# Beyond Retail Stores: Managing Product Proliferation along the Supply Chain

Işık Biçer
Schulich School of Business, York University, 111 Ian MacDonald Blvd, M3J1P3, Toronto, ON, Canada, bicer@schulich.yorku.ca,

Florian Lücker
Cass Business School, City, University of London, London EC1Y, UK, florian.lucker@city.ac.uk,

Tamer Boyacı
ESMT Berlin, Schlossplatz 1, 10178, Berlin, Germany, tamer.boyaci@esmt.org,

Product proliferation occurs in supply chains when manufacturers respond to diverse market needs by trying to produce a range of products from a limited variety of raw materials. In such a setting, manufacturers can establish market responsiveness and/or cost efficiency in alternative ways. Delaying the point of the proliferation helps manufacturers improve their responsiveness by postponing the ordering decisions of the final products until there is partial or full resolution of the demand uncertainty. This strategy can be implemented in two different ways: (1) redesigning the operations so that the point of proliferation is swapped with a downstream operation or (2) reducing the lead times. To establish cost efficiency, manufacturers can systematically reduce their operational costs or postpone the high-cost operations. We consider a multi-echelon and multi-product newsvendor problem with demand forecast evolution to analyze the value of each operational lever of the responsiveness and the efficiency. We use a generalized forecast-evolution model to characterize the demand-updating process, and develop a dynamic optimization model to determine the optimal order quantities at different echelons. Using anonymized data of Kordsa Inc., a global manufacturer of advanced composites and reinforcement materials, we show that our model outperforms a theoretical benchmark of the repetitive newsvendor model. We demonstrate that reducing the lead time of a downstream operation is more beneficial to manufacturers than reducing the lead time of an upstream operation by the same amount, whereas reducing the upstream operational costs is more favorable than reducing the downstream operational costs. We also indicate that delaying the proliferation may cause a loss of profit, even if it can be achieved with no additional costs. Finally, a decision typology is developed, which shows effective operational strategies depending on product/market characteristics and process flexibility.

*Key words*: Product proliferation; lead-time reduction; process redesign; delayed differentiation.

## 1. Introduction

Digital transformation in the retail industry (e.g., omni-channel retailing, recommendation systems, and user-oriented product development using social media) has led to an increase in demand for niche items in almost all product categories (Brynjolfsson et al. 2011). Retailers now carry more diverse product portfolios than in past decades in both online and physical stores. The expansion of product portfolios has a negative impact on supply-demand mismatches in the retail industry (Rajagopalan 2013). Arguably, the challenges associated with diverse product portfolios are not only limited to downstream sales channels (retailers, online channels), but start with upstream operations (Atalı and Özer 2012). In fact, it is not uncommon for manufacturers to attempt to fulfill customer demand for broad product lines by *using the same upstream resources and differentiating products over time* as they get closer to markets. This strategy helps them to benefit from economies of scale for upstream resources and to postpone product differentiation until acquisition of more accurate market demand forecasts.

Fashion apparel is perhaps the most celebrated industry where product proliferation is prominent and has a profound impact on profitability (Lee and Tang 1997). Global manufacturers like Zara, H&M, and Uniqlo sell a variety of clothes in each selling season, which are produced using the same textile but sewn and colored differently. These manufacturers typically follow a series of operations to make clothes, and the proliferation occurs along the production stages. After a product design team develops new designs, for example, yarns selected by the team are ordered. Production occurs sequentially involving the weaving, sewing, and dyeing processes. First, yarns are transformed into textile by the weaving process. Then, the textile is sewn into different models and sizes. Finally, the items are dyed into different colors to complete the production. Product proliferation occurs sequentially during the sewing and dyeing processes.

We observe similar dynamics at Kordsa Inc., a global manufacturer of advanced composites and reinforcement materials, such as tire cords, which actively operates in Turkey, the US, Brazil, Thailand, and Indonesia, with 11 production sites and around 4,500 employees. Figure 1 presents the production steps of tire cords. The cords are sold to tire manufacturers, and their demand is both volatile and seasonal due to the seasonality of tire sales. In the first production step, polypropylene is processed into polymer threads. The threads are first twisted before they undergo the weaving process. Finally, the woven products enter a chemical blending process in which they are dipped into chemical liquids to bring the products to the right level of thermal resistance and elasticity. Although the variety of polymer materials is limited, there exists a high variety of end products due to the product proliferation in the last three stages.

