Inventory Optimization for Fulfillment Integration in Omnichannel Retailing

Aravind Govindarajan
Stephen M. Ross School of Business
University of Michigan

Amitabh Sinha
Stephen M. Ross School of Business
University of Michigan

Joline Uichanco
Stephen M. Ross School of Business
University of Michigan

Ross School of Business Working Paper
Working Paper No. 1341
April 2017

Inventory Optimization for Fulfillment Integration in Omnichannel Retailing

Aravind Govindarajan, Amitabh Sinha and Joline Uichanco
Stephen M. Ross School of Business, University of Michigan, Ann Arbor, MI 48109
{arav,amitabh,jolineu}@umich.edu

With e-commerce growing at a rapid pace compared to traditional retail, many brick-and-mortar firms are supporting their online growth through an omnichannel approach, which integrates inventories across multiple channels. We analyze the inventory optimization of three such omnichannel fulfillment systems for a retailer facing two demand streams (online and in-store). The systems differ in the level of fulfillment integration, ranging from no integration (separate fulfillment center for online orders), to partial integration (online orders fulfilled from nearest stores) and full integration (online orders fulfilled from nearest stores, but in case of stockouts, can be fulfilled from any store). We obtain optimal order-up-to quantities for the analytical models in the two-store, single-period setting. We then extend the models to a generalized multi-store setting, which includes a network of traditional brick-and-mortar stores, omnichannel stores and online fulfillment centers. We develop a simple heuristic for the fully-integrated model, which is near optimal in an asymptotic sense for a large number of omnichannel stores, with a constant approximation factor dependent on cost parameters. We augment our analytical results with a realistic numerical study for networks embedded in the mainland US, and find that our heuristic provides significant benefits compared to policies used in practice. Our heuristic achieves reduced cost, increased efficiency and reduced inventory imbalance, all of which alleviate common problems facing omnichannel retailing firms. Finally, for the multiperiod setting under lost sales, we show that a base-stock policy is optimal for the fully-integrated model.

Key words: omnichannel; e-commerce; inventory management; pooling; heuristic; asymptotic analysis

1. Introduction

In 2015, e-commerce sales accounted for 7% of the total retail sales in the United States (U.S. Census Bureau 2016). Although this is a small portion of the total sales, online sales have been increasing at a rapid growth rate of around 15% each year, and constituted 60% of the total retail sales growth in 2015 (Zaroban 2016). With customers increasingly favoring the online channel, traditional brick-and-mortar (B&M) firms are compelled to develop their e-commerce capabilities to remain competitive against pure play e-commerce firms like Amazon (Leiser 2016), which alone accounted for 24% of the total retail sales

growth in 2015 (Enright 2015). In order to improve efficiency and flexibility, retailers resort to an omnichannel approach to integrate the online channel with their physical stores.

Omnichannel refers to the seamless integration of a retailer's sales channels, which may comprise of B&M stores, online stores (websites, mobile apps), etc. Customers can purchase an item through different channels, including placing an order through the online store, through mobile devices, as well as through the traditional practice of walking into physical stores. In addition, customers placing orders online can also choose how they receive the item: they can pick up their items from a nearby physical store or from designated places (e.g. Amazon Lockers), or simply have the item shipped directly to their homes. Firms provide such flexibility to the customers in the form of omnichannel initiatives such as pickup in-store, ship-to-store, curbside pickup, ship-to-customer, etc.

Providing an omnichannel customer experience is regarded as a brand differentiator by many retailers, and customers value the convenience offered by omnichannel retailers (Forrester 2014). In addition, integrating the online channel with the physical stores increases revenue, reduces shipping costs and improves customer satisfaction (Forrester 2014). Hence it is not surprising that both pure play e-commerce firms as well as traditional B&M firms are switching to omnichannel to reap its benefits. For instance, pure play e-commerce firms like Amazon and Warby Parker are setting up physical stores to reduce shipping costs and improve shopping experiences (Bensinger and Morris 2014), while onetime B&M firms like Macy's and Walmart have integrated their online channels with their physical stores to leverage their existing network of retail stores (Nash 2015).

One of the key aspects of this channel integration is the use of physical stores to fulfill online orders. *Store fulfillment* has now become indispensable for firms like Walmart, which have a network of physical stores close to population centers (Barr 2013), and Macy's, that rely on their physical stores to offer same day and next-day delivery options to customers (Giannopoulos 2014). Store fulfillment for online orders can be done in several ways: ship-to-customer (where items are shipped directly to the customer), pickup in-store (where customers can pick up their items from nearby stores), curbside pickup, etc. In our study, we focus on the ship-to-customer option of online fulfillment, which has several advantages over the use of online fulfillment centers, including reduced shipping costs, quicker deliveries and efficient use of store inventory (UPS Compass 2014). In this method of fulfillment,

items are picked off the shelf, packed, labeled and shipped to the customers' homes, and this requires dedicated floor space and store staff to handle these activities.

In spite of its numerous advantages, such channel integration comes at a price: from 2010 to 2014, even as retail and online sales increased, inventory turnover decreased (Kurt Salmon 2016). This highlights the opportunity in efficient inventory management due to complex omnichannel practices. Additionally, the increase in competition in the e-commerce channel can lead to firms holding excess inventory. As opposed to traditional B&M sales, online orders have an additional supply flexibility: if the store nearest to an order does not have the item in stock, the order can be fulfilled from another store which has the item. If managed well, this can lead to increased turnover by redistributing inventory and avoiding markdowns. In this study, we analyze how this virtual pooling of online demand can be exploited to achieve better inventory efficiency.

An integrated approach can involve significant investments, as a unified view of inventory across all stores and fulfillment centers is mandatory for implementing store fulfillment. As a result, different firms resort to varied methods of integration of online demand to their physical stores, based on their financial capabilities and business priorities. Our research focuses on inventory optimization in omnichannel retailing, for varying levels of this integration of the online channel to physical stores. We consider three basic systems involving a firm which has the ship-to-customer online channel in addition to its physical stores:

1. *No Integration (NI) system*, where a separate online fulfillment center is used to fulfill all online orders, and in-store demand is fulfilled from physical stores,

2. *Partially-integrated (PI) system*, where online orders are allocated to the nearest stores, and both in-store and online demands are fulfilled using the store inventory, and

3. *Fully-integrated (FI) system*, where online orders are allocated to the nearest stores, but in case of stockout, they can be rerouted to any store which has the item in stock.

In a way, these systems can be viewed as different stages in the omnichannel evolution. Our study is versatile in application, providing implications for firms which already have established omnichannel practices, as well as those venturing into omnichannel retailing. Our major contributions to academic literature and industry practice are as follows:

- First, we develop analytical models for the three systems mentioned above, and solve them to obtain *optimal order-up-to quantities* for the single-period, two-store problem.

• Second, we consider a generalized multi-store setting, and observe that the NI and PI models extend trivially. We develop a simple heuristic and lower bound for the computationally intractable *multi-store FI model* and show that under certain conditions, our heuristic solution is near optimal with a constant approximation factor dependent on cost parameters, for an asymptotic setting with large number of omnichannel stores.

• Third, we augment our analytical results with a realistic numerical study for the multi-store setting, and show that our heuristic achieves significant cost savings around 30% over policies used in practice, while reducing inventory imbalance and improving efficiency. We also study the two-store case, and find that an omnichannel approach is preferred over a centralized approach when there is a moderate mix of online and in-store demands, and planning for virtual pooling is valuable when overage and underage costs are unbalanced.

• Finally, we prove the optimality of a stationary base-stock policy for the multi-period setting under lost sales for the fully-integrated model.

We review the related literature in Section 2. In Section 3, we analyze the two-store problem, deriving analytic solutions to the proposed models. In Section 4, we extend our analysis to a generalized multi-store setting, developing a simple heuristic which is near optimal in an asymptotic sense. In Section 5, we perform numerical analysis to compare the models in the two-store case, as well as demonstrate the performance of the heuristic in a realistic multi-store setting. We discuss extensions to the multi-period setting in Section 6, and conclude by summarizing our findings and contributions, and discussing future research directions in Section 7. The appendices are relegated to the online supplement.

2. Literature Review

Omnichannel retailing is a relatively new area in literature, and has been gaining traction in recent years. Readers are referred to Rigby (2011) and Brynjolfsson et al. (2013) for comprehensive reviews of the topic. Recent studies have focused on the consequences of channel integration, such as customer migration due to product information (Bell, Gallino, and Moreno 2013, Ansari, Mela, and Neslin 2008, Gallino and Moreno 2014) and sales dispersion (Gallino, Moreno, and Stamatopoulos 2016) among others. Our study focuses on optimizing inventory in an omnichannel setting with integrated channels, and is thus closely related to multi-channel distribution studies.

Earlier studies on channel integration such as Chiang and Monahan (2005), Netessine and Rudi (2004) and Netessine and Rudi (2006) have focused on analyzing fulfillment

structures in dual-channel supply chains, where the supplier can drop-ship to fulfill online demand. Alptekinoğlu and Tang (2005) develop near-optimal distribution strategies for firms distributing products using multiple channels. Interested readers are referred to the review by Agatz et al. (2008) on multi-channel distribution and e-fulfillment.

As the retail industry progressed, firms gradually integrated online demand to their physical stores. The reflection in literature is seen in the form of various studies which addressed integrating inventory across channels, some of which are as follows. Seifert et al. (2006) study the integration of a virtual store into an existing supply chain, where excess inventory from stores can be transshipped to the virtual store. Bretthauer et al. (2010) consider the problem of selecting physical stores to fulfill online orders. Mahar et al. (2009) develop dynamic assignment policies to allocate online orders to physical stores. The current retail industry has since evolved to admit online fulfillment in most physical stores, and our research focuses inventory optimization given the fulfillment structure.

To the best of our knowledge, there are only two studies which come close to ours in emulating the problem setting, where online demand is integrated with the physical stores either through a separate warehouse for online orders (centralized), or through store fulfillment (decentralized). Bendoly et al. (2007) numerically compare fulfillment structures for handling online demand, identifying when a decentralized system is more beneficial as well as the factors influencing the choice, and Jalilipour Alishah et al. (2015) consider a single store with online and in-store demands, and analyzed decisions at three levels — fulfillment structure, inventory optimization and inventory rationing.

While Bendoly et al. (2007) focus on the centralized vs. decentralized debate with the help of numerical analyses, Jalilipour Alishah et al. (2015) consider only the single-store case, and their work does not extend to the more complex multi-store case. Our study provides analytical as well as numerical guarantees on the performance of our models, and focuses on optimizing the order-up-to levels for channel integration across multiple stores.

The virtual pooling of online demand, where orders at one store can be fulfilled from other stores can be seen as a form of transshipment. Yang and Qin (2007) refer to this as 'virtual lateral transshipment', where instead of items shipped between stores, a store directly ships the item to the customer at another location. Our problem is more sophisticated, as it involves a component of the demand (in-store demand) which cannot be

rerouted to other stores. Nevertheless, our problem can hence be related to the extensively-studied transshipment literature, particularly reactive transshipment, wherein transfer of goods occur after demands are realized, but before they have to be fulfilled.

Tagaras (1989) and Tagaras and Cohen (1992) consider a two-location periodic review problem with reactive transshipments, and show that the dominant strategy is to utilize transshipment as much as possible irrespective of inventory levels. There have also been numerous studies on reactive transshipment in a multi-location setting. Karmarkar and Patel (1977) studied the single-period problem by decomposing it into decoupled newsvendor problems and a transportation problem, while Robinson (1990) studied the multi-period problem and found that a base-stock policy is optimal in the case of identical costs. Bertrand and Bookbinder (1998) consider the case of non-identical costs and non-zero order lead-time, and develop a near-optimal algorithm for the redistribution. Transshipment models are usually complicated, and analytical solutions are seldom available, especially for multi-location problems. Our research adds to the complexity by introducing a component of the demand which cannot be subject to virtual transshipment.

Finally, our problem setting can be viewed as a form of newsvendor network, with virtual lateral transshipment as a 'discretionary policy' (Van Mieghem and Rudi 2002). Newsvendor networks have been analyzed in great detail by Van Mieghem and Rudi (2002) and Van Mieghem (2003), building up from the multi-dimensional newsvendor models proposed by Harrison and Van Mieghem (1999). Though the primary application of these models is capacity planning, they argue that no distinction can be made between inventory and capacity planning in a one-period case. However, the dual-demand setting makes the newsvendor network approach difficult even for two stores, let alone the multi-store case.

Due to the nature of omnichannel fulfillment, our work involves a culmination of problem features, such as demand pooling within a store and virtual transshipment across regions, and the ensuing complexity is no surprise. Our study is unique in the combination of the problem setting, analytic focus and research methodology.

3. The Two-Store Problem

We now formulate mathematical models for the two-store setting, for three systems of omnichannel fulfillment: 1) the no-integration (NI) system, 2) the partially-integrated (PI) system, and 3) the fully-integrated (FI) system as shown in Figure 1, where dashed lines

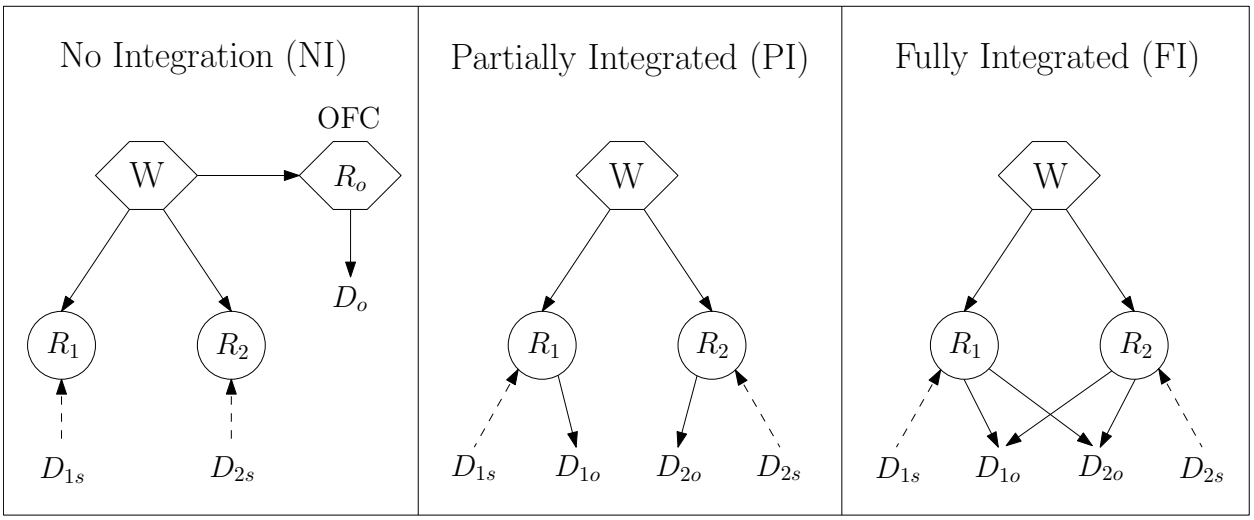


Figure 1 **Three systems of Omnichannel fulfillment a) no-integration (NI), b) partially-integrated (PI), and c) fully-integrated (FI).**

represent the flow of customers, and solid lines represent the flow of inventory. We will extend the formulations to the multi-store case in Section 4 for a more generalized setting.

