat{T}$ be a given deterministic delay at the warehouse
3: **for** $i \in N$ **do**
4: Set $\tilde{S}_i$ optimal base stock level at store i if $p_a(d) = 0$, $IC_i = 0$ and lead time $L_i + L_w$
5: Set the lead time for store i equal to $\frac{1}{\mu_i} = L_i + \hat{T}$
6: **for** $S_i \in [0, \tilde{S}_i]$ and $IC_i \in [0, S_i]$ **do**
7: Calculate $C_i(\hat{T}, S_i, IC_i, d)$ using (9)
8: **end for**
9: Find (S_i^*, IC_i^*) for which minimal costs at store i are achieved.
10: **end for**

Remark that in limiting the optimal base stock to $[0, \tilde{S}_i]$, we assumed that the optimal base-stock at a store is decreasing in d.

Knowing the base-stock and critical levels at the stores allows estimation of the arrival rate at the warehouse through (8) and, subsequently, the level R that results in the expected delay closest to $\hat{T}$. Finally, we choose the $\hat{T}^* \in \mathcal{T}$ that gives minimum total costs. A detailed description for finding the optimal delay for a given discount d and a discretization $\mathcal{T}$ is given in Algorithm 3.

Observe that in limiting the values of R in Line 9, we used Lemma 2. Assume the optimal value of the delay is $\hat{T}^* = ka$. We continue the search for a better delay by refining the discretization as follows. Consider the interval $I = [\max\{0, (k-1)a\}, \min\{(k+1)a, L\}]$. While the length of this interval is larger than a value ϵ, we discretize I further with a step $a = \lfloor \frac{length(I)}{\eta} \rfloor$ and continue the search for a value of the delay that gives a lower cost by applying Algorithm 3. Finally, we repeat all the steps for different values of the discount and choose the one that results in minimal total costs. The final optimization procedure is given in Algorithm 4.

Observe that, by searching in a discretized set of $E(T)$, we assumed that the objective function is quasi-convex. However, this discretized set is in general larger than the set of delays used in the procedure described in Andersson & Melchiors (2001), where, for a fixed up-to-level at the warehouse, the delays considered are the results of a recursive procedure iterating over optimal values of S_i^* at retailers. The procedure in Andersson & Melchiors (2001) stops when the set of optimal S_i^* converges, assuming that the S_i^* are monotonic in $E(T)$. By guiding the search by the discretized values of $E(T)$, the procedure has less chance to be trapped in a local minima. We show in Section 5.2 that this optimization heuristic obtains close to optimal solutions in most of the cases. Moreover, due to the fact that discretization allows for decomposition of the problem per retailer, the running times are very fast, as it will be seen in Section 5.3.

Algorithm 3 Optimization procedure for the retailer network for a given discount and discretization of the warehouse delay.

1: $\mathcal{T}$: discretization of $E(T)$ over an interval I with step a (assume the elements of $\mathcal{T}$ are in increasing order.
2: $R_{prev} = 0$
3: Set $R_{\max}$ to the optimal R in an (R, Q) system with lead time L_w and $\lambda = \lambda_o + \sum_{i \in N} \lambda_i$
4: **for** $\hat{T} \in \mathcal{T}$ **do**
5: **for** $i \in N$ **do**
6: Use Algorithm 2 to find (S_i^*, IC_i^*) for which minimal costs at store i are achieved
7: **end for**
8: Calculate λ_w by (8)
9: Find $R \in [R_{prev}, R_{\max}]$ that gives the closest value to $\hat{T}$ when using (6)
10: Set $R_{prev} = R$
11: Calculate the cost at the warehouse for R:

$$C_w(\hat{T}, R, d) = h_w E(IL_w^+) + bE(IL_w^-) + c_t(w, s)\lambda_w$$

12: Calculate the network total costs for $(d, \hat{T})$:

$$TC = C_w(\hat{T}, R, d) + \sum_{i=1}^{N} C_i(\hat{T}, S_i^*, IC_i^*, d)$$

13: Find the delay $\hat{T}^*(d)$ that results in minimum total costs when the discount is d
14: **end for**

Algorithm 4 Improved Optimization procedure for an omnichannel network.

1: $\mathcal{D}$: set of discounts
2: Choose $\eta \in \mathbb{N}$
3: **for** $d \in \mathcal{D}$ **do**
4: Let $I = [0, L_w]$ and the discretization step $a = \lfloor \frac{L_w}{\eta} \rfloor$
5: **while** $a \geq \epsilon$ **do**
6: $\mathcal{T}$: discretization of I with step a
7: Apply Algorithm 4 to find the delay $\hat{T}^*(d) \in \mathcal{T}$ that gives minimal total costs for discount d
8: If $\hat{T}^*(d) = ka$, set $I = [\max\{0, (k-1)a\}, \min\{(k+1)a, L\}]$
9: $a = \frac{length(I)}{\eta}$
10: **end while**
11: **end for**
12: Output the combination $(d^*, \hat{T}^*(d^*))$ that results in minimal total costs.

5. Numerical experiments

In this section, we describe the numerical experiments we conducted. First, we analyze the quality of the approximation proposed in Section 3.1, for the case of one store and one warehouse. Then we analyze the optimization procedure proposed in Section 4 and comment on the insights gained from the numerical experiments.

5.1. Quality of the approximation for one store and one warehouse

To assess the quality of the approximation proposed in Section 3.1, we conducted experiments by varying the problem parameters as described in Table 1. In total, we ran 22,814 parameter combinations.

We measured the relative differences between simulation and approximation $\left(\frac{Sim - Approx}{Sim}\right)$ for four measures: Inventory on hand and expected number of lost customers at the stores, expected inventory on hand, and expected backlog at the warehouse. As the

Table 1
Experimental settings for testing the quality of the approximation.

Param.	Value	Param.	Value
λ_s	$\{0.5, 1, 2, 4, 8\}$	λ_o	$\{1, 4\}$
L_s	1	L_w	$\{1, 3, 6\}$
d	0.1	$p_a(d)$	$\{0.2, 0.7\}$
S	$\{1, \ldots, \min\{6, 2\lambda_s L_s\}\}$	IC	$\{0, \ldots, S\}$
R	$\{\max\{1, (\lambda_s + \lambda_o)L_w\}, \ldots, \min\{20, 2(\lambda_s + \lambda_o)L_w\}\}$	Q	$\{1, \ldots, \min\{10, R\}\}$