Beyond fashion apparel manufacturers and Kordsa, the dynamics of product proliferation are also observed in different industries. According to an industry survey conducted by E2Open (2018), which covers global manufacturers across various industries, the effective management of proliferation along sequential production stages is considered one of the key success factors because the variety of products sold by the surveyed manufacturers has increased on average by 236% from 2010 to 2017, while total sales have grown by only 15% during the same time period.

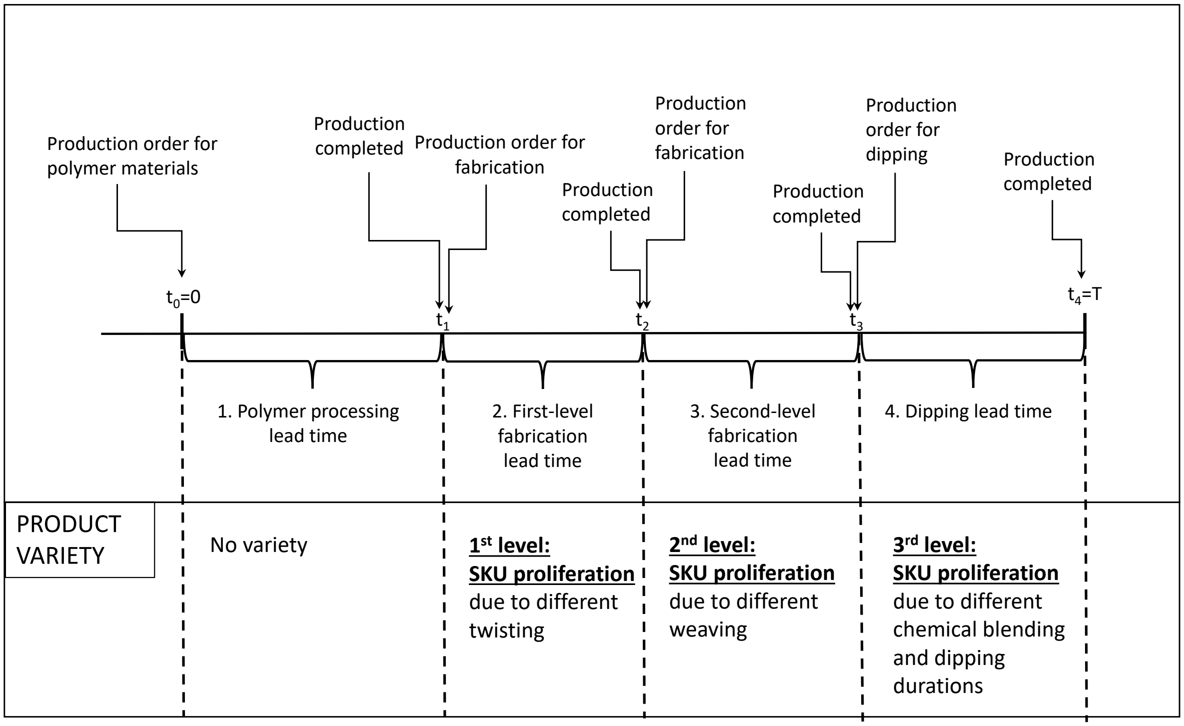


**Figure 1 Overview of Kordsa's production stages**

Manufacturers operating in such settings are often exposed to high demand uncertainty for upstream production orders. As production moves forward, demand uncertainty is partially resolved due to additional valuable demand information collected from the market. For downstream production orders, however, manufacturers are exposed to high product variety. When trading off cost efficiency in exchange for operational responsiveness, Fisher (1997) indicates that physically efficient supply chains are better aligned with products that have low demand uncertainty, whereas market-responsive supply chains are better aligned with products that have high demand uncertainty. Due to the evolutionary risk structure in a product proliferation model, the utilization of both market-responsive and cost-efficient strategies may improve profits depending on the supply chain structure and the cost, demand, and lead-time parameters.