We consider a system comprising of a firm with two retail stores R_1 and R_2 serving different regions, each selling a single product in a newsvendor fashion. There are two demand streams originating from each region — in-store demand (D_{1s}, D_{2s}) consisting of customers walking into the nearest store and buying the product, and online demand (D_{1o}, D_{2o}) consisting of customers ordering through the website or mobile app. The demands are assumed to be non-negative and continuous with well-behaved density functions.

The in-store and online demands within a region can be correlated as they represent the same segment of customers, while demands across regions can be independent. We only assume that the demands are jointly distributed, such that the total demands within each region and the total demand in the system have continuous and well-defined density functions. The demands are taken to be exogenous and independent of fulfillment structure.

As discussed earlier, customers can choose their type of fulfillment, such as ship-to-customer, pickup in-store, etc. The pickup in-store orders can be ignored without loss of generality as they can simply be included along with the in-store demand, since items are usually picked off the shelves as soon as the orders are received, and the orders are typically fulfilled on the same day. For our analyses, we consider only the ship-to-customer fulfillment, wherein an order placed online is delivered directly to the customer's home. We assume that unmet demand is lost, both for in-store and online demands.

In the no-integration (NI) model, the total online demand D_o $(= D_{1o} + D_{2o})$ is fulfilled by a dedicated online fulfillment center (OFC), designated by R_o. The retail stores R_1 and R_2 handle only in-store demands D_{1s} and D_{2s} respectively, and hence online and in-store demands are decoupled. This system is attractive to some retailers, as the decoupling of online and in-store demands makes managing the two channels easier by adoption of a silo-ed approach between the channels. This system can be useful in cases where staffing in stores is expensive, and the store staff cannot be used for fulfilling online orders, and also for cases where products are bulky (e.g. furniture) and require careful handling.

In the partially-integrated (PI) model, online orders are assigned to the nearest stores. In other words, online demand D_{1o} is assigned to store R_1, and D_{2o} to R_2. Thus in each region, in-store and online demands are pooled, and are fulfilled using the store inventory. In this system, stores can be independently managed due to decentralization, without investments for enterprise-wide inventory integration or setting up expensive fulfillment centers.

In the fully-integrated (FI) model, online orders are assigned to nearest stores, but in the event of stockout at the nearest store, an order can be fulfilled from another store with available inventory. Thus in addition to pooling of online and in-store demands within each region, there is *virtual pooling* of online demand across regions. This system can increase sales, reduce backorders and mitigate inventory levels across the stores. Note that unfulfilled in-store demand at one region cannot be fulfilled by stores in other regions. We will hereafter refer to these models as NI, PI and FI models respectively.

The FI model can also be extended to the case where some stores handle only one type of demand. The subsequent models can be derived by substituting $D_{is} = 0$ or $D_{io} = 0$ depending on the type of facility, with the NI and PI models extending trivially due to decentralization. Naturally, the results derived also hold for the pure play e-commerce case. We discuss this generalized setting for the multi-store case in Section 4.

We consider a one-period setting where stores R_1, R_2 and R_o (OFC) replenish their inventory to levels y_1, y_2 and y_o at the beginning of the period with zero replenishment leadtime from a supply with infinite capacity, and are subject to demands during the course of the period. All the inventory is available on the shelves of the stores, and the in-store demand consists of customers picking items off the shelves, with unmet store demand lost immediately. Online orders that are fulfilled from stores are handled in a different fashion: the store staff pick up the item from the shelf, followed by packing and labeling in the

back room of the store, and then the item is shipped to the customer. For the PI model, if inventory is unavailable, the demand is lost. However for the FI model, the order is rerouted to another store with available inventory, or if none exist, the demand is lost. In the NI model, all online orders are fulfilled from the OFC, and unmet demand is lost.

The sequence of events in store fulfillment is as follows: the in-store demand is fulfilled as it arrives, whereas the online demand is aggregated throughout the period, and fulfilled at the end of the period based on the model considered. We assume this sequence primarily for tractability purposes, because the model becomes intractable when in-store and online demands are fulfilled as they arrive randomly in tandem, as in addition to optimizing inventory levels, inventory rationing decisions also need to be made. Jalilipour Alishah et al. (2015) analyze this setting for a single store, and find that the optimal strategy is a threshold policy which depends on the cost of handling online orders and the mix of online and in-store demands. For our problem, virtual pooling among the different stores adds complexity by introducing rerouting of online orders as well as an additional rationing decision between online orders from the same region and those from other regions.

To balance the usefulness of the model with its complexity, we assume that in-store demand is prioritized over online demand. This assumption, although made for tractability, is justified in reality. Online orders have a fulfillment leeway — orders need not be fulfilled instantly as they arrive, and the delivery shipments can be expedited. In addition, one may argue that in the event of a stockout, disappointing a customer walking into the store can result in greater loss of goodwill than for a customer searching for the product online.

There have also been various studies analyzing benefits of postponing online fulfillment decisions in an e-commerce setting (Mahar and Wright 2009, Xu, Allgor, and Graves 2009). They find that delaying fulfillment decisions can lead to reduced fulfillment costs and number of shipments. Although this applies primarily to e-commerce fulfillment centers handling thousands of orders, a similar argument can be made for store fulfillment — aggregating online orders and fulfilling them together can lead to significant reductions in cost due to reduced labor hours involved in activities performed by store personnel.

We consider a per-unit service cost s_{ij} for online demand from region j fulfilled by the store from region i, and s_o for orders fulfilled by the OFC. This cost encapsulates the cost of picking the item off the shelf, packing and labelling, as well as the shipping cost for

delivery. We have $s_{ij} > s_{ii}, \forall j \neq i$, as fulfilling an online order from a different region is costlier than fulfilling an order from the same region because of longer distances.

In practice, the handling (pick-pack-and-label) component of the service cost is higher for stores fulfilling online demand, as it involves human labor, than for OFCs where the process can be automated. The shipping component of the service cost can be higher for the OFCs which are usually located farther away from population centers. With an abuse of terminology, we will refer to the service costs s_{ii} (within the same region) as *shipping* costs, and s_{ij} (across regions) as *cross-shipping* costs.

At the end of the period, for store R_i, each unit of unused inventory incurs an overage cost h_i, and each unit of unfulfilled in-store demand incurs a penalty cost p_{is}. The penalty cost for the online demand is p_o and is assumed to be constant irrespective of the fulfillment location. It follows that $p_o - s_{ii} > 0$, $\forall i$ and $p_o - s_o > 0$ as it is always better to fulfill online orders than not. We also have $p_{is} > p_o - s_{ii}$, $\forall i$, as the cost of not fulfilling a unit of in-store demand is higher than that of online demand. Similar assumptions are made by Seifert et al. (2006), who also assume that in-store demand is prioritized over online demand.

In the event that store R_i has a unit of available inventory after fulfilling the in-store and online demands from region i, we assume that it is always better to cross-ship to another region j with unfulfilled online demand. This is similar to the complete-pooling policy as proposed by Tagaras and Cohen (1992), which was found to be the dominant policy. The total cost incurred by not cross-shipping would be $h_i + p_o$, and hence $s_{ij} < h_i + p_o$, $j \neq i$.

We do not consider the purchasing cost of inventory, but this can be easily incorporated into the models as a linear term for the single period setting. The assumption that it is always better to cross-ship is valid for the one-period setting with lost sales, but for a multi-period setting with backlogging, it may be ex-post optimal to hold on to inventory for future periods rather than cross-ship to other regions, especially when purchasing costs are non-identical across regions. We discuss this interplay in more detail in Section 6.

3.1. The Three Basic Models

In this section, we formulate and solve the three proposed models of omnichannel retailing to obtain the optimal order-up-to quantities for a single period setting. The objective is to minimize the total expected one-period cost. The results obtained in the following sections inform the multi-store problem discussed in Section 4. We will later extend these models to the multi-period setting under lost sales in Section 6.

3.1.1. The No-Integration (NI) Model. In this model the in-store and online demands are decoupled, and by the independence of demands across regions, the problem becomes separable by region. This leads to solving the well-known newsvendor problem for each region, and the optimal order-up-to quantities are given by:

$$y_1^{NI} = F_{1s}^{-1}\Big(\frac{p_{1s}}{h_1+p_{1s}}\Big), \quad y_2^{NI} = F_{2s}^{-1}\Big(\frac{p_{2s}}{h_2+p_{2s}}\Big), \quad y_o^{NI} = F_o^{-1}\Big(\frac{p_o-s_o}{h_o+p_o-s_o}\Big) \tag{1}$$

where F_{1s}, F_{2s} and F_o are the cumulative distribution functions of the demands D_{1s}, D_{2s} and D_o respectively. Note that the underage cost for the online demand is the difference between the online penalty cost p_o and the service cost incurred in shipping the unit s_o.

3.1.2. The Partially-Integrated (PI) Model (Pooling within each Region). In this model, the in-store and online demands are pooled within each region. The stores replenish their inventories to quantities y_1, y_2 at the beginning of the period. The in-store demands are fulfilled first, followed by the fulfillment of online demand with the available inventory. This scenario is similar to the classic newsvendor model with multiple demand classes, as discussed by Şen and Zhang (1999). The total expected one-period cost function is:

$$\begin{aligned} C^{PI}(y_1,y_2) = \sum_{i=1,2}\Big[&\mathbb{E}h_i\left((y_i-D_{is})^+-D_{io}\right)^+ + \mathbb{E}p_{is}(D_{is}-y_i)^+ \\ &+\mathbb{E}p_o\left(D_{io}-(y_i-D_{is})^+\right)^+ + \mathbb{E}s_{ii}\min\left((y_i-D_{is})^+, D_{io}\right)\Big] \end{aligned} \tag{2}$$

where μ_{io} is the mean of the online demand D_{io}, and $x^+ = \max(x,0)$. The cost function is convex, which can be seen by expressing it in terms of the pooled demands $D_i = D_{is} + D_{io}$:

$$\begin{aligned} C^{PI}(y_1,y_2) = \sum_{i=1,2}\Big[&s_{ii}\mu_{io} + \mathbb{E}h_i\left(y_i-D_i\right)^+ \\ &+\mathbb{E}\left(p_o-s_{ii}\right)\left(D_i-y_i\right)^+ + \mathbb{E}\left(p_{is}-(p_o-s_{ii})\right)\left(D_{is}-y_i\right)^+\Big] \end{aligned} \tag{3}$$

The above form can be derived from Equation 2 using the identity $\min(x,y) = y-(y-x)^+$, and $(D_{is}-y_i)^+ + \left(D_{io}-(y_i-D_{is})^+\right)^+ = (D_i-y_i)^+$, which holds when demands are non-negative. The above representation can be explained as follows: the first two expectation terms correspond to a newsvendor cost function with the total demand having the same underage cost as the online demand. Since the store demand is fulfilled first, the final term makes up for the deficit in cost, as the underage cost is higher for the store demand.

This representation is useful, as it contains terms similar to the classic newsvendor cost function, with positive coefficients ($p_{is} > p_o - s_{ii}$). The pooled demand D_i has a well-defined

density function, and C^{PI} is thus convex as well as separable in its variables. The optimal order-up-to quantities (y_1^{PI}, y_2^{PI}) can be calculated from the implicit optimality equations:

$$(h_i + p_o - s_{ii})\, F_i\left(y_i^{PI}\right) + (p_{is} - (p_o - s_{ii}))\, F_{is}\left(y_i^{PI}\right) = p_{is} \qquad \forall i \tag{4}$$

where F_i is the cumulative distribution function of the pooled demand D_i. The equations provide unique order-up-to quantities, as the left hand side of the equation is an increasing function of the variable y_i^{PI}, and the right hand side is a constant. Similar to the classic newsvendor problem, the order-up-to quantities are increasing in the penalty costs, and decreasing in the overage cost as well as the shipping cost.

3.1.3. The Fully-Integrated (FI) Model (Pooling within and across Regions). This model is similar to the PI model, except that unfulfilled online orders from region j $(\neq i)$ can be fulfilled from region i provided there is available inventory in R_i, after R_i has fulfilled its own in-store and online demands. This *cross-shipping* is done at a per-unit cost of s_{ij} $(= s_{ji})$ and takes place after each store has attempted to fulfill its own in-store and online demands. In the two-store problem, the cross-shipped quantity from store R_i to region j can be explicitly calculated as the minimum of the inventory available at R_i and the unfulfilled online demand at R_j, after each store has attempted to fulfill its own in-store and online demands. The problem is not separable by region as was in the previous cases, and the total expected one-period cost function is:

$$\begin{aligned} C^{FI}(y_1, y_2) = \sum_i \Big[& \mathbb{E} h_i \left((y_i - D_{is})^+ - D_{io}\right)^+ + \mathbb{E} p_{is} (D_{is} - y_i)^+ \\ & + \mathbb{E} p_o \left(D_{io} - (y_i - D_{is})^+\right)^+ + \mathbb{E} s_{ii} \min\left((y_i - D_{is})^+, D_{io}\right) \Big] \\ & + (s_{12} - h_1 - p_o)\, \mathbb{E} \min\left(\left((y_1 - D_{1s})^+ - D_{1o}\right)^+, \left(D_{2o} - (y_2 - D_{2s})^+\right)^+\right) \\ & + (s_{12} - h_2 - p_o)\, \mathbb{E} \min\left(\left((y_2 - D_{2s})^+ - D_{2o}\right)^+, \left(D_{1o} - (y_1 - D_{1s})^+\right)^+\right) \end{aligned} \tag{5}$$

The only differences from the cost function of the PI model in Equation 2 are the last two terms, which represent the value of cross-shipping. By cross-shipping a unit from store R_i to region j, the total savings is $h_i + p_o$, as an excess unit is utilized to reduce unfulfilled demand, and a service cost of s_{ij} is incurred. Convexity does not directly ensue as in earlier cases, as the coefficients $s_{ij} - h_i - p_o$ are negative, and the cross-shipment quantities are non-convex. However, the total cost can be shown to be jointly convex in the order-up-to quantities (Proposition 1), for cost parameters satisfying assumptions made earlier:

$$\Phi = \Big\{ p_{is} > p_o - s_i > 0,\ \forall i; \quad h_i + p_o > s_{ij} > s_{ii},\ \forall i, j \neq i \Big\} \tag{6}$$

PROPOSITION 1. *Under the conditions on cost parameters in* Φ, $C^{FI}(y_1, y_2)$ *can be represented as the expectation of a linear program. As a consequence,* $C^{FI}(y_1, y_2)$ *is jointly convex in the order-up-to levels.*

Proof: To prove the proposition, we first observe that given a realization of the demands, the optimal cost can be obtained using a linear program. The proof follows in similar fashion to Seifert et al. (2006, Proposition 1). Consider the linear program in (7), where z_i represents the amount of inventory at R_i used to fulfill its in-store demand, and z_{ij} represents the amount of inventory of R_i used to fulfill online demand of R_j (z_{ii} is the amount of inventory used by R_i to fulfill its own online demand).

$$
\begin{aligned}
\underset{z_i, z_{ii}, z_{ij}}{\text{minimize}} \sum_i h_i(y_i - z_i - \sum_j z_{ij}) + \sum_i p_{is}(D_{is} - z_i) & \\
+ \sum_i p_o(D_{io} - \sum_j z_{ji}) + \sum_i s_{ii} z_{ii} + \sum_i \sum_{j \neq i} s_{ij} z_{ij} & \\
\text{subject to} \quad z_i + \sum_j z_{ij} \leq y_i, & \quad \forall i \\
z_i \leq D_{is}, & \quad \forall i \\
\sum_j z_{ji} \leq D_{io}, & \quad \forall i \\
z_i \geq 0, & \quad \forall i \\
z_{ij} \geq 0, & \quad \forall i, j
\end{aligned}
\tag{7}
$$

Let $P(y_1, y_2, \bar{D})$ be the optimal value of the linear program, where $\bar{D}$ represents the demand vector $(D_{1s}, D_{2s}, D_{1o}, D_{2o})$. We shall show that, given a realization of demands (fixed $\bar{D}$), the function P represents the cost of the FI model as a function of the order-up-to quantities y_1 and y_2. First, notice that the coefficients of the decision variables z_i, z_{ii}, $z_{ij,(j\neq i)}$ in the objective function are $(-h_i - p_{is})$, $(s_{ii} - h_i - p_o)$, $(s_{ij} - h_i - p_o)$ respectively.