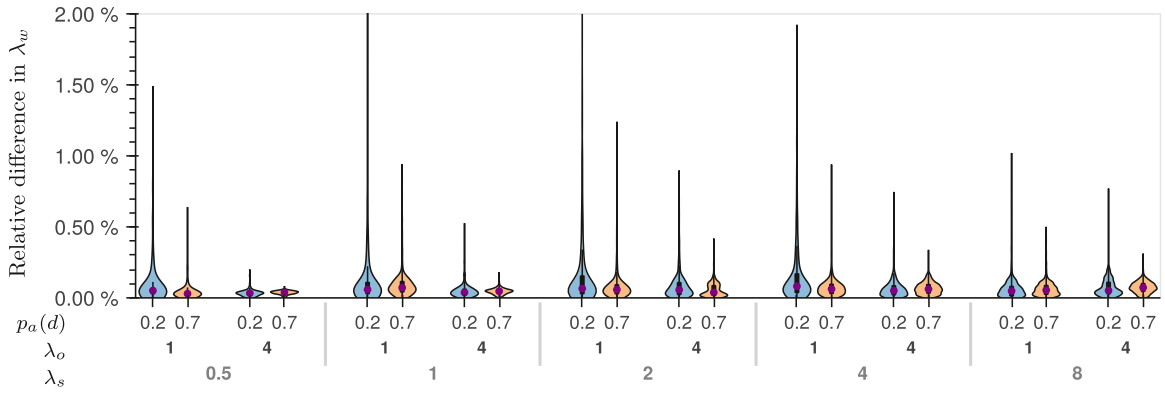

Fig. 2. Relative differences between simulation and approximation in the arrival rate λ_w at the warehouse.

total costs are linear in these quantities, errors in costs can be directly found from the errors of these quantities, for any combination of costs. The simulation experiments were limited to 50 replications for each combination, with 30,000 total demand events and 1000 warming up demand events in each replication. The summarized results are plotted using the violin plots that can show the average (the circle in the violin plot) and the full distribution (along the vertical axis) of the grouped data.

Fig. 2 presents the relative difference in the arrival rate at the warehouse between simulation and approximation $\left(\frac{\lambda_w^{sim} - \lambda_w^{app}}{\lambda_w^{sim}}\right)$. Each violin graph corresponds to a parameter setting mentioned below the horizontal axis. For example, the first violin graph in Fig. 2, describes the results for the experiments having ($p_a(d) = 0.2$, $\lambda_o = 1$, $\lambda_s = 0.5$). As the graphs show, most of the errors are below 0.2%, indicating that the arrival rate at the warehouse is well approximated. The relative average difference between simulation and approximation is 0.086%, and the standard deviation is 0.12%. The maximum relative difference is 1.98%, obtained for $\lambda_s = 2$, $\lambda_o = 1$, $p_a(d) = 0.2$, $L_s = 1$, $L_o = 6$, $S = 3$, $IC = 0$, $R = 18$, $Q = 1$.

The relative errors in inventory on hand at the store and at the warehouse for different parameter settings are presented in Figs. 3A and B. The graphs indicate that both the expected inventory on hand at the store and at the warehouse are well approximated for different combinations of parameters. The average relative difference for the inventory on hand at the store (see Fig. 3A) is 0.6%, and the standard deviation is 1.01%. The outliers in Fig. 3A correspond to low values of inventory on hand. The maximum relative difference is reached for a case where the simulated average inventory on hand is 2.02, while the approximated average inventory is 1.77. For the inventory on hand at the warehouse (see Fig. 3B), the average relative difference is 0.39%, and the standard deviation is 0.98%. The maximum relative difference is achieved for small inventory levels when the absolute difference between the simulated and approximated average inventory at the warehouse is 0.4.

The quality of the approximation for the expected number of lost sales is discussed in Fig. 4. This figure contains the relative differences between approximation and simulation (Fig. 4A) and the expected number of lost sales in the simulation (Fig. 4B) for different values of λ_o, λ_s, and S indicated in Table 1. In 92% of the cases, the relative difference is below 5%. The higher differences are caused by the fact that in many cases, losing customers is a rare event; hence it is difficult to capture it by simulation. To demonstrate this behavior, we refer to the expected number of lost sales in different cases (Fig. 4B). It can clearly be seen that the high relative differences (the outliers in Fig. 4A) appear only in the cases with higher stock levels (S) and lower expected numbers of lost sales (see Fig. 4B). In general, in the performed experiments, the expected numbers of lost customers are typically small (the average over all cases is 0.93), with the standard deviation of 1.22. As expected, the lost sales are decreasing in S and increasing in the arrival rates at the store. Also, they are higher for the lower acceptance rate of the discount ($p_a(d) = 0.2$).

A similar effect can be seen in the case of backorders at the warehouse (see Figs. 5 and 6). In the performed experiments, the expected backorder levels are rather small (in simulation, the average of all the cases is 0.068), with the standard deviation of 0.13. To understand better the behavior of backorders, we separated the cases with simulated average backorders larger than 1 (Fig. 5A and B) and the cases with simulated average backorders lower than 1 (Fig. 6A and B). For the cases with average backorders larger than 1, we report the relative difference in backorders between simulation and heuristic. For the cases with the simulated average number of backorders smaller than 1, we report the absolute difference, as for these cases, the relative difference can give a distorted image of the error.

For the cases where the simulated average number of backorders is larger than 1 (Fig. 5A and B), the average relative difference is 1.1%, with the standard deviation of 0.8%. As expected, Fig. 5B indicates that the total number of backorders is higher when the probability of accepting the discount is higher ($p_a(d) = 0.7$), and the arrival rate at the warehouse is higher.

For the cases where the average number of backorders is below 1.0, (Fig. 6A and B), the absolute differences are below 0.19, with