*Delaying differentiation* is an effective strategy for improving the responsiveness of supply chains in which product proliferation occurs (the terms "differentiation" and "proliferation" are used interchangeably). It enables manufacturers to take advantage of inventory pooling at upstream echelons, while ensuring that the proliferation at downstream occurs with more accurate demand information. There are two practical approaches to operationalizing delayed differentiation, both of which have been widely popularized by their implementation in the fashion apparel industry. The first approach is to *redesign the processes* so that the operations that cause proliferation are deferred to a later stage in the supply chain. Benetton, the Italian clothing company, is the first firm that successfully implemented this approach and reversed the order of the dyeing and knitting operations (Heskett and Signorelli 1989, Lee and Tang 1997). Traditionally Benetton spun and dyed the yarns first and then knitted the colored yarns. In 1972, the company began dyeing clothes rather than yarns to postpone the costly dyeing operation. This allowed Benetton to postpone product differentiation until it could observe accurate market demand information, leading to higher profits due to the decrease in supply-demand mismatches. Given the success of this approach, many other companies followed Benetton's lead (Parsons and Graves 2005, Viswanathan and Allampalli 2012, Kouvelis and Tian 2014).[1]

The second approach is to *reduce lead times* for each operation in the supply chain. Zara, the Spanish fashion apparel company, followed this strategy and became the market leader in 2008 (Ghemawat and Nueno 2006). When lead times are long, demand forecasts are often plagued with high uncertainty. Reducing lead times allows manufacturers to postpone the point of proliferation and actual ordering decisions closer to market demand, making it possible to place production orders based on more accurate demand forecasts. This, in turn, leads to a decrease in supply-demand mismatches (Caro and Martínez-de Albéniz 2015).

To establish cost efficiency, manufacturers can systematically reduce the cost of all operations along the supply chain or postpone the high-cost operations to a later stage. The former helps them reduce the unit product cost, whereas the latter makes it possible to avoid unnecessary overutilization of expensive resources. The value of each operational lever of the responsiveness and the efficiency depends on the supply chain structure and the cost and lead-time values of each operation along the supply chain. Our objective in this paper is to quantify their costs and benefits. To this end, we consider a multi-echelon and multi-product newsvendor model with demand forecast evolution. We make three important contributions to the extant literature. First, from a modeling perspective, we develop an analytical framework for dynamically optimizing inventory/ordering quantities in a multi-echelon and multi-product newsvendor setting with demand forecast evolution. This framework extends the existing inventory models in the literature (Wang et al. 2012, Biçer and Seifert 2017) by incorporating forecast evolution to multi-product and multi-echelon settings. Our framework takes the supply chain structure along with lead times and cost values for each echelon as inputs, incorporates the evolution of demand forecasts using a generalized model, and optimizes the ordering decisions at each echelon. We characterize the optimal strategy and investigate the effects

[1] We remark that process redesign does not necessarily require operations to be swapped; it may also be achieved by changing the way operations are performed (and associated costs). In the auto-tire industry, for example, it is possible to meet the technical specifications requested by a customer by changing either technical grades used during the fabrication process or the chemical recipes used during the blending process. The latter enables the postponement of the point of proliferation, but increases production costs.

on profits of salient parameters—costs, lead times, proliferation points. We calibrate our model by using anonymized sales data from Kordsa. Our approach outperforms a theoretical benchmark of the repetitive newsvendor model such that it increases profits by 135% over the benchmark.