Under the conditions in Φ in Equation 6, we see that $(-h_i - p_{is}) < (s_{ii} - h_i - p_o) < (s_{ij} - h_i - p_o)$. The linear program can be solved greedily, and it is easy to see that the optimal solution is given by $z_i = \min(y_i, D_{is})$, $z_{ii} = \min\left((y_i - D_{is})^+, D_{io}\right)$, $z_{ij} = \min\left(\left((y_i - D_{is})^+ - D_{io}\right)^+, \left(D_{jo} - (y_j - D_{js})^+\right)^+\right)$.

The sequence of fulfillment is clear: in-store demand is fulfilled first, followed by online demand from the same region, and finally cross-shipment to other regions. Hence, we have $C^{FI}(y_1, y_2) = \mathbb{E}_{\bar{D}}\left(P\left(y_1, y_2, \bar{D}\right)\right)$. The objective function is linear and the constraint set in (7) is a polyhedral convex set with linear constraints, and hence by the projection theorem

of Heyman and Sobel (2003), P is jointly convex in y_1, y_2, $\bar{D}$. As the expectation of a convex function is convex, it follows that $C^{FI}(y_1, y_2)$ is jointly convex in y_1 and y_2. □

In the case where demands follow discrete distributions taking values from finite sets, the cost function can be written as the weighted sum of optimal values of linear programs for each demand realization vector, with weights corresponding to realization probabilities. This can in turn be expressed as a single linear program, from which optimal order-up-to quantities y_1^{FI}, y_2^{FI} can be found. The downside to this approach is that as the number of discretizations of the demands grows, the linear program grows exponentially in size.

In the case where demands follow continuous distributions, we can employ convex optimization methods to obtain the optimal solution as the cost function is jointly convex in the variables. Proposition 2 provides an easy way to calculate the gradient of C^{FI}.

PROPOSITION 2. *There exist regions $\Omega_i(y_1, y_2)$ in the demand space, such that in each region the dual-price vector λ^i corresponding to the variables y_1, y_2 remains constant, and the gradient of the cost function of the FI model can be written as*

$$\nabla C^{FI}(y_1, y_2) = (h_1, h_2)^T - \sum_i \lambda^i \mathbb{P}\left(\Omega_i\left(y_1, y_2\right)\right) \tag{8}$$

Proof: The proof follows along the lines of the proof of Proposition 2 in Harrison and Van Mieghem (1999), as the cost function of the FI model has a similar structure to newsvendor networks. The gradient of the function $P\left(y_1, y_2, \bar{D}\right)$ with respect to (y_1, y_2) is

$$\nabla_{y_1, y_2} P\left(y_1, y_2, \bar{D}\right) = (h_1, h_2)^T - \lambda\left(y_1, y_2, \bar{D}\right) \tag{9}$$

where $\lambda\left(y_1, y_2, \bar{D}\right)$ is the dual-price vector corresponding to the constraints with y_1 and y_2 in (7), and the vector $(h_1, h_2)^T$ arises from the coefficients of y_1 and y_2 in the objective function. The 4-dimensional demand space can be divided into domains $\Omega_i\left(y_1, y_2\right)$ such that in each domain, the optimal values of the decision variables z_i, z_{ii} and z_{ij} are linear in y_1 and y_2, and the dual-price vector $\lambda\left(y_1, y_2, \bar{D}\right)$ is constant (refer to Appendix B for a discussion with a simplified case). The first-order conditions for the FI model are:

$$\nabla_{y_1, y_2} C^{FI}(y_1, y_2) = 0 = \nabla_{y_1, y_2} \mathbb{E}_{\bar{D}}\left(P\left(y_1, y_2, \bar{D}\right)\right) \tag{10}$$

We can interchange the gradient and expectation on the right hand side of Equation 10 for newsvendor networks, similar to Proposition 1 of Van Mieghem and Rudi (2002) (see Harrison and Van Mieghem (1999) for a proof), and thus Equation 10 becomes

$$\begin{aligned}\nabla_{y_1,y_2} C^{FI}(y_1,y_2) = 0 = \mathbb{E}_{\bar{D}} \nabla_{y_1,y_2} P\left(y_1,y_2,\bar{D}\right) = (h_1,h_2)^T - \mathbb{E}_{\bar{D}} \lambda\left(y_1,y_2,\bar{D}\right) \\ = (h_1,h_2)^T - \sum_i \lambda^i \mathbb{P}\left(\Omega_i\left(y_1,y_2\right)\right)\end{aligned} \tag{11}$$

where λ^i is the constant $\lambda\left(y_1,y_2,\bar{D}\right)$ for $\bar{D} \in \Omega_i\left(y_1,y_2\right)$. □

The optimality equation is given by

$$\sum_i \lambda^i(y_1^{FI},y_2^{FI}) \mathbb{P}\left(\Omega_i\left(y_1^{FI},y_2^{FI}\right)\right) = (h_1,h_2)^T \tag{12}$$

The optimal probabilities $P_i^* = \mathbb{P}\left(\Omega_i\left(y_1^{FI},y_2^{FI}\right)\right)$ are the 'critical fractiles' (Van Mieghem and Rudi 2002) of the multivariate distribution at which the underage and overage costs are optimally balanced for the system as a whole. Having obtained the gradient of the cost function, we can use the gradient descent algorithm to obtain the optimal solution.

In Section 5.1, we employ this approach in the two-store setting for a special case where the in-store and online demands are perfectly correlated, thereby reducing the number of random variables involved to two. This approach becomes complicated even when the number of random variables involved is more than two, as determination of the demand regions Ω_i's becomes non-trivial. The number of demand regions Ω_i's also increases exponentially with the number of stores. Hence we develop a simple heuristic in Section 4 to calculate order-up-to quantities for systems with large number of stores.

3.2. Partially-Integrated with Ex-Post Cross-Shipping (PICS) Model

The FI system has many advantages over the PI system due to cross-shipping, such as reduction in unfulfilled online demand, reduction in inventory held at the end of a period, etc. However in practice, unplanned cross-shipping does happen on an ad-hoc basis, where unfulfilled orders from one store are routed to a nearby store even though this is not taken into account while planning for inventory at either store. We refer to this system as the PICS system, and we model this as an extension of the PI model.

For the PICS system, the need for cross-shipping is not anticipated in advance, and hence inventory planning is done individually for each region based on the PI model, with cross-shipping done in an ex-post fashion after demands are realized. For notational brevity,

consider identical costs across stores, with s being the shipping cost for fulfilling online demand from the same region, and s' for fulfilling online demand from other regions. The optimal expected one-period cost function for the PICS model is as follows:

$$C^{PICS} = C^{PI}(y_1^{PI}, y_2^{PI}) + (s' - h - p_o)\,\mathbb{E}\left(\mathrm{CS}\left(y_1^{PI}, y_2^{PI}, \bar{D}\right)\right) = C^{FI}(y_1^{PI}, y_2^{PI}) \tag{13}$$

The function $CS(y_1, y_2, \bar{D})$ represents the total number of units cross-shipped given that the inventory levels were y_1, y_2 at the beginning of the period. The equality thus follows from Equation 5, as $C^{FI}(y_1, y_2) = C^{PI}(y_1, y_2) + (s' - h - p_o)\,\mathbb{E}\left(\mathrm{CS}\left(y_1, y_2, \bar{D}\right)\right)$. The PICS model gives us an easy way to compare the FI and PI models, and the optimal expected one-period cost functions of FI, PI and PICS models obey the following inequality:

$$C^{FI}(y_1^{FI}, y_2^{FI}) \leq C^{PICS} \leq C^{PI}(y_1^{PI}, y_2^{PI}) \tag{14}$$

The first inequality follows from the optimality of (y_1^{FI}, y_2^{FI}) under C^{FI}, and the second inequality follows from the fact that $CS(y_1, y_2, \bar{D}) \geq 0$, and $(s' - h - p_o) < 0$. The FI model has a lower cost than the PI model as well as the PICS model which can be attributed to the effect of pooling across regions and planning for cross-shipping in advance. As long as cross-shipping is beneficial ($s' - h - p_o < 0$), the FI model utilizes the option more as a result of planning for cross-shipping in advance, which leads us to Proposition 3.

Proposition 3. $\mathbb{E}\left(CS\left(y_1^{FI}, y_2^{FI}, \bar{D}\right)\right) \geq \mathbb{E}\left(CS\left(y_1^{PI}, y_2^{PI}, \bar{D}\right)\right)$

The proof is included in the Appendix. Naturally, the value of virtual pooling depends on the cost s', and high values of cross-shipping cost can diminish the advantage of the FI system over the PICS system. The NI model has a different cost structure compared to the other models because of the OFC. However, we compare the NI and FI models numerically in Section 5, for the case where the overage costs are identical for the stores and the OFC.

4. The Multi-Store Problem

We extend the two-store problem discussed so far to a generalized setting with multiple regions. Consider a system composed of a firm which owns N facilities $R_1, R_2, \ldots, R_N$ in different regions. These facilities are not necessarily all physical stores — we consider three different types of facilities described by the following sets:

- $\mathcal{S}_s$ - physical stores which handle only in-store demand, and do not cross-ship to other regions. Thus, $\forall i \in \mathcal{S}_s$, $D_{io} = 0$.

- $\mathcal{S}_o$ - online fulfillment centers (OFCs) which handle only online orders: $\forall i \in \mathcal{S}_o, D_{is} = 0$.
- $\mathcal{S}_{so}$ - omnichannel physical stores which handle both online and in-store demands.

The facilities have identical cost structures, with overage cost h, in-store penalty cost p_s, and online penalty cost p_o wherever applicable. The within-region service cost for online orders is s and is identical across regions. The per-unit cross-shipping cost from facility R_i to region j is s_{ij} $(= s_{ji})$. The cross-shipping costs are assumed to be directly dependent on the distances between the regions, with $s_{ij} = s + f(d_{ij})$, where d_{ij} is the distance between region i and region j, and f is a non-negative, increasing function such that $f(d) \to 0$ as $d \to 0$. Also, $\sup_{d \in \mathcal{D}} f(d) \leq h + p_o - s$ where $\mathcal{D}$ is the set of all distances between regions, so that the conditions on cost parameters as defined earlier in Equation 6 are satisfied.

The NI and PI models readily extend to the multiple-store case, as these cost functions are separable by region. However, an explicit cost function cannot be derived for the FI model as was done in Equation 5 for the two-store case. The cross-shipment quantities are set by a transportation linear program, where facilities having excess inventory act as supply points and regions with unfulfilled online demand act as demand points.

An easier alternative is to express the cost function as the expectation of a linear program, as was done in (7). Propositions 1 and 2 still hold, as the cost parameters satisfy the conditions in Equation 6 and the gradient descent method can be employed to obtain the optimal solution. As discussed earlier, this approach becomes impractical for larger number of stores, as the number of demand regions as described in the proof of Proposition 2 increases exponentially in the order of $(2N)^N$, and the calculation of probabilities become computationally expensive due to the number of random variables involved. To tackle the curse of dimensionality, we develop a heuristic derived from a lower bound to the multi-store FI model. We then show that under certain conditions, the lower bound is near optimal in an asymptotic sense when there are a large number of omnichannel stores.

4.1. Lower Bound (LB) and Heuristic for the Multi-Store Problem

For physical stores that are not involved in virtual pooling ($\mathcal{S}_s$), optimal order-up-to quantities can be easily found using the newsvendor model, and hence we ignore these stores in the following formulation. Let $\mathcal{S} = \mathcal{S}_o \cup \mathcal{S}_{so}$ be the set of facilities involved in virtual pooling. In order to derive a lower bound, we revisit the two-store FI model. An important feature which complicates the FI model is that the in-store demands are not pooled across regions, which in turn leads to complicated coupled terms in the cost function. We relax

this by treating unfulfilled in-store demand as online demand which can be fulfilled by cross-shipping. Keeping the sequence of fulfillment intact, we obtain the cost function for the case where all demands are pooled across regions, with identical costs across stores:

$$\begin{aligned} C^{FIp}(y_1,y_2) = s(\mu_{1o}+\mu_{2o}) + \Big[& \mathbb{E}h\left(y_1+y_2-D\right)^+ + \mathbb{E}\left(p_o-s'\right)\left(D-y_1-y_2\right)^+ \\ & \mathbb{E}\left(p_o-s-(p_o-s')\right)\left(D_1-y_1\right)^+ + \mathbb{E}\left(p_o-s-(p_o-s')\right)\left(D_2-y_2\right)^+ \\ & \mathbb{E}\left(p_s-(p_o-s)\right)\left(D_{1s}-y_1\right)^+ + \mathbb{E}\left(p_s-(p_o-s)\right)\left(D_{2s}-y_2\right)^+ \Big] \end{aligned} \tag{15}$$

where s' is the cross-shipping cost, and $D = D_1 + D_2$ is the total demand. We find that the above cost function is significantly less complex compared to C^{FI}, and has a similar structure to the cost function of the PI model in Equation 3. The first two terms represent the cost of a newsvendor setting where all the demand is fulfilled by cross-shipping. The remaining terms account for the underage costs for demands fulfilled within the same region, with modified coefficients to account for double-counting of unfulfilled demand.

Equation 15 is of a form which can be extended to multiple stores, with one consideration: the cross-shipping cost is different for different pairs of regions, as $s_{ij} = s + f(d_{ij})$. We circumvent this issue and extend the above formulation to multiple locations by lowering the cross-shipping cost s' to s, the constant within-region shipping cost, and this gives us a valid lower bound to the multi-store problem, as discussed in Proposition 4.