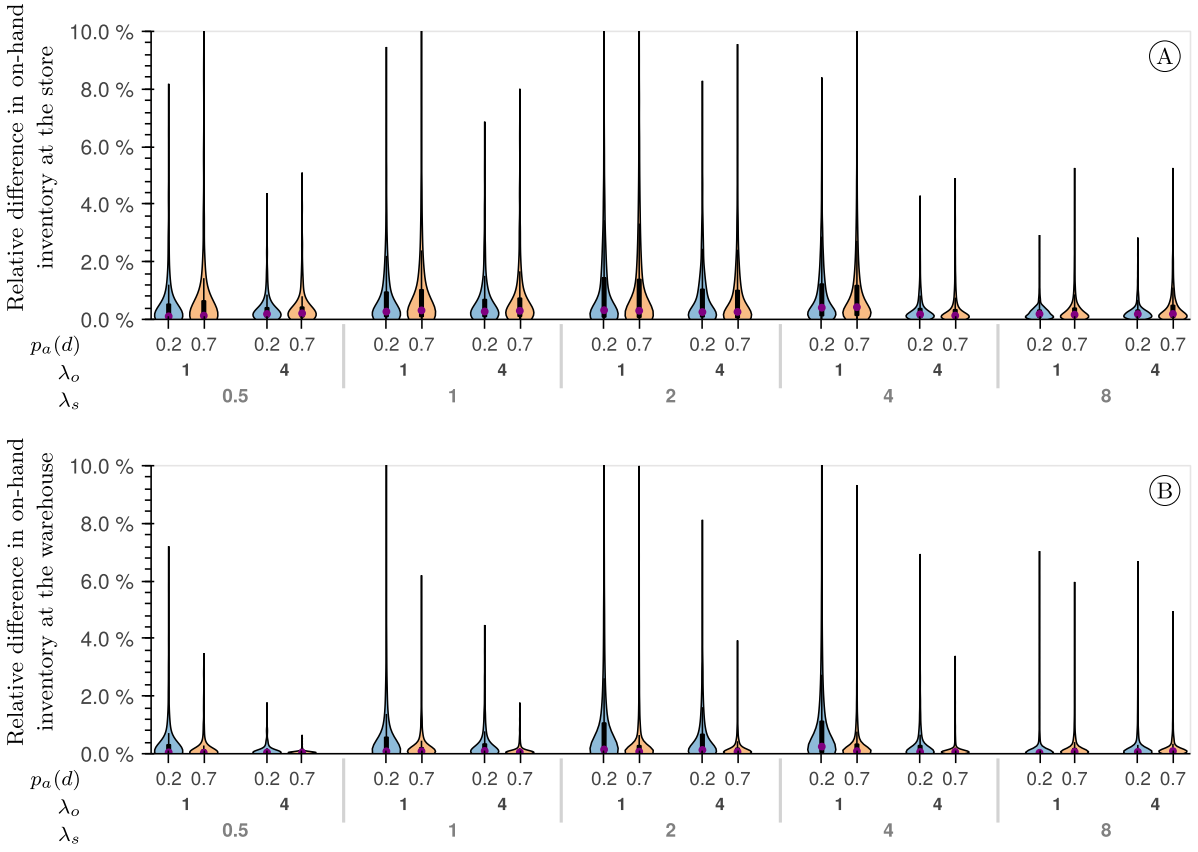

Fig. 3. Relative differences between simulation and approximation in on hand inventories at the store (A) and the warehouse (B).

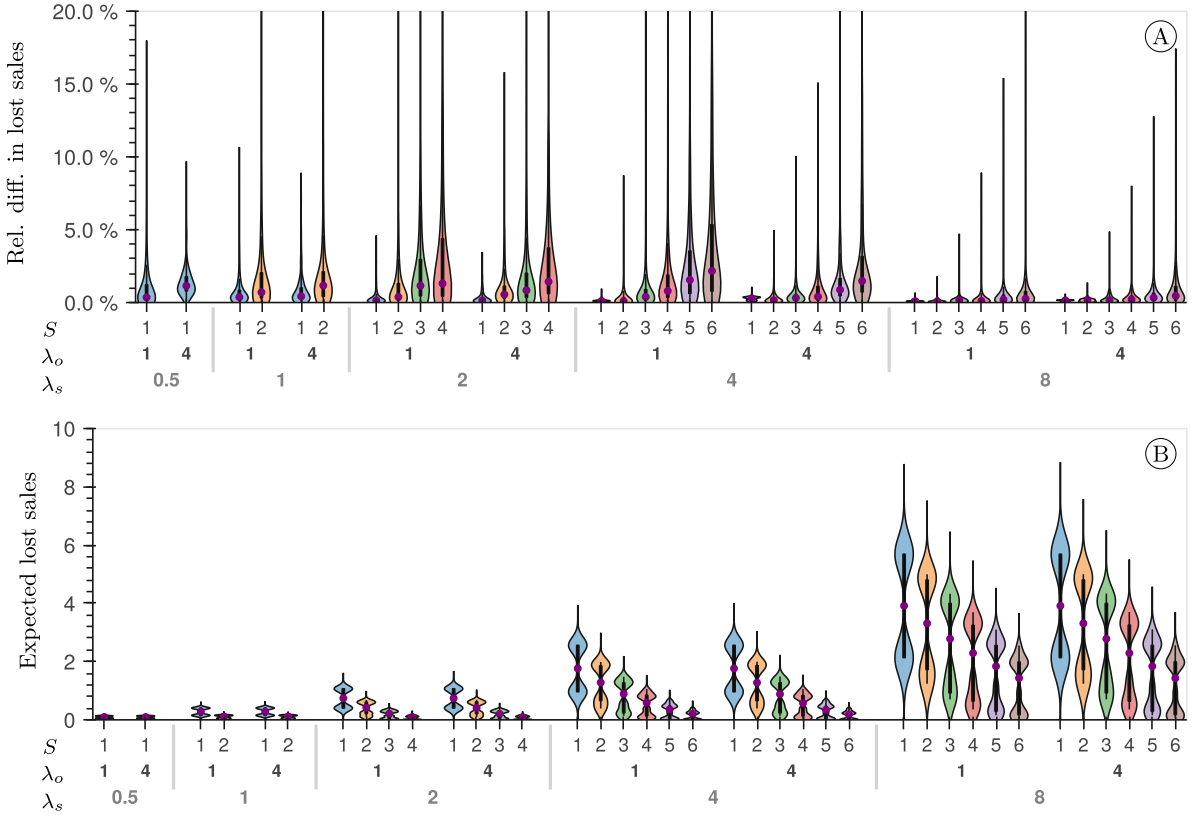

Fig. 4. Relative differences between simulation and approximation in lost sales at the store (A), and the expected simulated lost sales (B).

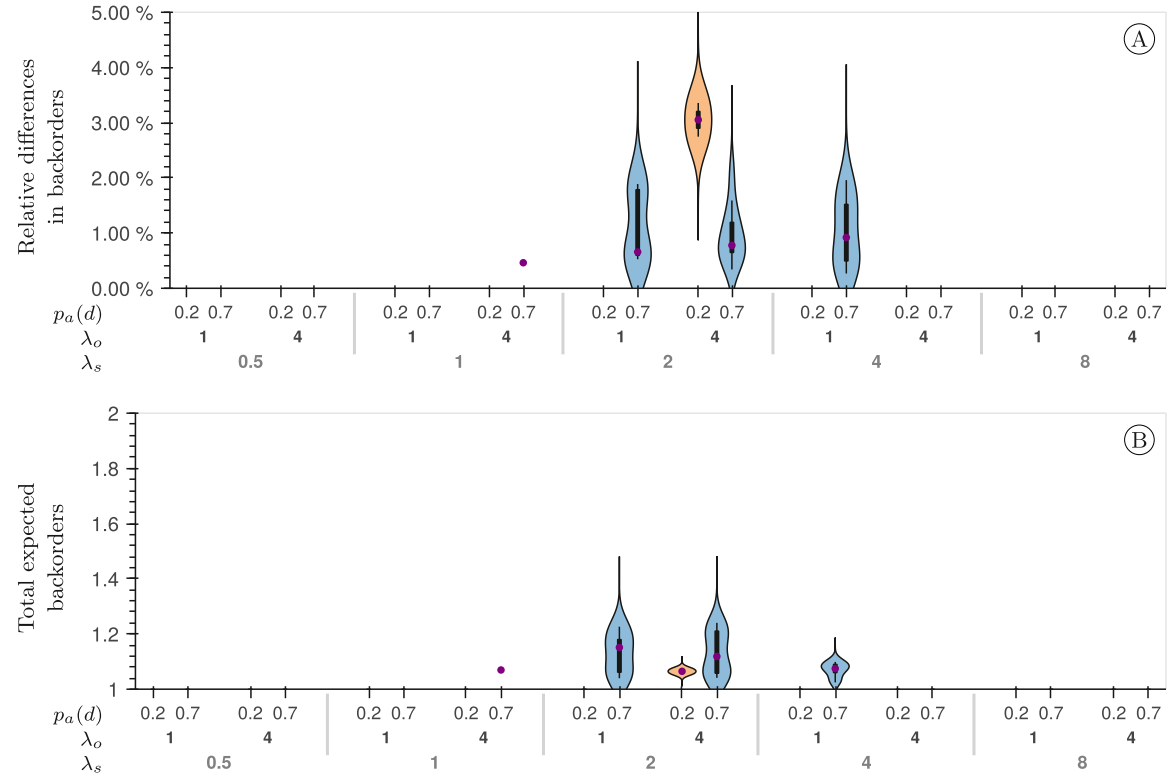

Fig. 5. Relative differences and the total backorders at the warehouse, when the expected backorders are greater than 1.