Second, utilizing this framework, we analytically demonstrate the critical impact of cost changes, lead-time reduction, and postponement points on optimal inventory levels and consequent profits. We establish that reducing the lead time of a downstream operation is more beneficial to manufacturers than reducing the lead time of an upstream operation by the same amount, whereas reducing the costs of upstream operations is more favorable than reducing the costs for downstream operations. We also show that delaying the proliferation may cause a loss of profit. Such a loss in profits may occur if its implementation requires swapping a high-cost downstream operation with a low-cost upstream operation. Further, even when all costs across echelons are equal, delaying the proliferation may still cause a loss of profit. This occurs when postponing the proliferation requires swapping an upstream operation with a long lead time with a downstream operation with a short lead time (see Theorem 3). These results are counterintuitive to conventional wisdom that the postponement strategy would always help reduce the cost of mismatches between supply and demand (please see Zinn (2019) for a historical review of the evolution of the postponement research). We analytically develop a threshold value such that deferring the proliferation by swapping it with a high-cost downstream operation causes a loss of profit when the cost of the downstream operation exceeds the threshold value. We reveal how the threshold value is affected by demand volatility, demand correlation, and the number of products in the portfolio.

Third, we translate the descriptive results into prescriptive insights for practicing managers, particularly with respect to the implementation of delayed differentiation. We provide normative support for when postponing the proliferation actually leads to a profit gain depending on the costs and lead times of the operations in the supply chain. We also provide a decision topology that indicates the most appropriate strategy based on product/market characteristics and process flexibility.

## 2. Literature Review

Our research has a natural connection with studies that focus on postponement strategies for delaying product differentiation. One stream within this literature focuses on the design of supply chain structures (Johnson and Anderson 2000, Lee and Tang 1997, 1998), capacity investments (Kouvelis and Tian 2014), and inventory levels at the decoupling points (i.e., vanilla boxes) (Swaminathan and Tayur 1998, Paul et al. 2015). Common to these papers is that demand is assumed to be random without an evolutionary form, so the benefits of postponement are only attributed to inventory pooling—benefits due to improved forecast accuracy are not incorporated. Another stream focuses precisely on demand evolution. In particular, Aviv and Federgruen (2001a,b) analyze the value of order postponement in a multi-period inventory setting where sales occur in each period and demand forecasts are updated according to a Bayesian model. In a similar vein, Atalı and Özer (2012) develop a two-stage production model with product differentiation occurring at the beginning of the second stage under a Markov-modulated demand model. They show that the value of postponement increases with operational flexibility (as measured by difference in minimum and maximum production limits). Our contribution to this literature is the development of a multi-echelon and multi-product newsvendor model with demand forecast evolution. Utilizing this model, we quantify the impact of supply chain structure, the cost and the lead-time values on profits in a product-proliferation setting. Thus, our results shed light on how to employ operational responsiveness and cost efficiency to improve the profits.

Our paper is also connected to the operations management literature that focuses on the multi-ordering inventory models with demand forecast evolution. The closest papers within this literature are Wang et al. (2012) and Biçer and Seifert (2017) because they develop integrated dynamic inventory models with the martingale model of forecast evolution (MMFE). Wang et al. (2012) model a newsvendor with multiple ordering opportunities and increasing costs over time, and characterize optimal base-stock levels. Biçer and Seifert (2017) extend Wang et al. by including capacity limitations and allowing for multiple products. In both papers, the ordering decisions are made only for the end products, not for the components or the raw materials at the upstream echelons. We contribute to the extant literature such that we optimize ordering decisions in a multi-echelon setting in which the order quantity of a given operation determines the capacity for the immediate downstream echelon. We also consider the possibility of product proliferation to occur at any echelon in the supply chain. For the same reasons, our model differs from single-item inventory models with evolving demand forecasts and multiple ordering opportunities. Song and Zipkin (2012) study such a setting where order quantities can be updated downwards (after paying the cost) as new demand information arrives. Cao and So (2016) consider an assembler ordering from two suppliers (effectively two ordering decisions) with demand forecasts updated over time.

In addition to the above-mentioned theoretical contributions, we calibrate our model using empirical data from Kordsa. In the company setting, evolution of demand forecasts occurs depending on the advance demand (firm orders) received from customers. We use a generalized demand model to capture the demand dynamics when the firm orders are used as the sole source of demand information. We show how to fit the modeling parameters to empirical data and compare our model with a theoretical benchmark of the repetitive newsvendor model. Our model outperforms the benchmark.