PROPOSITION 4. *$C^{LB}(y_1,\ldots,y_N) \le C^{FI}(y_1,\ldots,y_N)$, $\forall y_1,\ldots,y_N \ge 0$, where*

$$\begin{aligned} C^{LB}(y_1,\ldots,y_N) = s\sum_{i\in\mathcal{S}}\mu_{io} + \Bigg[& \mathbb{E}h\left(\sum_{i\in\mathcal{S}} y_i - D_{\mathcal{S}}\right)^+ + \mathbb{E}\left(p_o-s\right)\left(D_{\mathcal{S}} - \sum_{i\in\mathcal{S}} y_i\right)^+ \\ & + \mathbb{E}\left(p_s-p_o+s\right)\sum_{i\in\mathcal{S}_{so}}\left(D_{is}-y_i\right)^+ \Bigg] \end{aligned} \tag{16}$$

with $D_{\mathcal{S}} = \sum_{i\in\mathcal{S}} D_{is} + D_{io}$.

The proof is relegated to the Appendix. This lower bound is useful, as C^{LB} is convex in the order-up-to quantities, and can be solved to yield a heuristic solution. However, due to the assumption that all cross-shipping can take place at a cost s, the order quantities at the OFCs are zero in the optimal solution. This is because a unit of inventory at the OFC can lead to a decreased cost if it was rather at a store, as in-store demand has a higher underage cost than online demand. The implicit optimality equations for C^{LB} are:

$$(h+p_o-s)F_{D_{\mathcal{S}}}\left(\sum_{j\in\mathcal{S}} y_j^{LB}\right) + (p_s-p_o+s)F_{D_{is}}(y_i^{LB}) = p_s, \quad \forall i\in\mathcal{S}_{so} \tag{17}$$

where $(y_1^{LB}, \ldots, y_N^{LB})$ is the optimal solution with $y_i^{LB} = 0$, $\forall i \in \mathcal{S}_o$. We revise y^{LB} to obtain the heuristic solution y^{FIH} by calculating order quantities for the OFCs separately, and using them in Equation 17 to compute order quantities for the omnichannel stores. The order-up-to quantities for OFCs are calculated from the pooled total order quantity for OFCs, which is determined using the newsvendor quantity for the combined online demand.

$$\sum_{j \in \mathcal{S}_o} y_j^{FIH} = F_{D_{\mathcal{S}_o}}^{-1}\left(\frac{p_o - s}{h + p_o - s}\right) \tag{18}$$

where $D_{\mathcal{S}_o} = \sum_{i \in \mathcal{S}_o} D_{io}$. To calculate the individual order quantities y_i^{FIH}, $i \in \mathcal{S}_o$, we use the method of obtaining order-up-to quantities for multiple products with capacity constraints, as described in Chopra and Meindl (2007, p. 367). The total capacity is the total order-up-to quantity calculated from Equation 18, and the order-up-to quantity for each product corresponds to the order-up-to quantity for each OFC. Each unit from $\sum_{j \in \mathcal{S}_o} y_j^{FIH}$ is allocated incrementally to the OFCs based on the individual expected marginal costs. Once the order-up-to quantities for the OFCs are obtained, they are used in Equation 19 to determine order-up-to levels for other omnichannel stores.

$$(h + p_o - s) F_{D_{\mathcal{S}}}\left(\sum_{j \in \mathcal{S}} y_j^{FIH}\right) + (p_s - p_o + s)\, F_{D_{is}}\left(y_i^{FIH}\right) = p_s, \quad \forall i \in \mathcal{S}_{so} \tag{19}$$

The cost of the heuristic solution is calculated as $C^{FIH} = C^{FI}(y_1^{FIH}, \ldots, y_N^{FIH})$, and the lower bound is $C^{LB}(y_1^{LB}, \ldots, y_N^{LB})$. We thus capture the effect of virtual pooling among the facilities, and the systematic approach is shown in Algorithm 1.

The performance of the heuristic clearly depends on the value $\max_{i,j}(\frac{s_{ij}}{s})$, however, in practice, this ratio is quite small, and is less than 2 for locations in the mainland US based on UPS ground shipping rates. For cases where the in-store channel dominates, the heuristic performance can also be affected by the fact that it assumes that all demands are pooled across locations, when in reality, the in-store demands are not pooled.

As the problem scale increases, and the number of stores grows large within a given area to accommodate the increase in demand, it is highly likely that a store with unfulfilled online demand can find a close-by store with available inventory. Hence, almost all cross-shipping takes place over short distances, at a cost close to s, the within-region shipping cost. Hence, we can expect the heuristic solution to be close to the optimal solution, as the heuristic assumes a constant cross-shipping cost s. As a consequence of this notion, Proposition 5 shows that the heuristic is near optimal in an asymptotic sense.

Algorithm 1 Procedure to calculate the heuristic solution $(y_1^{FIH}, \ldots, y_N^{FIH})$

1: For physical stores in set $\mathcal{S}_s$, set $y_i^{FIH} = F_{is}\left(\frac{p_s}{h+p_s}\right), \forall i \in \mathcal{S}_s$.

2: **for** $i \in \mathcal{S}_o$ (OFCs) **do**

3: Calculate total order quantity: $y^{TOT} = F_{D_{\mathcal{S}_o}}\left(\frac{p_o - s}{h+p_o-s}\right)$, where $D_{\mathcal{S}_o} = \sum_{i\in\mathcal{S}_o} D_{io}$.

4: Set $y_i^{FIH} = 0, \forall i \in \mathcal{S}_o$, and $rem = \lfloor y^{TOT} \rfloor$.

5: Calculate the marginal cost of store $i \in \mathcal{S}_o$ as $MC_i\left(y_i^{FIH}\right) = -(p_o - s)\left(1 - F_{D_{io}}\left(y_i^{FIH}\right)\right) + hF_{D_{io}}\left(y_i^{FIH}\right)$

6: Choose $i^* = \min_{i\in\mathcal{S}_o} MC_i(y_i^{FIH})$. Set $y_{i^*}^{FIH} \leftarrow y_{i^*}^{FIH} + 1$

7: Set $rem \leftarrow rem - 1$. If $rem > 0$, go to Step 3.

8: **for** $i \in \mathcal{S}_{so}$ **do**

9: Calculate order quantities implicitly from the optimality equations: $(h + p_o - s) F_{D_{\mathcal{S}}}\left(\sum_{j\in\mathcal{S}} y_j^{FIH}\right) + (p_s - p_o + s) F_{D_{is}}\left(y_i^{LB}\right) = p_s, \ \forall i \in \mathcal{S}_{so}$.

Proposition 5. *As the number of omnichannel stores in a given area increases, with demands bounded and i.i.d. across regions, for sufficiently small $h > 0$, the heuristic is near optimal in an asymptotic sense with a constant approximation factor, i.e.*

$$\frac{C^{FIH}}{C^{LB}(y^{FIH})} \leq \frac{h + p_s}{p_s - p_o + s}, \text{ as } N \to \infty$$

The proof is relegated to the Appendix. The proposition holds only when all locations have omnichannel stores, and $y^{LB} = y^{FIH}$. The heuristic deviates from optimal by two assumptions - the cost of shipping an item is s irrespective of distance, and all demands are pooled across locations. We first show that the first assumption does not lead to any changes from the optimal cost, by considering a simplified setting where the stores are uniformly distributed in the given region, which is in-turn divided into identical sub-regions. As the number of stores grows large, each sub-region has sufficient supply to fulfill its demands, and hence cross-shipping takes place only within the sub-regions. The size of these sub-regions grow smaller as N increases, and the per-unit cross-shipping cost for every unit approaches the value s due to the functional form of s_{ij}.

The heuristic solution still suffers from the second assumption, specifically in cases where there are unfulfilled in-store demands at various locations, and the system as a whole has enough inventory to fulfill all demands. We bound the heuristic performance by a constant

approximation factor, which is dependent on cost parameters as one would expect. We evaluate the performance of the heuristic under various conditions in the numerical studies in Section 5.2. We find that the ratio of heuristic to the lower bound is low in most cases: for example, when online demand dominates, the ratio of heuristic to the lower bound was found to be less than 1.03 (Figure 5b).

5. Numerical Analysis

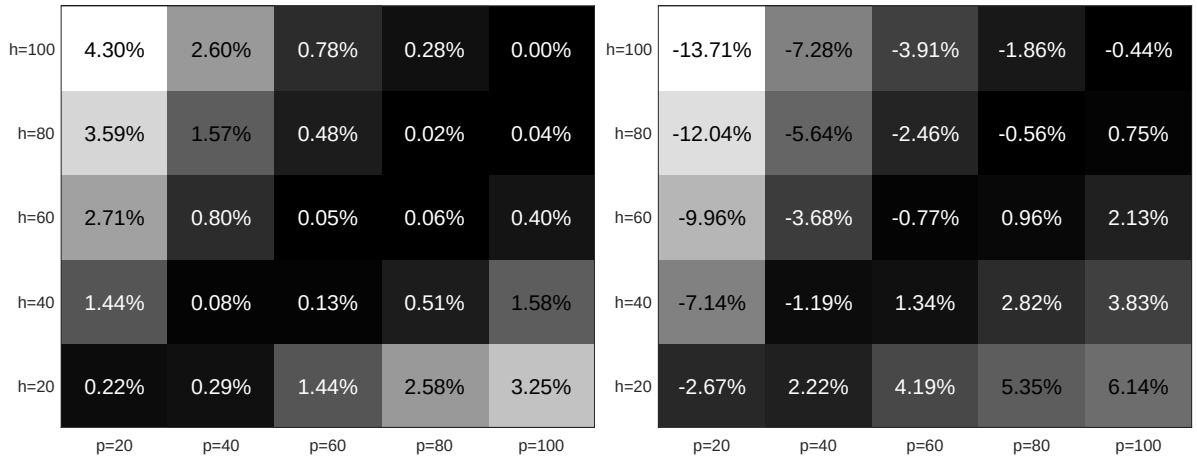
In this section, we first numerically compare the NI, PICS and FI models for the two-store case, and evaluate the performance of the heuristic in the multi-store case for a network of facilities embedded in the mainland US. The results from the two-store case address the centralized vs. omnichannel debate as well as the pooling benefits reaped by the fully-integrated FI model, whereas the results from the multi-store case showcase the improvements wrought by the heuristic as compared to the PICS model used in practice.

5.1. Two-Store Case

We perform sensitivity analysis to understand the effects of cost and demand parameters on the NI, PICS and FI models. Obtaining optimal order-up-to quantities for the FI model involves non-trivial identification of demand region where dual prices are constant, and cumbersome calculation of their probabilities (Appendix B). Hence, we seek to reduce the computational effort by assuming perfect positive correlation between the in-store and online channels, thereby reducing the number of random variables in the two-store case.

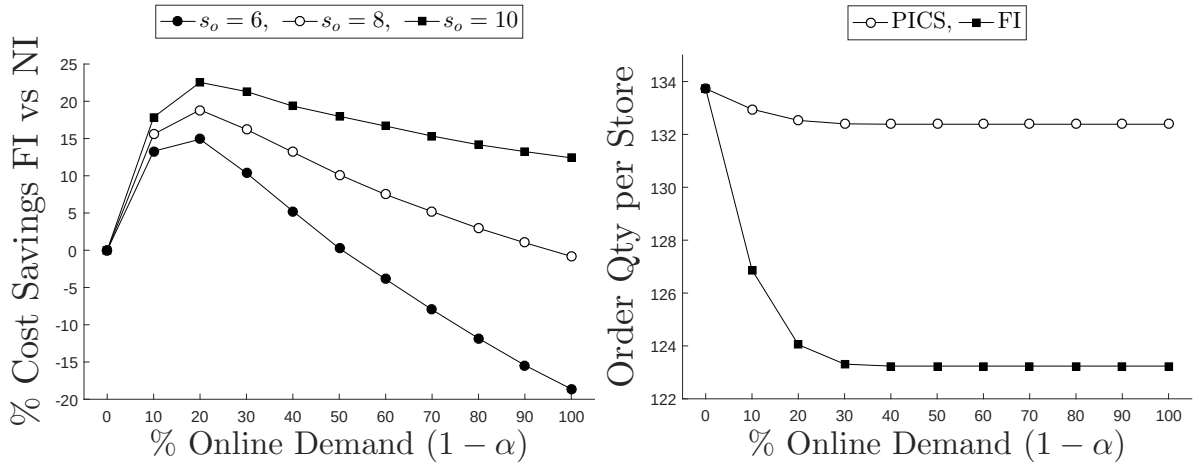
Let M_i be the random market size of region i (total number of customers purchasing either online or in-store), and let α_i be the fixed proportion of the customer demand which accounts for in-store sales. M_1 and M_2 are assumed to be independent as they represent different segments of customers. The in-store demands are hence $\alpha_1 M_1$, $\alpha_2 M_2$ and the online demands are $(1-\alpha_1)M_1$, $(1-\alpha_2)M_2$. Further details can be found in Appendix A. We take identical demand and cost parameters across regions: $M_1, M_2 \sim \mathcal{N}(100, 30)$, $\alpha_1 = \alpha_2 = 0.75$, $h_o = h = 15$, $p_s = 100$, $p_o = 100$, $s_o = s = 8$, $s_{12} = 12.5$. The penalty costs are higher than the overage cost, which is common in newsvendor settings.

5.1.1. Holding and Penalty Costs. We study the differences between the Fi and PICS models for different values of overage and underage costs keeping $p_s = p_o = p$ (Figure 2). Darker colors indicate that there is little difference between the FI and PICS models, whereas lighter colors show the cases where the FI model most differs from the PICS model.



(a) $(C^{PICS} - C^{FI})/C^{PICS}$ (b) $(y^{PICS} - y^{FI})/y^{PICS}$

Figure 2 **Shows the effect of holding and penalty costs on % cost savings and % change in order quantity between FI and PICS models**



(a) $(C^{NI} - C^{FI})/C^{NI}$ (b) y^{PI}, y^{FI}

Figure 3 **Shows the effect of increasing % of online demand on % cost savings achieved by FI compared to NI (left) and the Order Quantity per store (right)**

The FI model as a newsvendor network balances the overage and underage costs for the system as a whole, and hence benefits from pooling across regions in addition to pooling across channels within each region. However, when these costs are already balanced (seen in the non-leading diagonal in Figure 2a), the pooling benefit from reduction in variability is ineffectual, and there is little difference between the FI and PICS models. The FI model naturally performs best when the holding and penalty costs are unbalanced, by adjusting the order quantities to reflect the availability of the cross-shipping option (Figure 2b).

5.1.2. Market Share of E-commerce ($1-\alpha$). E-commerce sales as a proportion of total retail sales in the US is still very small (around 7%), however certain products have large e-commerce market shares. For instance, the e-commerce market share for computers and consumer electronics is around 50%, whereas for groceries which are purchased mainly in-store, it is close to 5% (FTI Consulting 2015). We compare the centralized (NI) and omnichannel (FI) approaches in handling online demand for varying α (Figure 3).