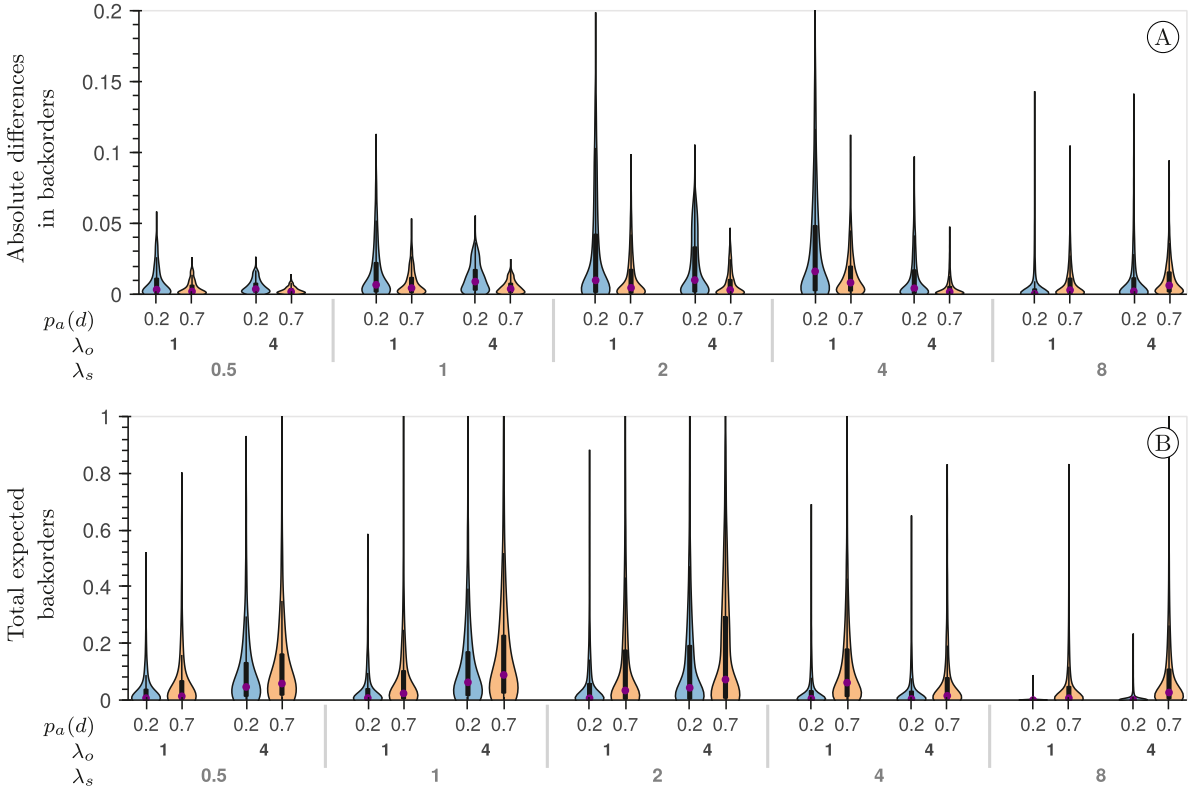

Fig. 6. Absolute differences and the total backorders at the warehouse, when the expected backorders are less than 1.

Table 2
Experimental settings for testing the quality of the optimization heuristic.

Param.	Value	Param.	Value
λ_s	$\{0.5, 1, 2\}$	λ_o	1
L_s	1	L_w	$\{1, 3, 6\}$
h_s	$h_w \times \{1, 1.5, 2\}$	h_w	10
l_s	$p \times \{0.5, 1, 2\}$	b	$h_w \times \{2, 10\}$
p	$h_s \times \{10, 20\}$	N	$\{1, 2\}$

Table 3
Experimental settings for testing the total cost reduction created by discounts.

Param.	Value	Param.	Value
λ_s	$\{0.5, 1, 2\}$	λ_o	$\{0.5, 1, 2\}$
L_s	1	L_w	$\{1, 3, 6\}$
h_s	$h_w \times \{1, 1.5, 2\}$	h_w	10
l_s	$p \times \{0.5, 1, 2\}$	b	$h_w \times \{2, 10\}$
p	$h_s \times \{10, 20\}$	N	$\{1, 4, 10\}$

the average absolute difference of 0.013 and the standard deviation of 0.021. As for a higher number of backorders, the expected backorders are better approximated in the simulation, and the errors are lower for $p_a(d) = 0.7$.

5.2. Quality of the optimization heuristic for more stores

To assess the quality of the optimization heuristic described in Algorithms 2–4, we compared the solutions provided by the optimization heuristic with the optimal solutions found by brute-force enumeration of the optimized variables (S, R, IC, d) for a set of combinations of the problem parameters. As the costs are essential in discussing the impact of discounts, we added the holding, lost sales, and backlog costs to the parameters that are being varied. Due to the computational time required by the brute force calculations used to benchmark the proposed heuristic, we had to restrict the number of the tested values for each parameter. The values used for each parameter are described in Table 2. Note that keeping the parameters λ_s, L_s, and h_w fixed allows us to study essentially different situations, corresponding to different values of the ratios $\frac{\lambda_s}{\lambda_o}$, $\frac{L_w}{L_s}$, and $\frac{h_s}{h_w}$. In total, we ran 432 different parameter combinations.

In all these combinations, the order quantity Q and the transportation costs ($c_t(w)$ and $c_t(w, s)$) are fixed to 10. The discounts take values in $\mathcal{D} = \{0.0, 0.05p, 0.1p, 0.15p, 0.2p, 0.25p\}$, with the probability of accepting a discount being equal to $p_a(d) = 3d$, $d \in \mathcal{D}$. The stores have identical characteristics and the number of stores is $N \in \{1, 2\}$. For each set, we found the optimal reorder levels S_i, the critical levels IC_i $(i = 1, \ldots, N)$, the reorder point at the warehouse R, and the optimal discount d. In the optimization procedure, the discretization parameter for the delay at the warehouse is $\eta = 10$.

In most of the studied cases (91% of the single-store cases and 90% of the two-store cases), the total costs obtained via the optimization procedure were almost equal (with $\leq 0.1\%$ difference) to the total costs obtained by brute-force enumeration (see Fig. 7A and B). The maximum difference between the approximated total costs and the optimal costs was 2.23% for $N = 1$ and 2.27% for $N = 2$, both are reached for small loss penalties ($l_s = 0.5p$).