## 3. Model Preliminaries

Consider a supply chain with $(n+1)$ echelons, where the most downstream echelon $n$ is closest to the customer and echelon 0 is the farthest from the customer. Echelon $i+1$ is considered to be the downstream and echelon $i-1$ the upstream of echelon $i$. Supply chain activities occur sequentially such that the operation at echelon $i$ uses the output of echelon $i-1$ as input and transforms it into output. The output of echelon $i$ is then used as input for echelon $i+1$. Without loss of generality, we assume that one unit of input is transferred into one unit of output. The manufacturer has to make $n$ ordering decisions at the time epoch $t_i$ for $i \in \{0, 1, \cdots, n-1\}$. Hence, there is a positive lead time at each echelon; $t_{i+1} - t_i > 0$ for $i \in \{0, 1, \cdots, n-1\}$ and expediting is not allowed. For ease of exposition, suppose for now that there is a single final product, and let $Q_i$ denote the order quantity at echelon $i$. The order quantity $Q_i$ for $i \in \{1, \cdots, n-1\}$ is constrained by the order quantity at the previous echelon (i.e., $Q_i \leq Q_{i-1}$), while the first order quantity $Q_0$ is unrestricted. We use $D_i$ to denote the demand forecast at time $t_i$ for $i \in \{0, \cdots, n\}$, with the end demand forecast $D_n$ representing the actual market demand. The timeline of ordering decisions for this single-product model without any product proliferation is depicted in Figure 2.


**Figure 2** **Timeline of ordering decisions for a single-product model without product proliferation**

The following sequence of events occurs at each decision epoch $t_i$ for $i \in \{0, \cdots, n-1\}$: i) manufacturer observes the demand forecast $D_i$; ii) the order quantity of the previous operation $Q_{i-1}$ is reviewed; iii) the order quantity $Q_i$ is determined, and the manufacturer incurs an operational cost $c_i Q_i$.

We model the evolution of demand forecasts $D_i$ from $t_0$ to $t_n$ according to a generalized demand model that can be developed in either an additive or a multiplicative form. In practice, there are two common mechanisms employed to update demand forecasts. The first mechanism is the judgmental forecast updating mechanism such that demand planners dynamically update the point forecasts for future periods (Diermann and Huchzermeier 2017). They collect information from different sources and estimate the demand for future months. They dynamically update the forecasts in each period until the actual demand is realized. The de-biased estimates in this approach can be modeled by the martingale model of forecast evolution (MMFE) (Hausman 1969, Heath and Jackson 1994).

The second mechanism is the advance demand information mechanism such that firm orders received from customers are the only source of demand information. Firm orders serve as a lower bound for the actual demand, and they accumulate over time until the actual demand is realized (i.e., occurs at time $t_n$). When the demand model is constructed in such a way that $D_i$ denotes the amount of accumulated advance demand that has to be fulfilled at time $t_n$, the demand process cannot satisfy the martingale property. In this case, the accumulated demand structure should be captured by a submartingale with a positive drift rate because advance demand serves as a lower bound for the actual demand. Our generalized demand model involves the drift rate as a modeling parameter, so it can be used for both mechanisms.

In addition to the distinction between martingale and submartingale forms, there is another dimension to categorize demand models: either (1) additive or (2) multiplicative. Both versions fit well the empirical data under different circumstances. For example, the multiplicative model fits well when demand uncertainty is relatively high or the forecasting horizon is long; while the additive version fits well when the forecasting horizon is short (Wang et al. 2012, Biçer et al. 2018, Oh and Özer 2013).

In the following two subsections, we present the additive and the multiplicative versions of the demand model that can be utilized in martingale or submartingale forms. We develop our analytical results in Sections 4 and 5 based on a generalized demand distribution form, so each model can be applied to our analytical framework.