We see that pooling online demands at a central location does not add much value when the online demand is small compared to the in-store demand, whereas the FI model additionally benefits from cross-channel pooling. However, as the online channel becomes dominant, a centralized approach becomes valuable dependent on the service cost of fulfilling online orders from the OFC, as seen in Figure 3a.

Another interesting observation is that the optimal order quantities for the PI and FI models converge to the pure play e-commerce case for $\alpha < 0.5$ (Figure 3b). When the online channel dominates the store channel, the model treats the in-store demand as if it were online demand. This happens because the in-store demand is always fulfilled first, and all of the small in-store demand can be fulfilled with a high probability. Hence, there is no need to account for the underage cost for in-store demand separately, and the in-store demand can be accounted for along with the online demand.

5.2. Multiple-Store Case

We employ a more realistic setting for the multiple-store case to test the performance of the heuristic. We take the locations of the stores to be at the most populous cities in mainland US (Wikipedia 2016) and the OFCs are located according to the list of most efficient warehouses in the US, in terms of possible transit lead-times (Chicago Consulting 2013). The shipping costs are calculated using the cost equation estimated by Jasin and Sinha (2015) based on UPS Ground shipping rates: $s_{ij} = 9.182 + 0.000541 d_{ij}$, where d_{ij} is the distance in miles from region i to region j. We assume the constant within-region shipping cost to be $s = 9.182$. Other cost parameters used are: $h_o = h = 15$, $p_s = 100$, $p_o = 90$.

We assume that the in-store and online channels are independent and normally distributed, and all demands have a coefficient of variation of 0.3. The average values of the demands are calculated as proportions of the population of the cities, with α being the average market share of the in-store channel. The OFCs fulfill online demand from cities without physical stores. Further details can be found in Appendix A.

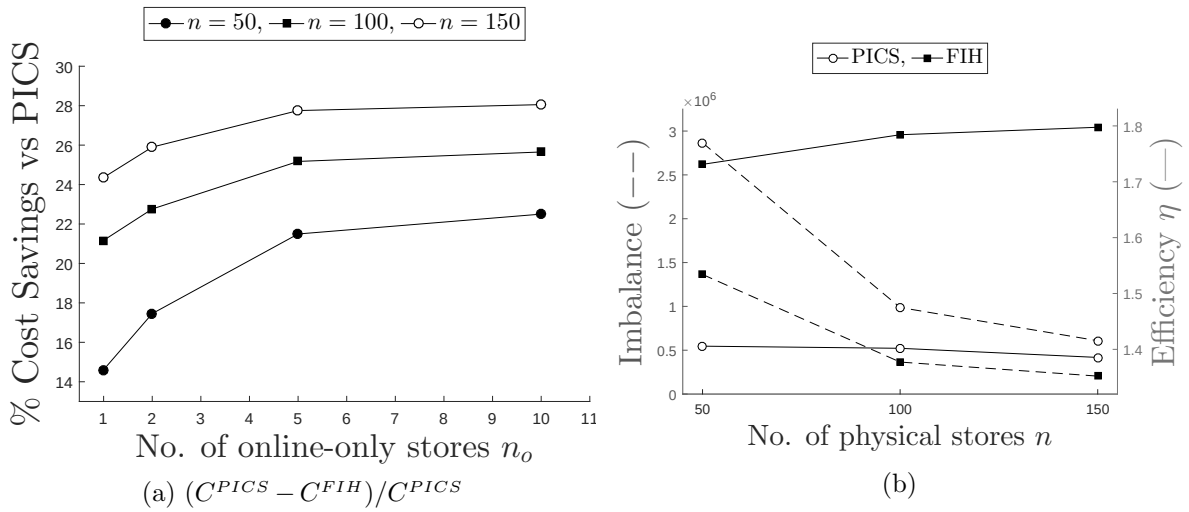


Figure 4 **Shows the effect of increasing number of stores and online warehouses on % cost savings achieved by heuristic against the PICS model (left) and the improvements in inventory efficiency achieved by the heuristic compared to the PICS model(right)**

There are n physical stores, located at the n most populous cities. Of these cities, the top 80% of cities in terms of population have omnichannel stores ($n_{so} = |\mathcal{S}_{so}|$), and the remaining cities have stores which handle only in-store demand and do not handle any online orders ($n_s = |\mathcal{S}_s|$). This is because store fulfillment capabilities are usually rolled out to physical stores in more populous cities first, as they can cater to a larger market.

5.2.1. Number of Stores. Let $n = n_s + n_{so}$ be the number of physical stores and n_o be the number of OFCs, so that $n + n_o = N$. We first compare the heuristic FIH to the PICS model to evaluate potential savings which can be achieved by the heuristic over current practices, and the results are shown in Figure 4. From Figure 4a, we find that the heuristic provides significant cost savings of 30% compared to the PICS model, for a network of 150 stores and 10 OFCs, with $\alpha = 50\%$. We also observe that increasing the number of stores with store fulfillment facilities, or increasing the number of OFCs improves cost savings by adding more avenues for virtual pooling, albeit in a marginal fashion as expected.

We find similar results in comparison to the NI model. The heuristic outperformed the NI model and achieved significant cost savings around 60% for a network of 150 stores and 10 OFCs. In our study, we assume the same service costs for both OFCs and physical stores, and the cost savings achieved by the heuristic does depend on the service cost from OFCs which can be different from the service cost from physical stores.

The results in Figure 4b address improvements in efficiency of inventory management for an omnichannel setting brought by our heuristic. As an equivalent measure for inventory turnover in a single-period setting, we calculate the inventory efficiency through the efficiency index η, which we define as the ratio of the total filled demand to the average inventory level of the system, which is in turn calculated as the mean of the total order-up-to quantity and the expected total ending inventory in the system. Keeping $n_o = 2$, we see that the heuristic is more efficient compared to the PICS model, and the increase in efficiency is higher for larger networks as a result of increased flexibility due to pooling across more regions. Higher efficiency stems from a reduction in inventory levels, albeit without a considerable decrease in service levels due to effective utilization of cross-shipping.

Another important performance metric is the level of inventory imbalance across stores. Higher imbalance can lead to costly spillovers and local stockouts (Acimovic and Graves 2015), which in turn can cause markdowns in stores. We calculate inventory imbalance as the variance of the ending inventory positions across stores at the end of the period. Although this is different from the metric used by Acimovic and Graves (2015) for e-commerce fulfillment centers, it captures the essence of imbalance among stores in an omnichannel network. From Figure 4b, we see that the heuristic achieves significant reduction in inventory imbalance compared to the PICS model. Planning for cross-shipping in advance lowers inventory imbalance by mitigating inventory levels across facilities.

For a network with $n = 150$ and $n_o = 10$, calculating the order-up-to quantities using the heuristic takes around 6 minutes, whereas calculating expected costs through simulation with 15,000 iterations takes around 4 hours. In real life, the networks are bigger in size — out of its 1800 stores, Target has more than 1000 stores which have store fulfillment capabilities (Lindner 2016). For larger networks, potential cost savings could be much higher, and computations can be run much faster with better computational facilities.

5.2.2. Market Share of E-commerce. We study the effect of market share of e-commerce sales for a network of 50 stores and 2 OFCs (Figure 5). The heuristic derives maximum benefit from virtual pooling when the online market share is moderate (Figure 5a), similar to the two-store case. When the online channel dominates, the heuristic is within 2% of the lower bound (Figure 5b), but performs poorly when the online market share is low, as the assumption that all demands are pooled across locations for the heuristic becomes costly for products with dominant in-store demand. However, for such

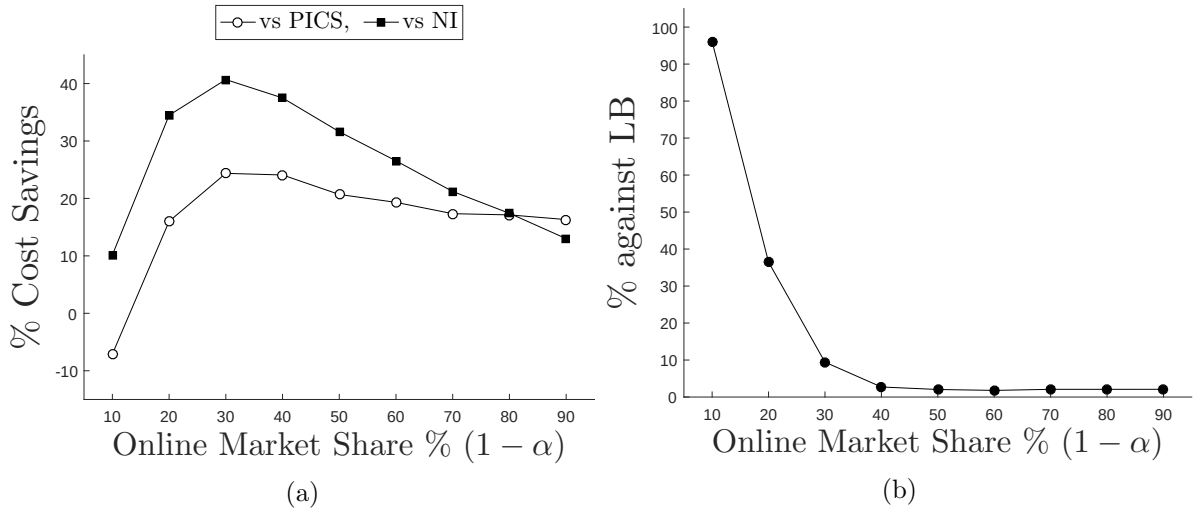


Figure 5 **Shows the effect of increasing online market share on cost savings achieved by heuristic against the PICS and NI models (left), as well as the performance of the heuristic against the lower bound (right)**

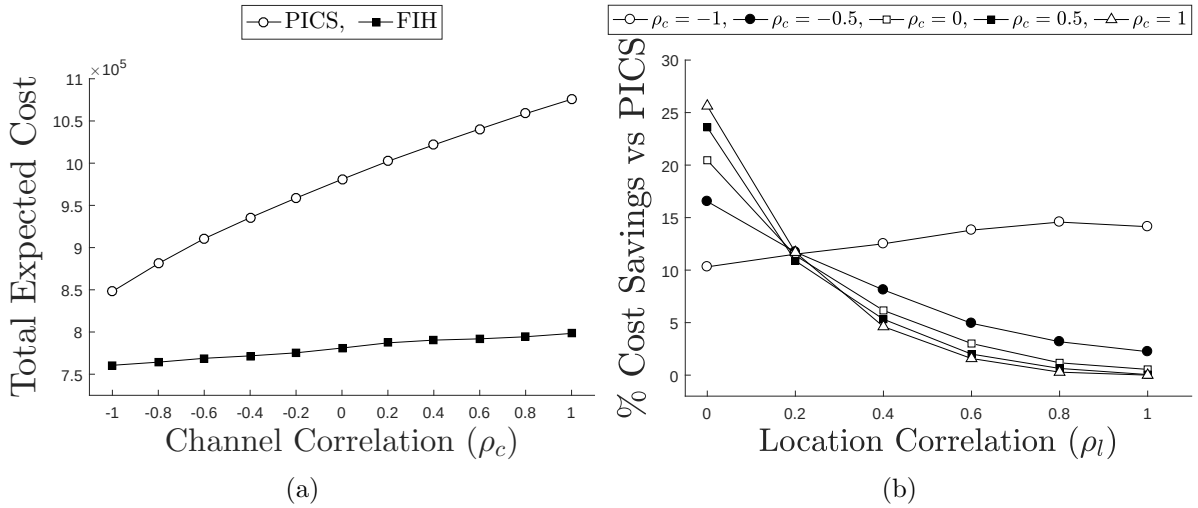


Figure 6 **Shows the effect of channel and location correlations on the heuristic and PICS models. (a) shows the effect of channel correlation on the expected cost, and (b) shows the effect of location correlation on % cost savings achieved by the heuristic for different values of the channel correlation coefficient.**

products, we observe that the cost savings increase steeply with $1-\alpha$ (Figure 5a). This is important from a practical standpoint, as e-commerce sales have been increasing at a rapid pace of 15% across product categories.

5.2.3. Effect of Correlation - between Channels and Locations. Correlation between in-store and online channels can be complex, as customer migration and fulfillment structure can influence the extent to which channels are correlated. Let ρ_c be the coefficient of correlation between in-store and online channels, a constant for all locations. We vary ρ_c

from -1 (perfect negative correlation) to +1 (perfect positive correlation) for a network of 50 physical stores and 2 OFCs, with $\alpha = 70\%$, and the results are shown in Figure 6a.

The NI model is unaffected by channel correlation, and is thus ignored. We observe that pooling benefits decrease when there is increased correlation, leading to an increase in expected costs. However, the heuristic is not affected as much as the PICS model, as it is additionally aided by pooling across locations. This naturally changes when there is *positive location correlation*, as seen in Figure 6b. Any pair of in-store demands or online demands from different locations have a coefficient of correlation ρ_l. The coefficient of correlation between the in-store demand at one location and online demand at the other location is given by ρ_{lc} calculated as the product of ρ_l and ρ_c.

The heuristic is aided by pooling across both channels and locations, and hence can make use of one type of pooling when there are correlation effects in the other. We also observe that correlation among locations has a stronger effect on the performance of the heuristic, as it undermines the value of virtual pooling — correlated demands lead to a lower imbalance in ending inventories across locations, thereby reducing the need for cross-shipping. However, other than extreme cases, we see that there is considerable cost savings achieved by the heuristic over the PICS model used in practice.

6. Extension to Multiple Periods

In this section, we discuss the extension of the FI model to the multi-period setting. The arguments provided in this section also hold for the other models discussed in this study. We have seen that the one-period setting is complex in itself, and naturally the extension to multiple periods can be done only for simplified cases, such as assuming lost sales for both in-store and online demands. The results obtained in this section are fairly straightforward for anyone versed in the multi-period inventory management literature.

The analyses closely follow the newsvendor network theory as discussed by Van Mieghem and Rudi (2002). We first consider the case where there is an initial level of inventory x_i at region i before ordering. For notational simplicity, we shall restrict our problem to two-stores. The cost function now becomes:

$$V^{FI}(x_1, x_2) = \min_{y_1 \geq x_1, y_2 \geq x_2} C^{FI}(y_1, y_2) \tag{20}$$

As (y_1^{FI}, y_2^{FI}) minimizes the function C^{FI}, for any $\{x = (x_1, x_2) : x_1 \leq y_1^{FI}, x_2 \leq y_2^{FI}\}$ it is optimal to order up to (y_1^{FI}, y_2^{FI}). We ignore cases where $x_i > y_i^{FI}$, as eventually the system is brought to the state $x \leq y^{FI}$. Hence, a base-stock policy is optimal for $x \leq y^{FI}$, and

$$V^{FI}(x_1, x_2) = C^{FI}(y_1^{FI}, y_2^{FI}) \tag{21}$$

We now consider the multiple period case, where we have T time periods: $1, 2, .., T$. Inventory replenishment is done at the beginning of each period, and the orders are received with zero leadtime. The in-store demands $\{D_{is}^t, t > 0\}$ and online demands $\{D_{io}^t, t > 0\}$ are assumed to be i.i.d. The demand shortages are modeled as lost sales, where unfulfilled demands are not carried over to future periods. The fulfillment sequence is as defined earlier, with in-store demand prioritized over online demand. The available inventory at the end of a period serves as the initial inventory for the next period, and we assume zero purchasing costs. Future costs are discounted using a factor $\delta \in (0, 1]$.