5.3. Running time of the optimization heuristic

To study the running times of the optimization heuristic proposed in Algorithms 2–4, we ran all the cases described in Table 2, with $N \in \{10, 40, 70, 100\}$. All the experiments are performed using Python 3.7 implementation on Intel® Xeon® X5650 @ 2.67GHz processors. The results regarding the running times are summarized using violin plots in Fig. 8. The highest running times are for $N = 100$, with the average of 833.5 s, the standard deviation of 736.6 s and the maximum of 3958 s. This is a reasonable time for this application, regarding that the inventory parameters are calculated once for a longer period of time. The factors that have the most impact on the running time are the factors that affect the delay at the warehouse $E(T)$ (the lead time at the warehouse and the arrival rates at the stores) and the number of retailers N. Observe that the delay at the warehouse impacts the number of the discretized values for which an optimization per retailer is solved, while the number of retailers impacts the number of optimizations solved. The presented results indicate that the running times are almost linear in the number of stores (N) and the arrival rates at the stores (λ_s). The lead time at the warehouse seems to have the largest impact on the running time, especially for a larger number of stores N. Note that in case shorter running times are desirable, the optimization procedure we propose can be easily parallelized due to the fact that after discretizing $E(T)$, the optimizations per retailer become independent.

5.4. Impact of offering discounts

To evaluate the importance of offering discounts and its sensitivity to the input parameters, we conducted a number of optimization experiments (using Algorithm 4) by varying the problem parameters as described in Table 3. In total, we ran 1944 parameter combinations. In all these combinations, the transportation costs ($c_t(w)$ and $c_t(w, s)$) are fixed to 10. The discounts take values in $\mathcal{D} = \{k \times p | k \in \{0, 0.05, 0.10, 0.15, 0.20, 0.25\}$, where p as indicated in Table 3. The probability of discount acceptance was set to be equal to three times the level of the discount ($p_a(d) = 3d$). The stores have identical characteristics and their number $N \in \{1, 4, 10\}$.

Fig. 9 illustrates the total cost reduction compared to the situation with no discounts for different combinations of N and λ_s, λ_o (Fig. 8A), N, λ_s, h_s/h_w (Fig. 8B) and N, p/h_w and l_s/p (Fig. 8C). Over all cases, when using discounts, the total cost is reduced by 8.5% on average, with the maximum of 19.8% (obtained for $N = 10$ $\lambda_s = \lambda_o = 0.5$, $L_s = L_w = 1$, $h_s = 20$, $h_w = 10$, $l_s = 200$, $b = 20$, $p = 200$). This reduction can be explained by the fact that discounts allow lower stock at the stores while reducing the risk of losing customers by redirecting them to the warehouse.

All graphs in Fig. 9 show that the higher the number of stores, the higher the average benefit of discounts. For $N = 1$, the average improvement is 6%, while for $N = 10$ the average improvement is 10.3%. This indicates the important role of pooling inventory at the warehouse, as a result of offering discounts. Figures 9A and 9B indicate that the benefits decrease as λ_s increases, as more items have to be kept in stock at the store. This effect is especially visible for a higher number of stores, $N = 4$ and $N = 10$.

Fig. 9A shows that the cost reductions are higher for smaller values of the online and store demand rates when discounts result in higher relative stock reduction at the retailers. For a fixed value of λ_s, and N, the benefits of offering discounts remain relatively constant when the values of λ_o are changed. This is expected, as the discounts are only offered to the customers arriving at the store.

Furthermore, Fig. 9B illustrates that the higher the store holding costs, compared to the warehouse, the more beneficial is to offer discounts. For example, when $h_s = h_w$, the average benefit is 6.7%, while for $h_s = 2h_w$, the average benefit is 10% over all the cases. For a large number of stores ($N = 10$), discounts still result in

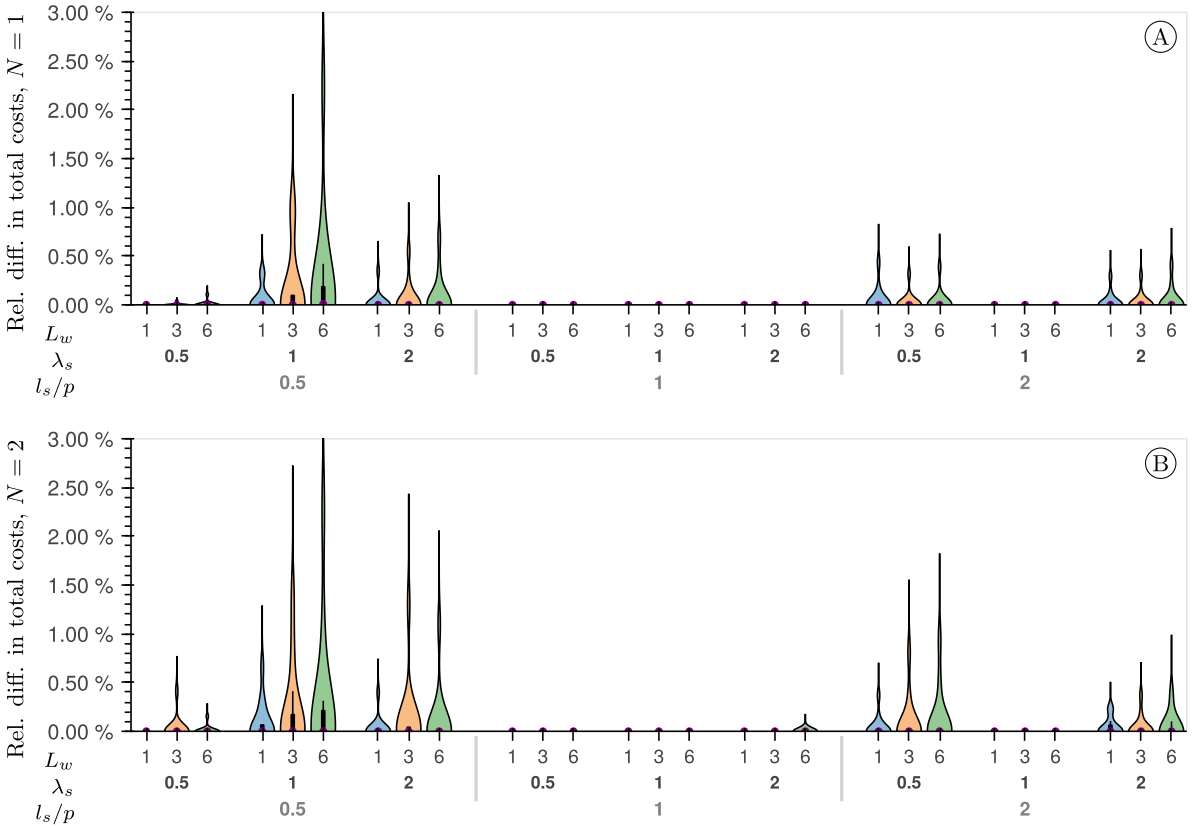

Fig. 7. Relative differences in optimal total costs (obtained by brute force) and the optimized total costs, for $N = 1$ (A) and $N = 2$ (B).