### 3.1. Additive Demand Model

In the additive model, the demand forecasts are updated additively such that $D_i = D_0 + \mu(t_i - t_0) + \epsilon_1 + \epsilon_2 + \cdots + \epsilon_i$ for $i \in \{1, \cdots, n\}$, where $\mu$ is the drift rate and $\epsilon_i$ is the incremental forecast adjustment at time $t_i$ that follows a normal distribution:

$$\epsilon_i \sim \mathcal{N}(0, \sigma\sqrt{t_i - t_{i-1}}), \qquad \forall i \in \{1, \cdots, n\}. \tag{1}$$

The end demand conditional on the demand forecast at time $t_i$ follows the normal distribution:

$$D_n | D_i \sim \mathcal{N}(D_i + \mu(t_n - t_i), \sigma\sqrt{t_n - t_i}). \tag{2}$$

Demand uncertainty for the distribution (2) depends on not only the volatility parameter $\sigma$ but also the time when the forecast is made. As $t_i$ approaches to $t_n$, the forecast horizon is shortened and the accuracy of forecasts is improved, which is relevant in practice. The (positive) drift rate captures the upward trend in the demand process. If advance orders are the sole source of demand information, the drift rate should be positive and set equal to the arrival rate of advance orders. If forecasts are formed according to the judgmental forecasting, the drift rate should be set equal to

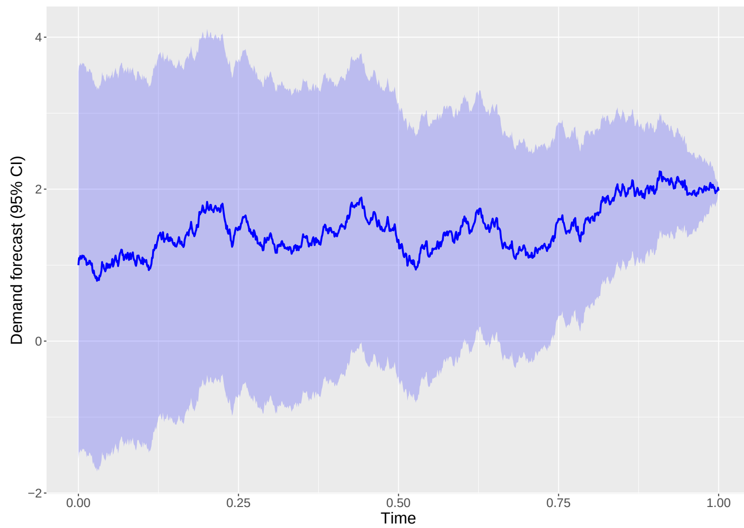


**Figure 3 A sample path of the demand forecast according to the a-MMFE with 95% two-sided confidence intervals**

zero. The case with the zero drift rate reduces to the additive version of the martingale model of forecast evolution (a-MMFE).

In Figure 3, we present an example of the evolution of demand forecasts according to the a-MMFE. We simulate a random path assuming that the initial demand forecast is scaled to one. We set the $\sigma$ parameter to a high value (i.e., $\sigma = 1.3$) to highlight the problems related to the a-MMFE. The forecast evolves from $t_0 = 0$ until $t_n = 1$. The solid curve represents the mean forecast, and the shaded area shows the 95% confidence interval. As time approaches the realization of market demand ($t \to 1$), the forecast accuracy increases significantly as indicated by a reduction of the distance between the upper and the lower bounds of the confidence interval. The lower bound is negative when $t \leq 0.60$, although, true demand values cannot be negative. For this reason, the a-MMFE fits poorly empirical data when the forecast horizon is long.

The additive model causes some problems when an ordering decision is made for low-margin products with a long lead time. The newsvendor solution indicates that the in-stock probability for the optimal decision should be low for low-margin products. In Figure 3, for example, %10 in stock-probability is around "$-0.5 \times D_0$" when the ordering decision is made at $t = 0$. Thus, the optimal order quantity can be found negative if a-MMFE is used for a such product when the lead time is long.

### 3.2. Multiplicative Demand Model

The multiplicative model is structurally different from the additive one such that the incremental forecast adjustments are accumulated multiplicatively. In mathematical terms, $D_i = D_0 \exp(\mu(t_i - t_0) + \varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_i)$, where $\varepsilon_t$ follows a normal distribution:

$$\varepsilon_i \sim \mathcal{N}(-\sigma^2(t_i - t_{i-1})/2, \sigma\sqrt{t_i - t_{i-1}}), \qquad \forall i \in \{1, \cdots, n\}. \tag{3}$$