Proposition 6 *For the finite horizon problem with lost sales, a stationary base-stock policy is optimal, with order-up-to levels* (y_1^{FI}, y_2^{FI}).

The proof is relegated to the Appendix, and is similar to traditional multi-period inventory problems involving lost-sales. Thus, in the case of lost-sales for both in-store and online demand, the multi-period problem reduces to solving the one period problem to obtain the optimal order-up-to quantities. In practice, the choice of lost-sales for in-store demand is apt, but the online demand can usually be backlogged to be fulfilled at a later time. The problem becomes complicated once backlogging is introduced.

Consider the case where a stationary policy may be optimal with backlogging. The backorder in a period t needs to be mapped directly to a deterministic component added to the online demand for the next period, so that the order quantity for the next period becomes $y^{FI} - x + b$ where b is the vector of backordered online demand. The structure of our problem may inhibit the use of designated b units to fill the backlogs from the previous period, as it may be ex-post optimal to use these units to fulfill the in-store demand in the next period as it may lead to lower costs. The situation becomes more complex if purchasing costs are included, as it may not always be beneficial to cross-ship a unit of inventory at the end of a period, if the purchasing costs are non-identical across regions.

Thus, a stationary policy may no longer be optimal for the case where online demand is backordered. If we ensure that backorders are filled first in the following period, a stationary

base-stock policy may be optimal, and the order quantity in a period becomes $y^{FI} - x + b$. One way to model this is with variable cost for each unit of online demand backordered, which increases steeply after the first period so that it is always optimal to fill online backorders first. However, modeling complications arise with the introduction of such time-variant backorder costs. This is an interesting direction for future work, as one can also incorporate a 'delivery window' within which online orders have to be fulfilled.

7. Conclusion

Despite numerous retailers struggling with the operational problems posed by omnichannel retailing, the area has received comparatively less attention in literature. Our research addresses an important facet of omnichannel retailing — inventory management, by demonstrating the value of effective inventory policies in utilizing the pooling benefits offered by omnichannel retailing. We also develop a simple heuristic for the multi-store case which is asymptotically near optimal, and provides significant cost savings over current practices.

A plethora of promising directions for future research remain. In practice, a lot of firms fulfill store and online orders as they arrive even though a strong case can be made for prioritizing in-store demand. Some firms do not assign online orders to stores running low on inventory, while certain others have no such rule. It remains to be seen whether there is a dominant strategy for such fulfillment decisions.

With increasingly complex fulfillment practices, network design plays an important role in reducing costs and increasing efficiency. The choice of stores for store fulfillment can also contribute to balancing inventory levels, and is a topic which demands attention. Future research may also focus on classical extensions to the multi-period problem, such as leadtimes and capacities. However, one needs to be diligent in balancing the complexity induced by these extensions and the value of insights derivable from the emergent models.

It is obvious that there isn't one fulfillment strategy which fits all, as firms are constantly innovating to compete with others - Nordstrom provides free shipping to customers' homes in case of a store stockout when a customer walks in, and firms like Walmart and Target have multiple store fulfillment options including in-store pickup, curbside pickup, etc. It is unclear whether there exists a single model which can capture all of the complexity in omnichannel retailing, but we believe that our models provide a fair attempt at characterizing this complexity for a primitive study in omnichannel inventory management.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Number 1561791. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Acimovic, Jason, Stephen C Graves. 2015. Mitigating spillover in online retailing via replenishment. Working Paper.

Agatz, Niels AH, Moritz Fleischmann, Jo AEE Van Nunen. 2008. E-fulfillment and multi-channel distribution–a review. *Eur. J. Oper. Res.* **187**(2) 339–356.

Alptekinoğlu, Aydın, Christopher S Tang. 2005. A model for analyzing multi-channel distribution systems. *Eur. J. Oper. Res.* **163**(3) 802–824.

Ansari, Asim, Carl F Mela, Scott A Neslin. 2008. Customer channel migration. *J. of Marketing Res.* **45**(1) 60–76.

Barr, Alistair. 2013. Retail stores become shipping hubs to battle amazon. *USA Today.* Retrieved from: `http://www.usatoday.com/story/tech/2013/09/28/retailers-ship-from-store/2862405/`.

Bell, David, Santiago Gallino, Antonio Moreno. 2013. Inventory showrooms and customer migration in omni-channel retail: The effect of product information. Working Paper.

Bendoly, Elliot, Doug Blocher, Kurt M Bretthauer, MA Venkataramanan. 2007. Service and cost benefits through clicks-and-mortar integration: Implications for the centralization/decentralization debate. *Eur. J. Oper. Res.* **180**(1) 426–442.

Bensinger, Greg, Keiko Morris. 2014. Amazon to open first brick-and-mortar site. *The Wall Street Journal.* Retrieved from: `http://www.wsj.com/articles/amazon-to-open-first-store-1412879124`.

Bertrand, Louise P, James H Bookbinder. 1998. Stock redistribution in two-echelon logistics systems. *J. of the Oper. Res. Society* **49**(9) 966–975.

Bretthauer, Kurt M, Stephen Mahar, MA Venakataramanan. 2010. Inventory and distribution strategies for retail/e-tail organizations. *Computers & Industrial Engineering* **58**(1) 119–132.

Brynjolfsson, Erik, Yu Jeffrey Hu, Mohammad S Rahman. 2013. Competing in the age of omnichannel retailing. *MIT Sloan Management Review* **54**(4) 23.

Chiang, Wei-yu Kevin, George E Monahan. 2005. Managing inventories in a two-echelon dual-channel supply chain. *Eur. J. Oper. Res.* **162**(2) 325–341.

Chicago Consulting. 2013. Ten best warehouse networks. URL `http://www.chicago-consulting.com/ten-best-warehouse-networks/`.

Chopra, Sunil, Peter Meindl. 2007. Supply chain management. strategy, planning & operation. *Das Summa Summarum des Management*. Springer, 265–275.

Enright, Allison. 2015. Amazon's growth accelerates. *InternetRetailer.* Retrieved from: https://www.internetretailer.com/2015/12/23/amazons-growth-accelerates.

Forrester. 2014. Customer desires vs. retailer capabilities: Minding the omnichannel commerce gap. URL http://global.sap.com/asia/campaigns/2014-07/hybris-accenture-forrester-tlp-omni-channel.pdf.

FTI Consulting. 2015. The U.S. online retail forecast. URL http://www.fticonsulting.com/~/media/Files/us-files/insights/featured-perspectives/fti-2015onlineretailforecast.pdf.

Gallino, Santiago, Antonio Moreno. 2014. Integration of online and offline channels in retail: The impact of sharing reliable inventory availability information. *Management Sci.* **60**(6) 1434–1451.

Gallino, Santiago, Antonio Moreno, Ioannis Stamatopoulos. 2016. Channel integration, sales dispersion, and inventory management. *Management Sci.* Forthcoming.

Giannopoulos, Nicole. 2014. The macy's advantage: Secrets to same-day delivery, omnichannel. *RIS News.* Retrieved from: http://risnews.edgl.com/retail-news/The-Macy-s-Advantage--Secrets-to-Same-Day-Delivery,-Omnichannel97157.

Harrison, J Michael, Jan A Van Mieghem. 1999. Multi-resource investment strategies: Operational hedging under demand uncertainty. *Eur. J. Oper. Res.* **113**(1) 17–29.

Heyman, Daniel P, Matthew J Sobel. 2003. *Stochastic models in operations research: stochastic optimization*, vol. 2. Courier Corporation.

Hoeffding, Wassily. 1963. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association* **58**(301) 13–30.

Jalilipour Alishah, Elnaz, Kamran Moinzadeh, Yong-Pin Zhou. 2015. Inventory fulfillment strategies for an omni-channel retailer. *Manufacturing Service Oper. Management* Under review.

Jasin, Stefanus, Amitabh Sinha. 2015. An lp-based correlated rounding scheme for multi-item ecommerce order fulfillment. *Oper. Res.* **63**(6) 1336–1351.

Karmarkar, Uday S, Nitin R Patel. 1977. The one-period, n-location distribution problem. *Naval Res. Logist. Quarterly* **24**(4) 559–575.

Kurt Salmon. 2016. Building a solid omnichannel foundation with effective inventory management. URL http://www.kurtsalmon.com/en-us/Retail/vertical-insight/1478/Building-a-Solid-Omnichannel-Foundation-with-Effective-Inventory-Management.

Leiser, Jordy. 2016. Think tank: Why an omnichannel approach can help retail escape the amazon. *WWD.* Retrieved from: http://wwd.com/retail-news/forecasts-analysis/blazing-the-amazon-and-why-an-omnichannel-approach-can-help-retail-10312172/.

Lindner, Matt. 2016. Target now ships online orders from more than 1,000 stores. *InternetRetailer.* Retrieved from: https://www.internetretailer.com/2016/11/16/target-now-ships-online-orders-more-1000-stores.

Mahar, Stephen, Kurt M Bretthauer, MA Venkataramanan. 2009. The value of virtual pooling in dual sales channel supply chains. *Eur. J. Oper. Res.* **192**(2) 561–575.

Mahar, Stephen, P Daniel Wright. 2009. The value of postponing online fulfillment decisions in multi-channel retail/e-tail organizations. *Computers & Oper. Res.* **36**(11) 3061–3072.

Nash, Kim. 2015. Walmart build supply chain to meet e-commerce demands. *The Wall Street Journal.* Retrieved from: `http://www.wsj.com/articles/wal-mart-builds-supply-chain-to-meet-e-commerce-demands-1431016708`.

Netessine, Serguei, Nils Rudi. 2004. Supply chain structures on the internet. *Handbook of Quantitative Supply Chain Analysis*. Springer, 607–641.

Netessine, Serguei, Nils Rudi. 2006. Supply chain choice on the internet. *Management Sci.* **52**(6) 844–864.

Rigby, Darrell. 2011. The future of shopping. *Harvard Business Review* **89**(12) 65–76.

Robinson, Lawrence W. 1990. Optimal and approximate policies in multiperiod, multilocation inventory models with transshipments. *Oper. Res.* **38**(2) 278–295.

Seifert, Ralf W, Ulrich W Thonemann, Marcel A Sieke. 2006. Relaxing channel separation: Integrating a virtual store into the supply chain via transshipments. *IIE Trans.* **38**(11) 917–931.

Şen, Alper, Alex X Zhang. 1999. The newsboy problem with multiple demand classes. *IIE Trans.* **31**(5) 431–444.

Tagaras, George. 1989. Effects of pooling on the optimization and service levels of two-location inventory systems. *IIE Trans.* **21**(3) 250–257.

Tagaras, George, Morris A Cohen. 1992. Pooling in two-location inventory systems with non-negligible replenishment lead times. *Management Sci.* **38**(8) 1067–1083.

UPS Compass. 2014. Ship from store: A smart competitive strategy for retailers. URL `https://compass.ups.com/ship-from-store-benefits/`.

U.S. Census Bureau. 2016. Quarterly retail e-commerce sales. Retrieved from: `https://www.census.gov/retail/mrts/www/data/pdf/ec_current.pdf`.

Van Mieghem, Jan A. 2003. Commissioned paper: Capacity management, investment, and hedging: Review and recent developments. *Manufacturing Service Oper. Management* **5**(4) 269–302.

Van Mieghem, Jan A, Nils Rudi. 2002. Newsvendor networks: Inventory management and capacity investment with discretionary activities. *Manufacturing Service Oper. Management* **4**(4) 313–335.

Wikipedia. 2016. List of united states cities by population. URL `https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population`. [Online; accessed 12-December-2016].

Xu, Ping Josephine, Russell Allgor, Stephen C Graves. 2009. Benefits of reevaluating real-time order fulfillment decisions. *Manufacturing Service Oper. Management* **11**(2) 340–355.

Yang, Jian, Zhaoqiong Qin. 2007. Capacitated production control with virtual lateral transshipments. *Oper. Res.* **55**(6) 1104–1119.

Zaroban, Stefany. 2016. U.S. e-commerce grows 14.6% in 2015. *InternetRetailer*. Retrieved from: `https://www.internetretailer.com/2016/02/17/us-e-commerce-grows-146-2015`.

Appendices

Appendix A: Additional Details for Numerical Analyses

All numerical analyses were done on a desktop computer (i7-3770 CPU @3.7GHz, 16GB RAM). For all numerical studies, we run Monte-Carlo simulations using 15000 random samples, and use the sample average as an approximation of the expected costs for each model.

A.1. Two-Store Setting

Given M_1, M_2 are the random market sizes, and α_1, α_2 are the fraction of total demand which occurs in-store, the in-store demands are thus $D_{is} = \alpha_i M_i$, and the online demands are $D_{io} = (1-\alpha_i)M_i$. With this alternate demand formulation, the optimal order-up-to quantities for the NI and PI models are re-calculated using the optimality equations:

$$y_1^{NI} = \alpha_1.F_1^{-1}\Big(\frac{p_s}{h+p_s}\Big)\ ,\ y_2^{NI} = \alpha_2.F_2^{-1}\Big(\frac{p_s}{h+p_s}\Big)\ ,\ y_o^{NI} = F_o^{-1}\Big(\frac{p_o-s_o}{h+p_o-s_o}\Big) \tag{22}$$

$$(h+p_o-s)\,F_i\left(y_i^{PI}\right) + (p_s-(p_o-s))\,F_i\left(\frac{y_i^{PI}}{\alpha_i}\right) = p_s, \qquad \forall i \tag{23}$$

where F_1, F_2 are the cumulative distribution functions of the market sizes M_1, M_2 respectively, and F_o is the cumulative distribution function of the total online demand $D_o = (1-\alpha_1)M_1 + (1-\alpha_2)M_2$. As M_1 and M_2 are normally distributed and independent, D_o is normally distributed with mean $\mu_o = (1-\alpha_1)\mu_1 + (1-\alpha_2)\mu_2$ and standard deviation $\sqrt{(1-\alpha_1)^2\sigma_1^2 + (1-\alpha_2)^2\sigma_2^2}$.

For the FI model, we employ the gradient descent method to obtain the optimal order-up-to quantities using the formula in Equation 8. The demand regions in which the dual-prices are constant are described in Appendix B, including the method to calculate the probabilities of these regions in the demand space.

A.2. Multi-Store Setting

The total market is assumed to be the top 300 most populous cities in mainland US. The means of the in-store and online demands at physical stores are calculated as follows: the mean in-store demand of a city is taken to be α times mean market size, whereas the proportion of the mean online demand of a city is $(1-\alpha)$. The mean market size of a city is in turn taken to be a fixed proportion of the population in the city. The coefficient of variation of all demands are taken to be 0.3. The random demands are generated from a random multivariate normal vector with a correlation matrix according to the coefficients of channel and location correlations. When the demands are independent, all correlation coefficients are zero.