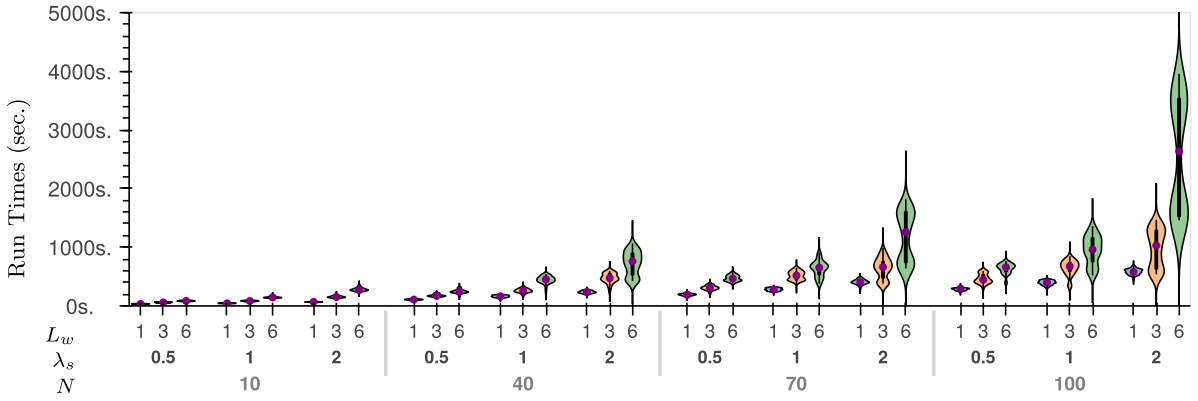

Fig. 8. Running times of the optimization heuristic.

large savings due to the pooling effect at the warehouse. The two modes of the distributions in 9B are due to the additional impact of l_s/p.

Fig. 9C shows that the savings are increasing in l_s/p, as losing customers is expensive and thus is beneficial to prevent it through discounts. When $l_s = 0.5p$, the average savings in our experiments are on average 4%. However, when $l_s = 2p$, the savings increase to 11%, over all the cases, and to 10% for $N = 10$. The figure also confirms that discounts are more beneficial for higher holding costs at the store ($p = 10h_s$); however, the impact of p/h_s does not seem to be significant. For $p = 10h_s$, the average improvement is 8.7%, and for $p = 20h_s$, the average improvement is 8.2%.

Based on the analysis above, we conclude that the main factors that impact the profitability of the discount policy are the number of stores N, λ_s, h_s/h_w, and l_s/p. The policy is especially beneficial for a larger number of stores N, low arrival rates λ_s, high holding costs at the retailers (compared to the warehouse), and high lost sales costs.

The analysis of the optimal discounts (Fig. 10) indicates that the optimal discounts (that can be offered) increase in l_s/p. In our experiments, when $l_s/p \in \{1, 2\}$, the optimal discounts were on average 25%. As expected, the more expensive is too loose customers, the higher is the discount that can be offered in order to convince the customer to switch to the online channel.

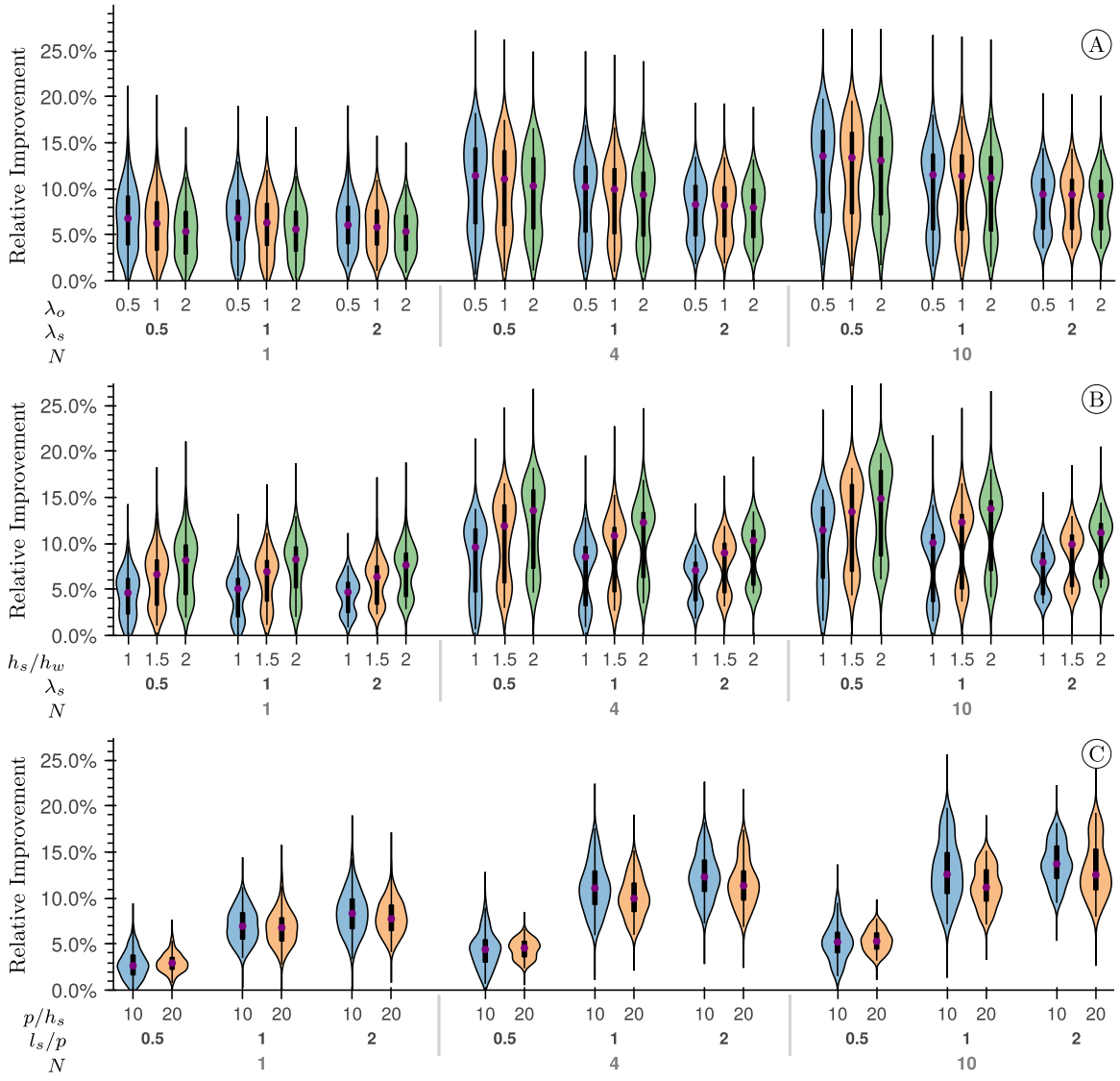

Fig. 9. Total Cost improvements when discounts are used.