The term $\exp(\epsilon_i)$ follows a lognormal distribution and $\mathbb{E}(\exp(\epsilon_i)) = 1$. The end demand conditional on the demand forecast at $t_i$ follows a lognormal distribution:

$$\ln(D_n)|D_i \sim \mathcal{N}(\ln(D_i) + (\mu - \sigma^2/2)(t_n - t_i), \sigma\sqrt{t_n - t_i}), \qquad \forall i \in \{0, \cdots, n-1\}. \tag{4}$$

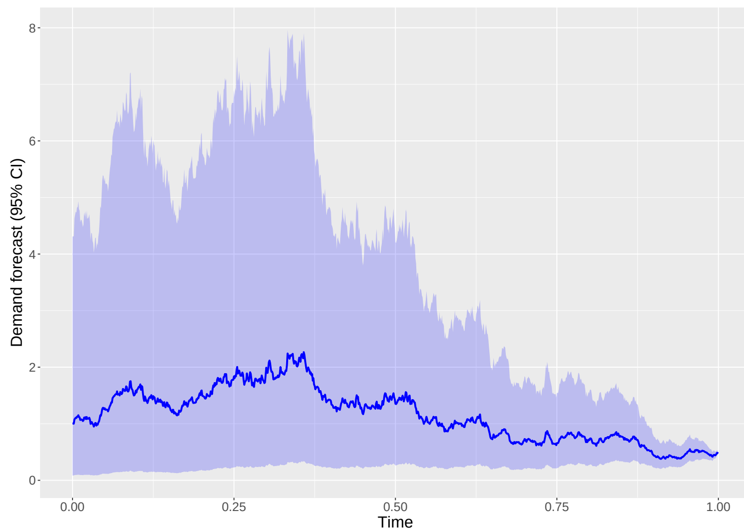


**Figure 4 A sample path of the demand forecast according to the m-MMFE with 95% two-sided confidence intervals**

In Figure 4, we present an example of the evolution of demand forecasts according to the multiplicative martingale model of forecast evolution (m-MMFE). We simulate a random path assuming that the initial demand forecast is scaled to one and the $\sigma$ value is set to one. The coefficient of variation for $\sigma = 1$ is equal to 1.3, which is the same as in Figure 3. The forecast evolves from $t_0 = 0$ until $t_n = 1$. The solid curve represents the mean forecast, and the shaded area shows the 95% confidence interval. As time approaches the realization of market demand ($t \to 1$), the forecast accuracy increases significantly as indicated by a reduction in the distance between the upper and the lower bounds of the confidence interval. Contrary to the a-MMFE, the m-MMFE avoids negative forecasts such that the lower bound cannot be negative due to the lognormal property. Thus, it provides a better fit with data than the a-MMFE when the forecast horizon is long.

## 4. Single-Product Model

In this section, we formulate the manufacturer's optimization problem and derive its solution for the single-product case. Consider the single-product model shown in Figure 2, where the final product is sold in a single market. The product is processed from raw materials through a sequence of operations and sold in the market at a price of $p$ per unit. We assume that there is no salvage value for the excess inventory. Thus, a revenue of $p\min(D_n, Q_{n-1})$ is collected at time $t_n$. Let $c_i$ denote the cost of processing the $i^{th}$ operation per unit input. This includes all the cost elements such as labor, utility, material, and other operational costs that the manufacturer incurs only from $t_i$ until $t_{i+1}$.

We formulate the manufacturer's optimization problem as a dynamic program (DP). At each decision epoch $t_i$, the manufacturer observes the state, which consists of the available supply $Q_{i-1}$ at the upstream echelon and demand forecast $D_i$, and then determines the order quantity $Q_i$ that maximizes expected profits. For the last decision epoch $t_{n-1}$, the ordering decision is a constrained newsvendor problem:

$$V_{n-1}(Q_{n-2}, D_{n-1}) = \max_{Q_{n-1} \leq Q_{n-2}} \left\{ \mathbb{E}_{D_n|D_{n-1}} \left[ p \min(D_n, Q_{n-1}) \right] - c_{n-1} Q_{n-1} \right\}. \tag{5}$$