The OFCs are initially designated to fulfill all online orders other than those from cities at which physical stores are present, and the mean total online demand for the OFCs is estimated in a similar fashion to that of physical stores, from the population not covered by the physical stores. The rationing of online demand for each OFC is done based on the optimal throughput rates estimated by Chicago Consulting (2013). The

online demand in cities with stores which cannot handle online orders ($\mathcal{S}_s$) are also allocated to the OFCs. In the NI model, the online demands at the physical stores are allocated to the nearest OFC, and there is no store fulfillment.

The increase in number of physical stores discussed in Section 5.2.1 corresponds to the situation where the firm opens new stores in the cities which do not have physical stores. The online demand arising from these cities are now allocated to the physical stores, as they were previously fulfilled from an OFC.

Appendix B: Demand Regions for the FI Model

We illustrate the identification of demand regions in which the dual vector λ is constant (as discussed in Section 3.1.3) and the calculation of the corresponding probabilities with the help of a simplified case where the in-store and online demands within a region are perfectly correlated — we will use the (M, α) demand formulation as discussed in the two-store numerical study in Section 5.1. Using these probabilities, the gradient of the cost function of the FI model can be calculated according to Equation 8, and can be used as input to the gradient descent algorithm to obtain the optimal order-up-to quantities.

For any given (y_1, y_2), the demand space (M_1, M_2) can be divided into a number of independent regions. Based on the values taken by the variables in the optimal solution in (7), Table 1 shows the different cases that are possible for the order-up-to levels y_1 and y_2. From these cases, the independent demand regions are listed in Table 2 along with the constant dual prices in those regions. We use identical costs across regions, similar to the notation used in Section 3.2. The underlined cases are redundant, and can be discarded while calculating the probability for each region. The dual prices λ_1, λ_2 are the shadow prices of the constraints

Table 1 Table showing the various demand cases based on the values of y_1, y_2

	A	B	C	D
1	$y_1 < \alpha_1 M_1$	$\alpha_1 M_1 \le y_1 < M_1$	$M_1 \le y_1 < M_1 + (1-\alpha_2)M_2$	$y_1 \ge M_1 + (1-\alpha_2)M_2$
2	$y_2 < \alpha_2 M_2$	$\alpha_2 M_2 \le y_2 < M_2$	$M_2 \le y_2 < M_2 + (1-\alpha_1)M_1$	$y_2 \ge M_2 + (1-\alpha_1)M_1$
3	$y_1 + y_2 < M_1 + M_2$	$y_1 + y_2 \ge M_1 + M_2$		

which contain y_1 and y_2 respectively, namely the first set of constraints $z_i + \sum_{j=1}^{2} z_{ij} \le y_i, \forall i$ in the linear program in (7). For example, for the demand regions with the case D1, that is, $y_1 \ge M_1 + (1-\alpha_2)M_2$, irrespective of the value of y_2, there will be inventory left over at retail store 1 at the end of the period. Thus the constraint $z_1 + \sum_{j=1}^{2} z_{1j} \le y_1$ will not bind, and hence the dual price $\lambda_1 = 0$. There are a total of 20 valid demand regions, as shown in Figure 7.

Table 2 Table showing the various demand regions and the corresponding constant dual-prices.

Region	Case	λ_1	λ_2	Region	Case	λ_1	λ_2
Ω_1	A1,A2,A3	$h+p_s$	$h+p_s$	Ω_{11}	C1,A2,A3	$h+p_o-s'$	$h+p_s$
Ω_2	A1,B2,A3	$h+p_s$	$h+p_o-s'$	Ω_{12}	C1,B2,A3	$h+p_o-s'$	$h+p_o-s$
Ω_3	A1,C2,A3	$h+p_s$	$h+p_o-s'$	Ω_{13}	C1,B2,B3	0	$s'-s$
Ω_4	A1,D2,A3	$h+p_s$	0	Ω_{14}	C1,C2,B3	0	0
Ω_5	A1,D2,B3	$h+p_s$	0	Ω_{15}	C1,D2,B3	0	0
Ω_6	B1,A2,A3	$h+p_o-s$	$h+p_s$	Ω_{16}	D1,A2,A3	0	$h+p_s$
Ω_7	B1,B2,A3	$h+p_o-s$	$h+p_o-s$	Ω_{17}	D1,A2,B3	0	$h+p_s$
Ω_8	B1,C2,A3	$h+p_o-s$	$h+p_o-s'$	Ω_{18}	D1,B2,B3	0	$s'-s$
Ω_9	B1,C2,B3	$s'-s$	0	Ω_{19}	D1,C2,B3	0	0
Ω_{10}	B1,D2,B3	$s'-s$	0	Ω_{20}	D1,D2,B3	0	0

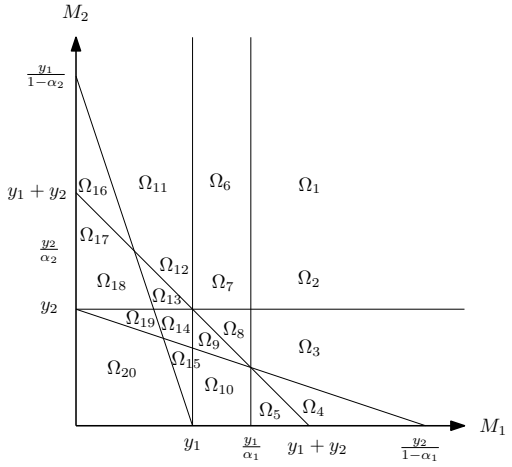


Figure 7 Demand regions with constant dual-price vector

The probability for each region is calculated based on the values which the corresponding random variables take. A few example calculations are shown:

$$\mathbb{P}(\Omega_1) = \mathbb{P}(A1, A2) = \mathbb{P}(y_1 < \alpha_1 M_1, y_2 < \alpha_2 M_2) = \left(1 - F_1\left(\frac{y_1}{\alpha_1}\right)\right)\left(1 - F_2\left(\frac{y_2}{\alpha_2}\right)\right)$$

$$\begin{aligned}\mathbb{P}(\Omega_3) &= \mathbb{P}(A1, C2) = \mathbb{P}\left(y_1 < \alpha_1 M_1, M_2 \le y_2 < M_2 + (1-\alpha_1) M_1\right) \\ &= \mathbb{E}\left(\mathbb{P}\left(y_1 < \alpha_1 M_1, M_2 \le y_2 < M_2 + (1-\alpha_1) M_1\right) | M_1\right) \\ &= \int_0^\infty \mathbb{P}\left(M_2 \le y_2 < M_2 + (1-\alpha_1) m\right) \mathbb{1}\left(m > \frac{y_1}{\alpha_1}\right) f_1(m)\, dm \\ &= F_2(y_2)\left(1 - F_1\left(\frac{y_1}{\alpha_1}\right)\right) - \int_{\frac{y_1}{\alpha_1}}^{\frac{y_2}{1-\alpha_2}} F_2\left(y_2 - (1-\alpha_1) m\right) f_1(m)\, dm\end{aligned}$$

where f_1 denotes the probability density function of the random variable M_1. Probabilities involving both M_1 and M_2 have to be calculated using conditional expectations as shown above. These expressions can be evaluated quickly for normal distributions using a computational software like MATLAB, and calculations of probabilities for demand regions with zero dual-prices can be skipped. These calculations become computationally intensive when more random variables are involved. Note that for the case where $s' > h + p_o$, the dual-prices need to be recalculated, as cross-shipping ceases to occur in this case.

Appendix C: Proofs of Propositions

C.1. Proof of Proposition 3

From Equation 14, we have

$$\begin{aligned} C^{FI}(y_1^{FI}, y_2^{FI}) &= C^{PI}(y_1^{FI}, y_2^{FI}) + (s' - h - p_o)\mathbb{E}\left(\mathrm{CS}\left(y_1^{FI}, y_2^{FI}, \bar{D}\right)\right) \\ &\leq C^{PICS} = C^{PI}(y_1^{PI}, y_2^{PI}) + (s' - h - p_o)\mathbb{E}\left(\mathrm{CS}\left(y_1^{PI}, y_2^{PI}, \bar{D}\right)\right) \end{aligned}$$

As $C^{PI}(y_1^{PI}, y_2^{PI}) \leq C^{PI}(y_1^{FI}, y_2^{FI})$ (y_1^{PI},y_2^{PI} are optimal under C^{PI}), and $s' - h - p_o < 0$, it follows that $\mathbb{E}\left(\mathrm{CS}\left(y_1^{FI}, y_2^{FI}, \bar{D}\right)\right) \geq \mathbb{E}\left(\mathrm{CS}\left(y_1^{PI}, y_2^{PI}, \bar{D}\right)\right)$

C.2. Proof of Proposition 4

First, we define the cost function of the FI model in the case where all the cross-shipping costs s_{ij} are lowered to the within-region shipping cost s as follows:

$$\begin{aligned} C^{LB'}(y_1, \ldots, y_N) = s\sum_i \mu_{io} + \mathbb{E}h\left(\sum_i (y_i - D_{is})^+ - \sum_i D_{io}\right)^+ + \mathbb{E}p_s \sum_i (D_{is} - y_i)^+ \\ + (p_o - s)\,\mathbb{E}\left(\sum_i D_{io} - \sum_i (y_i - D_{is})^+\right)^+ \end{aligned}$$

The above cost function can be understood as follows: after the in-store demand has been fulfilled at each store, the total supply available in the system is $\sum_i (y_i - D_{is})^+$ to fulfill the total online demand $\sum_i D_{io}$. The cost function follows a similar structure to the newsvendor cost function with this supply and demand, including an underage cost for the in-store demand.

From Equation 16, we have

$$\begin{aligned} C^{LB'}(y_1, \ldots, y_N) - C^{LB}(y_1, \ldots, y_N) = h\mathbb{E}\left[\left(\sum_i (y_i - D_{is})^+ - \sum_i D_{io}\right)^+ - \left(\sum_i y_i - D\right)^+\right] \\ + (p_o - s)\,\mathbb{E}\left[\left(\sum_i D_{io} - \sum_i (y_i - D_{is})^+\right)^+ + \sum_i (D_{is} - y_i)^+ - \left(D - \sum_i y_i\right)^+\right] \end{aligned} \tag{24}$$

As $\left(\sum_i (y_i - D_{is})^+ - \sum_i D_{io}\right)^+ \geq \left(\sum_i (y_i - D_{is}) - \sum_i D_{io}\right)^+ = \left(\sum_i y_i - D\right)^+$, the first term is non-negative. For the second term, we have:

$$\begin{aligned} &\left(\sum_i D_{io} - \sum_i (y_i - D_{is})^+\right)^+ + \sum_i (D_{is} - y_i)^+ - \left(D - \sum_i y_i\right)^+ \\ &\qquad \geq \left(\sum_i D_{io} - \sum_i (y_i - D_{is})^+ + \sum_i (D_{is} - y_i)^+\right)^+ - \left(D - \sum_i y_i\right)^+ \\ &\qquad = \left(\sum_i D_{io} + \sum_i (D_{is} - y_i)\right)^+ - \left(D - \sum_i y_i\right)^+ \\ &\qquad = 0 \end{aligned}$$

The second term is also non-negative, and hence, $C^{LB'}(y_1,\ldots,y_N) \geq C^{LB}(y_1,\ldots,y_N)$. Since $C^{LB'}(y_1,\ldots,y_N) \leq C^{FI}(y_1,\ldots,y_N)$ as $s_{ij} \geq s, \forall i,j$, we have $C^{LB}(y_1,\ldots,y_N) \leq C^{FI}(y_1,\ldots,y_N)$.

C.3. Proof of Proposition 5

Consider a square of unit area in which N stores are uniformly distributed. Let the square be divided into $\sqrt{N}$ identical cells, such that each cell contains $\sqrt{N}$ stores. The dimensions of each cell are thus $\frac{1}{N^{\frac{1}{4}}} \times \frac{1}{N^{\frac{1}{4}}}$. The superscript l for a demand variable denotes that the demand belongs to a store in cell l. For example, D^l_{is} represents the in-store demand of the i^{th} store in cell l, whereas D_{is} represents the i^{th} store in the overall system. Similar to earlier definitions, $D^l_i = D^l_{is} + D^l_{io}$.

As defined in the proof of Proposition 4, let LB' be the model obtained from the FI model by lowering all cross-shipping costs to the within-region shipping cost s. Let FI_c and LB'_c be the models obtained by restricting the FI and LB' models respectively, so that cross-shipments can only be made between two stores belonging to the same cell. Clearly, $C^{FI}(y) \leq C^{FI_c}(y)$ and $C^{LB'}(y) \leq C^{LB'_c}(y)$ for any $y \geq 0$, as allowing cross-shipments across different cells leads to reduced costs. Let $\Phi(y,N)$ denote the cost incurred by N stores ordering up-to y each, without the option of cross-shipping:

$$\Phi(y,N) = \sum_{i=1}^{N}\left[h\left(y-D_i\right)^+ + p_s\left(D_{is}-y\right)^+ + p_o\left(D_{io}-\left(y-D_{is}\right)^+\right)^+ + s\min\left(D_{io},\left(y-D_{is}\right)^+\right)\right]$$

Note that $\Phi(y,N)$ represents the sum of costs incurred by individual stores, and hence, $\mathbb{E}\Phi(y,N) = \mathbb{E}\sum_{l=1}^{\sqrt{N}}\Phi(y,\sqrt{N}) = \sqrt{N}\Phi(y,\sqrt{N})$. Let $CS_{ij}(y,N)$ denote the cross-shipped quantity between stores i and j, when there are N stores with order-up-to quantity y each (the notation is CS^l_{ij} when defined within a cell). Note that both the functions Φ and CS_{ij} also depend on the demand vector, but the dependency is ignored for notational convenience. As the cells are identical in terms of demands and costs, we can represent the costs C^{FI_c} and $C^{LB'_c}$ as follows:

$$\begin{aligned}
C^{FI_c}(y^{FIH}) = &\quad \mathbb{E}\left(\sum_{l=1}^{\sqrt{N}}\left[\Phi(y^{FIH},\sqrt{N}) + \sum_{i=1}^{\sqrt{N}}\sum_{j=1,j\neq i}^{\sqrt{N}}(s^l_{ij}-h-p_o)CS^l_{ij}(y^{FIH},\sqrt{N})\right]\right)\\
= &\quad \mathbb{E}\Phi(y^{FIH},N) + \mathbb{E}\left(\sum_{l=1}^{\sqrt{N}}\left(\sum_{i=1}^{\sqrt{N}}\sum_{j=1,j\neq i}^{\sqrt{N}}(s^l_{ij}-h-p_o)CS^l_{ij}(y^{FIH},\sqrt{N})\right)\right)\\
C^{LB'}(y^{FIH}) = &\quad C^{LB'_c}(y^{FIH})\\
&+(s-h-p_o)\mathbb{E}\left[\sum_{l=1}^{\sqrt{N}}\left(\sum_{i=1}^{\sqrt{N}}D^l_{io}-\left(y^{FIH}-D^l_{is}\right)^+\right)^+ - \left(\sum_{i=1}^{N}D_{io}-\left(y^{FIH}-D_{is}\right)^+\right)^+\right]\\
= &\quad \mathbb{E}\Phi(y^{FIH},N) + \mathbb{E}\left(\sum_{l=1}^{\sqrt{N}}\left(\sum_{i=1}^{\sqrt{N}}\sum_{j=1,j\neq i}^{\sqrt{N}}(s-h-p_o)CS^l_{ij}(y^{FIH},\sqrt{N})\right)\right)\\
&+(s-h-p_o)\left[\sqrt{N}\mathbb{E}\left(\sum_{i=1}^{\sqrt{N}}D^l_{io}-\left(y^{FIH}-D^l_{is}\right)^+\right)^+ - \mathbb{E}\left(\sum_{i=1}^{N}D_{io}-\left(y^{FIH}-D_{is}\right)^+\right)^+\right]
\end{aligned}$$