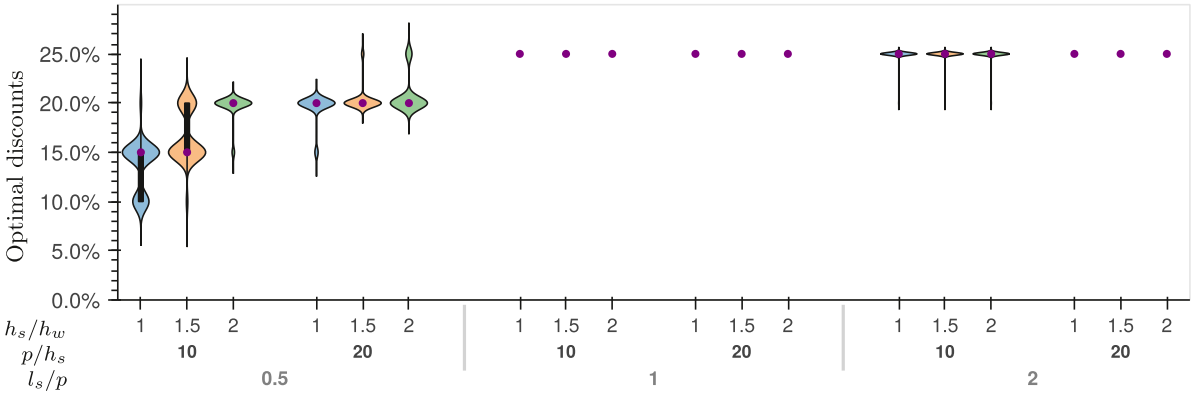

Fig. 10. Optimal discounts.

Finally, we have analyzed the behavior of the rationing levels *IC*, see Fig. 11. Observe that the critical levels become essential when the lost sales costs are high. In our experiments, 12% of the cases with $l_s/p = 2$ (118 out of 972) had non-zero *IC* levels. In an omnichannel setting, high lost sales can occur when sold items lead to sales-related subscriptions or sales of other related products.

Remark 2. Observe that we assumed that the transportation costs for the customers who switch to the online channel are part of the discount. This means that the retailer should run the optimization procedure only with levels of discount that cover the transportation costs in order to judge in which situations a discount is profitable.

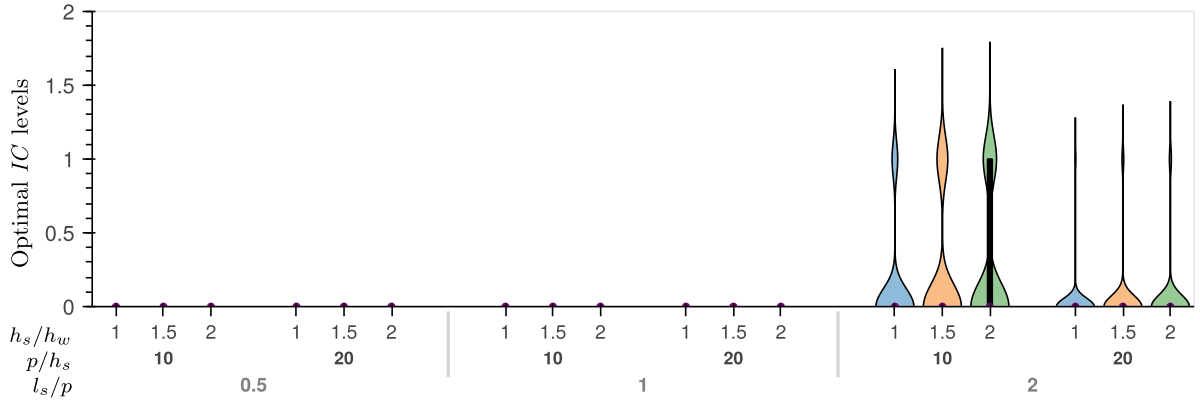

Fig. 11. Optimal *IC* levels.

6. Conclusions and discussion

In this paper, we studied the impact of offering financial incentives to customers for switching the purchasing channel from brick-and-mortar stores to online. We assumed a base-stock policy with critical level and lost sales at stores and an (R, Q) policy with backlog at the warehouse. The warehouse fulfills the online orders and replenishes the stores. We assumed Poisson distributed online and in-store demand. For the case of a single store, we proposed a recursive approximation to estimate the delay at the warehouse. We prove that the procedure converges and show via extensive experiments that the errors obtained are very low. For the case of more stores, we propose an optimization procedure based on the assumption that optimal base stock levels at stores are decreasing in the discount offered.

The results in this paper indicate that retailers of expensive, slow-moving items can benefit substantially by redirecting the customers that are willing to wait to the online channel. In our experiments, by adopting this policy, the average cost reduction was 8.5%, with the maximum of 19.5%. Discounts are especially beneficial if the arrival rates at the store are low, as in this case, an inventory reduction of a few items represents a considerable reduction in inventory cost relative to the total costs. The same beneficial effects occur when holding costs in stores are high compared to the holding cost at the warehouse. The policy results in high cost reductions also when the number of stores is high, as in this case, there is a pooling effect of inventory at the warehouse. As expected, it is also profitable to offer discounts when the lost sales costs are high. In this case, discounts reduce the number of lost customers, as some of them accept to be served by the warehouse. The larger the lost sales costs, the higher the critical level at which discounts should be offered.

An interesting venue for future research would be to study in more detail the properties of the optimal solution of this model. The quasi-convexity of the costs at the stores, although noticed in all the experiments, remained unproven. Similarly, it would be interesting to prove monotonicity results regarding the optimal stock levels, similar to the papers of Song et al. (2010) and Federgruen & Wang (2013) for the standard one-echelon models.

As it assumes base-stock policies at the stores, the model discussed in this paper is appropriate for slow-moving items. It would be interesting to develop models to study pricing (discount) policies for faster moving items, where other inventory policies are more appropriate.

Acknowledgments

This publication is based upon work supported by the Khalifa University of Science and Technology under Award No. RC2 DSO and Grant Number FSU 2019-11. We would like to thank Dr. Youssef Boulaksil and Tjebbe Bodewes for discussions on a related model and for preliminary simulations.

References

Acimovic, J., & Graves, S. C. (2015). Making better fulfillment decisions on the fly in an online retail environment. *Manufacturing & Service Operations Management, 17*(1), 34–51. https://doi.org/10.1287/msom.2014.0505.

Acimovic, J., & Graves, S. C. (2017). Mitigating spillover in online retailing via replenishment. *Manufacturing & Service Operations Management, 19*(3), 419–436. https://doi.org/10.1287/msom.2016.0614.

Andersson, J., & Melchiors, P. (2001). A two-echelon inventory model with lost sales. *International Journal of Production Economics, 69*(3), 307–315.

Arslan, H., Graves, S. C., & Roemer, T. A. (2007). A single-product inventory model for multiple demand classes. *Management Science, 53*(9), 1486–1500. https://doi.org/10.1287/mnsc.1070.0701.

Axsäter, S. (1990). Simple solution procedures for a class of two-echelon inventory problems. *Operations Research, 38*(1), 64–69. https://doi.org/10.1287/opre.38.1.64.

Axsäter, S. (1993). Exact and approximate evaluation of batch-ordering policies for two-level inventory systems. *Operations Research, 41*(4), 777–785. https://doi.org/10.1287/opre.41.4.777.

Axsäter, S. (1998). Evaluation of installation stock based (r, q)-policies for two-level inventory systems with poisson demand. *Operations Research, 46*(3-supplement-3), S135–S145.

Axsäter, S. (2000). Exact analysis of continuous review (r, q) policies in two-echelon inventory systems with compound poisson demand. *Operations research, 48*(5), 686–696.

Axsäter, S. (2015). *Inventory control*. Springer International Publishing. https://doi.org/10.1007/978-3-319-15729-0.

Bell, D. R., Gallino, S., & Moreno, A. (2018). Offline showrooms in omnichannel retail: Demand and operational benefits. *Management Science, 64*(4), 1629–1651. https://doi.org/10.1287/mnsc.2016.2684.

Bell, D. R., Gallino, S., & Moreno, A. (2020). Customer supercharging in experience-centric channels. *Management Science, 66*(9), 4096–4107. https://doi.org/10.1287/mnsc.2019.3453.

Bendoly, E., Blocher, D., Bretthauer, K. M., & Venkataramanan, M. (2007). Service and cost benefits through clicks-and-mortar integration: Implications for the centralization/decentralization debate. *European Journal of Operational Research, 180*(1), 426–442. https://doi.org/10.1016/j.ejor.2006.03.043.

Bhargava, H. K., Sun, D., & Xu, S. H. (2006). Stockout compensation: Joint inventory and price optimization in electronic retailing. *INFORMS Journal on Computing, 18*(2), 255–266.

Brumelle, S. L. (1978). A generalization of Erlang's loss system to state dependent arrival and service rates. *Mathematics of Operations Research, 3*(1), 10–16.

Cheung, K. L. (1998). A continuous review inventory model with a time discount. *IIE transactions, 30*(8), 747–757.

DeCroix, G. A., & Arreola-Risa, A. (1998). On offering economic incentives to backorder. *IIE transactions, 30*(8), 715–721.

Deshpande, V., Cohen, M. A., & Donohue, K. (2003). A threshold inventory rationing policy for service-differentiated demand classes. *Management Science, 49*(6), 683–703. https://doi.org/10.1287/mnsc.49.6.683.16022.

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

Fadloğlu, M. M., & Bulut, Ö. (2010). An embedded Markov chain approach to stock rationing. *Operations Research Letters, 38*(6), 510– 515. https://doi.org/10.1016/j.orl.2010.08.004.

Federgruen, A., & Wang, M. (2013). Monotonicity properties of a class of stochastic inventory systems. *Annals of Operations Research, 208*(1), 155–186.

Gabor, A. F., van Vianen, L. A., Yang, G., & Axsäter, S. (2018). A base-stock inventory model with service differentiation and response time guarantees. *European Journal of Operational Research, 269*(3), 900–908.

Govindarajan, A., Sinha, A., & Uichanco, J. (2018). Joint inventory and fulfillment decisions for omnichannel retail networks. *Naval Research Logistics (NRL)*.

Ha, A. Y. (1997). Inventory rationing in a make-to-stock production system with several demand classes and lost sales. *Management Science, 43*(8), 1093–1103. https://doi.org/10.1287/mnsc.43.8.1093.

Harsha, P., Subramanian, S., & Uichanco, J. (2019). Dynamic pricing of omnichannel inventories. *Manufacturing & Service Operations Management, 21*(1), 47–65. https://doi.org/10.1287/msom.2018.0737.

Hill, R., Seifbarghy, M., & Smith, D. (2007). A two-echelon inventory model with lost sales. *European Journal of Operational Research, 181*(2), 753–766. https://doi.org/10.1016/j.ejor.2006.08.017.

van Houtum, G.-J. (2006). Multiechelon production/inventory systems: Optimal policies, heuristics, and algorithms. *Models, Methods, and Applications for Innovative Decision Making*, 163–199. https://doi.org/10.1287/educ.1063.0026.

Hübner, A., Holzapfel, A., & Kuhn, H. (2016). Distribution systems in omni-channel retailing. *Business Research, 9*(2), 255–296.

Jasin, S., & Sinha, A. (2015). An lp-based correlated rounding scheme for multi-item ecommerce order fulfillment. *Operations Research, 63*(6), 1336–1351.

Kleijn, M. J., & Dekker, R. (1999). An overview of inventory systems with several demand classes. *New Trends in Distribution Logistics*, 253–265. https://doi.org/10.1007/978-3-642-58568-5_13.

de Kok, T., Grob, C., Laumanns, M., Minner, S., Rambau, J., & Schade, K. (2018). A typology and literature review on stochastic multi-echelon inventory models. *European Journal of Operational Research, 269*(3), 955– 983. https://doi.org/10.1016/j.ejor.2018.02.047.

Lei, Y. M., Jasin, S., & Sinha, A. (2018). Joint dynamic pricing and order fulfillment for e-commerce retailers. *Manufacturing & Service Operations Management, 20*(2), 269–284. https://doi.org/10.1287/msom.2017.0641.

Mahar, S., Bretthauer, K. M., & Venkataramanan, M. (2009). The value of virtual pooling in dual sales channel supply chains. *European Journal of Operational Research, 192*(2), 561–575. https://doi.org/10.1016/j.ejor.2007.09.034.

Melchiors, P., Dekker, R., & Kleijn, M. J. (2000). Inventory rationing in an (s,q) inventory model with lost sales and two demand classes. *Journal of the Operational Research Society, 51*(1), 111–122. https://doi.org/10.1057/palgrave.jors.2600844.

Messerli, E. J. (1972). Proof of a convexity property of the Erlang b formula. *The Bell System Technical Journal, 51*(4), 951–953. https://doi.org/10.1002/j.1538-7305.1972.tb01956.x.

Nahmias, S., & Demmy, W. S. (1981). Operating characteristics of an inventory system with rationing. *Management Science, 27*(11), 1236–1245. https://doi.org/10.1287/mnsc.27.11.1236.

Nahmias, S., & Smith, S. A. (1994). Optimizing inventory levels in a two-echelon retailer system with partial lost sales. *Management Science, 40*(5), 582–596. https://doi.org/10.1287/mnsc.40.5.582.

Shaked, M., & Shanthikumar, J. G. (2007). *Stochastic orders*. Springer Science & Business Media.

Sherbrooke, C. C. (1968). Metric: A multi-echelon technique for recoverable item control. *Operations Research, 16*(1), 122–141. https://doi.org/10.1287/opre.16.1.122.

Simchi-Levi, D., & Zhao, Y. (2011). Performance evaluation of stochastic multi-echelon inventory systems: A survey. *Advances in Operations Research, 2012*.

Song, J.-S., Zhang, H., Hou, Y., & Wang, M. (2010). The effect of lead time and demand uncertainties in (r, q) inventory systems. *Operations Research, 58*(1), 68–80.

Tijms, H. C. (2003). *A first course in stochastic models*. John Wiley and sons.

Vicil, O., & Jackson, P. (2016). Computationally efficient optimization of stock pooling and allocation levels for two-demand-classes under general lead time distributions. *IIE Transactions, 48*(10), 955–974. https://doi.org/10.1080/0740817x.2016.1146421.