The expression for $C^{LB'}$ is written as the sum of the cost of the LB'_c model which restricts cross-shipping to within each cell, and the cost of the additional cross-shipped units with this restriction removed. We know

that $C^{LB}(y^{FIH}) \leq C^{LB'}(y^{FIH}) \leq C^{FI}(y^{FIH}) \leq C^{FI_c}(y)$, where the first inequality follows from Proposition 4. We first show that $\frac{C^{FI_c}(y^{FIH})}{C^{LB'}(y^{FIH})} \to 1$ as $N \to \infty$. We have:

$$\frac{C^{FI_c}(y^{FIH})}{C^{LB'}(y^{FIH})} - 1 = \frac{\mathbb{E}\left(\sum_{l=1}^{\sqrt{N}}\left(\sum_{i=1}^{\sqrt{N}}\sum_{j=1,j\neq i}^{\sqrt{N}}(s_{ij}^l - s)CS_{ij}^l(y^{FIH},\sqrt{N})\right)\right)}{C^{LB'}(y^{FIH})} + \frac{(h+p_o-s)\left[\sqrt{N}\mathbb{E}\left(\sum_{i=1}^{\sqrt{N}} D_{io}^l - (y^{FIH} - D_{is}^l)^+\right)^+ - \mathbb{E}\left(\sum_{i=1}^{N} D_{io} - (y^{FIH} - D_{is})^+\right)^+\right]}{C^{LB'}(y^{FIH})}$$

We have $s_{ij}^l - s = f(d_{ij}^l) \leq f\left(\frac{\sqrt{2}}{N^{\frac{1}{4}}}\right)$, as the maximum distance within a cell is $\frac{\sqrt{2}}{N^{\frac{1}{4}}}$. Thus, using $C^{LB'}(y^{FIH}) \geq \mathbb{E}\left(\sum_{l=1}^{\sqrt{N}}\left(\sum_{i=1}^{\sqrt{N}}\sum_{j=1,j\neq i}^{\sqrt{N}}(s)CS_{ij}^l(y^{FIH},\sqrt{N})\right)\right)$ for the first term, and $C^{LB'}(y^{FIH}) \geq s\mu_o N$ for the second term, we have

$$\frac{C^{FI_c}(y^{FIH})}{C^{LB'}(y^{FIH})} - 1 \leq \frac{f\left(\frac{\sqrt{2}}{N^{\frac{1}{4}}}\right)}{s} + \left(\frac{h+p_o-s}{s\mu_o\sqrt{N}}\right)\mathbb{E}\left(\sum_{i=1}^{\sqrt{N}} D_{io} - \left(y^{FIH} - D_{is}\right)^+\right)^+ \tag{25}$$

The first term on the right hand side vanishes to zero as $N \to \infty$, as $f(d) \to 0$ as $d \to 0$. To simplify the second term, we need the following lemmas.

LEMMA 1. *When $h < p_o - s$, $y^{FIH} > \mu$ where $\mu = \mu_s + \mu_o$, and if additionally $h < p_s - (p_o - s)$,*

$$y^{FIH} \to F_s^{-1}\left(\frac{p_s - p_o + s - h}{p_s - p_o + s}\right) \in (0,\infty), \text{ as } N \to \infty \tag{26}$$

Lemma 1 can be proved by observing the optimality equations of the LB model (Equation 17) for the case where stores are identical, which is as follows:

$$(h+p_o-s)\mathbb{P}(\sum_{i=1}^{N} D_i \leq N y^{FIH}) + (p_s - p_o + s)F_{D_{1s}}(y^{FIH}) = p_s$$

From the above equation, when $h < p_o - s$, we have $p_s < 2\,(p_o - s)\,\mathbb{P}\left(\sum_{i=1}^{N} D_i \leq N y^{FIH}\right) + (p_s - p_o + s)$. This simplifies to yield $y^{FIH} > \mu$. Now, by applying the central limit theorem as $N \to \infty$, $\mathbb{P}\left(\frac{\sum_{i=1}^{N} D_i}{N} \leq y^{FIH}\right) \to 1$ when $y^{FIH} > \mu$, and the result follows.

LEMMA 2. *When $h < p_o - s$ and $h < p_s - (p_o - s)$, and the demands are bounded above as $D_{is} \leq M_s$ and $D_{io} \leq M_o$ for all i,*

$$\mathbb{P}\left(\sum_{i=1}^{\sqrt{N}} D_{io} > \sum_{i=1}^{\sqrt{N}}\left(y^{FIH} - D_{is}\right)^+\right) \leq \exp\left\{\frac{-2\sqrt{N}(y^{FIH} - \mu)^2}{M_o + M_s}\right\} \tag{27}$$

and hence, $\mathbb{P}\left(\sum_{i=1}^{\sqrt{N}} D_{io} > \sum_{i=1}^{\sqrt{N}}\left(y^{FIH} - D_{is}\right)^+\right) \to 0$, as $N \to \infty$.

The proof is as follows:

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^{\sqrt{N}} D_{io} > \sum_{i=1}^{\sqrt{N}} \left(y^{FIH} - D_{is}\right)^+\right) &= \mathbb{P}\left(\sum_{i=1}^{\sqrt{N}} \left(D_i - \left(D_{is} - y^{FIH}\right)^+\right) > \sqrt{N} y^{FIH}\right) \\ &\leq \mathbb{P}\left(\sum_{i=1}^{\sqrt{N}} D_i > \sqrt{N} y^{FIH}\right) \\ &\leq \exp\left\{\frac{-2\sqrt{N}(y^{FIH}-\mu)^2}{M_o + M_s}\right\} \to 0, \text{ as } N \to \infty\end{aligned}$$

The final inequality follows from the Hoeffding bound for tail probabilities Hoeffding (1963), as $y^{FIH} > \mu$ and demands are bounded, and the limit exists as y^{FIH} approaches a finite positive quantity as $N \to \infty$ by Lemma 1.

The expectation in the second term of Equation 25 can be bounded as follows:

$$\begin{aligned}&\mathbb{E}\left(\sum_{i=1}^{\sqrt{N}} \left(D_{io} - \left(y^{FIH} - D_{is}\right)^+\right)\right)^+ \\ &= \mathbb{E}\left[\left(\sum_{i=1}^{\sqrt{N}} \left(D_{io} - \left(y^{FIH} - D_{is}\right)^+\right)\right)^+ \middle| \sum_{i=1}^{\sqrt{N}} D_{io} > \sum_{i=1}^{\sqrt{N}} \left(y^{FIH} - D_{is}\right)^+\right] \mathbb{P}\left(\sum_{i=1}^{\sqrt{N}} D_{io} > \sum_{i=1}^{\sqrt{N}} \left(y^{FIH} - D_{is}\right)^+\right) \\ &\leq \mathbb{E}\left[\sum_{i=1}^{\sqrt{N}} D_{io} \middle| \sum_{i=1}^{\sqrt{N}} D_{io} > \sum_{i=1}^{\sqrt{N}} \left(y^{FIH} - D_{is}\right)^+\right] \mathbb{P}\left(\sum_{i=1}^{\sqrt{N}} D_{io} > \sum_{i=1}^{\sqrt{N}} \left(y^{FIH} - D_{is}\right)^+\right) \\ &\leq M_o \sqrt{N} \exp\left\{\frac{-2\sqrt{N}(y^{FIH}-\mu)^2}{M_o + M_s}\right\}\end{aligned}$$

The last inequality follows from Lemma 2 and the boundedness of the demands as $D_{is} \leq M_s$, and $D_{io} \leq M_o$ for all i with $0 < M_s, M_o < \infty$. Thus, we have:

$$\begin{aligned}\frac{C^{FI_c}(y^{FIH})}{C^{LB'}(y^{FIH})} &\leq 1 + \frac{f\left(\frac{\sqrt{2}}{N^{\frac{1}{4}}}\right)}{s} + \left(\frac{h + p_o - s}{s\mu_o}\right)\left(M_o\sqrt{N}\exp\left\{\frac{-2\sqrt{N}(y^{FIH}-\mu)^2}{M_o+M_s}\right\}\right) \\ &\to 1, \text{ as } N \to \infty\end{aligned} \tag{28}$$

The next step is to show the the LB cost function is off by a constant factor from the cost function of LB'. From Equation 24, the two terms within the square parantheses can be shown to be equal, and hence we have:

$$\begin{aligned}&C^{LB'}(y^{FIH}) - C^{LB}(y^{FIH}) \\ &\qquad = (h + p_o - s)\,\mathbb{E}\left[\left(\sum_{i=1}^{N} D_{io} - \left(y^{FIH} - D_{is}\right)^+\right)^+ + \sum_{i=1}^{N}\left(D_{is} - y^{FIH}\right)^+ - \left(D - \sum_{i=1}^{N} y^{FIH}\right)^+\right]\end{aligned}$$

where $D = \sum_{i=1}^{N} D_{is} + D_{io}$.

Similar to what was done to bound the second term in Equation 25, we can show that whenever the conditions in Lemma 2 are satisfied, $\mathbb{E}\left(\sum_{i=1}^{N} D_{io} - (y^{FIH} - D_{is})^+\right)^+ \leq M_o N \exp\left\{\frac{-2N(y^{FIH}-\mu)^2}{M_o+M_s}\right\}$. Thus, we have:

$$C^{LB'}(y^{FIH}) - C^{LB}(y^{FIH}) \leq (h + p_o - s)\left[M_o N \exp\left\{\frac{-2N(y^{FIH}-\mu)^2}{M_o+M_s}\right\} + \sum_{i=1}^{N}\left(D_{is} - y^{FIH}\right)^+\right]$$

Using $C^{LB}(y^{FIH}) \geq s\mu_o N$ and $C^{LB}(y^{FIH}) \geq (p_s - p_o + s)\sum_{i=1}^{N}(D_{is} - y^{FIH})^+$, we have:

$$\frac{C^{LB'}(y^{FIH})}{C^{LB}(y^{FIH})} - 1 \leq \left(\frac{h + p_o - s}{s\mu_o}\right)\left(M_o \exp\left\{\frac{-2N(y^{FIH} - \mu)^2}{M_o + M_s}\right\}\right) + \left(\frac{h + p_o - s}{p_s - p_o + s}\right) \tag{29}$$

Thus, from Equations 28 and 29, as $N \to \infty$, we have

$$\begin{aligned} &\frac{C^{FI_c}(y^{FIH})}{C^{LB}(y^{FIH})} \leq 1 + \frac{h + p_o - s}{p_s - p_o + s} \\ \Rightarrow &\frac{C^{FIH}}{C^{LB}(y^{FIH})} \leq \frac{h + p_s}{p_s - p_o + s} \end{aligned}$$

The final step follows from $C^{FI_c}(y^{FIH}) \geq C^{FI}(y^{FIH}) = C^{FIH}$. Thus, our heuristic achieves a constant approximation factor of $\frac{h+p_s}{p_s-p_o+s}$ with respect to the lower bound. Although the proof has been formulated for a simplified setting with stores distributed uniformly within a unit square, the result may still hold when subject to a few generalizations. The unit square can be replaced with any finite area, and the cells need not be identical, as long as the number of stores in each cell grows to infinity as $N \to \infty$. The resulting case may call for a more complicated proof, and is outside the scope of this study.

C.4. Proof of Proposition 6

The proof is by induction, and similar to the proof of Proposition 4 in Van Mieghem and Rudi (2002). Let $V_t^{FI}(x^t)$ be the expected cost-to-go function evaluated at period t, with the initial inventory $x^t = (x_1^t, x_2^t)$. If we show that V_t^{FI} is convex and affine in the initial inventory x, a stationary base stock policy would be optimal. For the $T+1^{th}$ period, the optimal cost function is $V_{T+1}^{FI}(x^{T+1}) = 0$ (assuming zero purchasing costs) which is trivially convex and affine in x^{T+1}. Let V_{t+1}^{FI} be convex and affine in x^{t+1}. The cost function for period t can be written as:

$$\begin{aligned} V_t^{FI}(x_1, x_2) &= \min_{y_1 \geq x_1, y_2 \geq x_2} \left[C^{FI}(y_1, y_2) + \delta \mathbb{E} V_{t+1}^{FI}\left(f_1(y_1, y_2, D), f_2(y_1, y_2, D)\right)\right] \\ &= \min_{y_1 \geq x_1, y_2 \geq x_2} U_t^{FI}(y_1, y_2) \end{aligned} \tag{30}$$

where $f_1(y_1, y_2, D), f_2(y_1, y_2, D)$ are the ending inventories at regions 1 and 2 respectively. D is the demand vector constituting the in-store and online demands for both the regions. As taking expectation preserves convexity, and the sum of convex functions is convex, $U_t^{FI}(y_1, y_2)$ is jointly convex in y_1, y_2. It only remains to be shown that V_t^{FI} is affine in x. To show this, consider any $y \leq y^{FI}$, so that $\left(f_1(y_1, y_2, D), f_2(y_1, y_2, D)\right)^T \leq y \leq y^{FI}$. We have

$$\begin{aligned} U_t^{FI}(y_1, y_2) &= C^{FI}(y_1, y_2) + \delta \mathbb{E} V_{t+1}^{FI}\left(f_1(y_1, y_2, D), f_2(y_1, y_2, D)\right) \\ &= C^{FI}(y_1, y_2) + \delta \mathbb{E} V_{t+1}^{FI}\left(y_1^{FI}, y_2^{FI}\right) \end{aligned} \tag{31}$$

as V_{t+1}^{FI} is affine in x^{t+1} and the purchasing cost is zero. Clearly, $y = y^{FI}$ minimizes U_t^{FI} for $y \leq y^{FI}$. Thus, $V_t^{FI}(x) = \max_{y \geq x} U_t^{FI}(y)$ is affine (constant) in x for all $x \leq y^{FI}$, and hence a stationary base-stock policy (y^{FI}) is optimal if $x \leq y^{FI}$. If there is some $x_i > y_i^{FI}$, the optimal policy will be more complicated, but eventually, the system comes back to $x \leq y^{FI}$. □