The ordering decisions at the earlier decision epochs (i.e., $\forall i \in \{0, \cdots, n-2\}$) can be determined dynamically according to the following Bellman equation:

$$V_i(Q_{i-1}, D_i) = \max_{Q_i \leq Q_{i-1}} \left\{ \mathbb{E}_{D_{i+1}|D_i} \left[ V_{i+1}(Q_i, D_{i+1}) - c_i Q_i \right] \right\}. \tag{6}$$

The order quantity at $t_0$ is not constrained, so we set $Q_{-1} = +\infty$. Let the functions to be maximized in Equations (5) and (6) be denoted respectively as:

$$G_{n-1}(Q_{n-1}, D_{n-1}) = \mathbb{E}_{D_n|D_{n-1}} \left[ p \min(D_n, Q_{n-1}) \right] - c_{n-1} Q_{n-1}, \tag{7}$$

$$G_i(Q_i, D_i) = \mathbb{E}_{D_{i+1}|D_i} \left[ V_{i+1}(Q_i, D_{i+1}) - c_i Q_i \right], \tag{8}$$

with $g_i(Q_i, D_i) = \partial G_i(Q_i, D_i)/\partial Q_i$.

Observe that the optimal value of $Q_i$ in Equation (6) depends on the demand forecasts in all future decision epochs. We define a new parameter $\overline{D}_j$ for $j \in \{i+1, \cdots, n-1\}$ to represent the critical demand forecast values at time $t_j$. If $D_j \geq \overline{D}_j$ $\forall j \in \{i+1, \cdots, n-1\}$, the optimal order quantities in all the remaining decision epochs become equal to $Q_i$. If $D_j < \overline{D}_j$ for $j > i$, the optimal value of $Q_j$ becomes less than $Q_i$. Therefore, $\overline{D}_j$ values for $j \in \{i+1, \cdots, n-1\}$ determine the lower bounds for demand forecasts that make optimal order quantity at time $t_j$ equal to $Q_i$. Solving the DP model by backward induction, we characterize the optimal ordering policy at each decision epoch, which is presented in the next theorem.[2]

THEOREM 1. *The optimal order quantity, denoted by $q_i$ for $i \in \{0, \cdots, n-1\}$, satisfies:*

$$q_i = \min(Q_{i-1}, Q_i^*), \tag{9}$$

[2] The proofs of all results are presented in our online appendix.

*where $Q_i^*$ is the optimal order quantity for the unconstrained problem (without "$Q_i \le Q_{i-1}$"), which is found by the following expressions:*

$$Q_i^* = \{Q_i \mid g_i(Q_i, D_i) = 0\}, \tag{10}$$

$$g_i(Q_i, D_i) = pPr(D_n > Q_i, \mathbf{D}_{\{i+1,n-1\}} > \overline{\mathbf{D}}_{\{i+1,n-1\}}) - c_{n-1}Pr(\mathbf{D}_{\{i+1,n-1\}} > \overline{\mathbf{D}}_{\{i+1,n-1\}})$$
$$-c_{n-2}Pr(\mathbf{D}_{\{i+1,n-2\}} > \overline{\mathbf{D}}_{\{i+1,n-2\}}) - \cdots - c_{i+1}Pr(D_{i+1} > \overline{D}_{i+1}) - c_i = 0 \tag{11}$$

*with $\mathbf{D}_{\{i+1,n-1\}}$ denoting the vector of demand forecasts from $i+1$ to $n-1$ and $\overline{\mathbf{D}}_{\{i+1,n-1\}} = (\overline{D}_{i+1}, \cdots, \overline{D}_{n-1})$ denoting the vector of critical demand forecasts from $i+1$ to $n-1$.*

It can be easily verified that equation (11) reduces to the newsvendor solution for $i = n-1$ such that:

$$g_{n-1}(Q_{n-1}, D_{n-1}) = pPr(D_n > Q_i) - c_{n-1} = 0. \tag{12}$$

For $i < n-1$, the solution is still in the spirit of the newsvendor solution. The first term on the right-hand side of Equation (11) gives the expected value of the marginal revenue generated by ordering one additional unit after $(Q_i - 1)$ units are already ordered. The marginal revenue depends on not only the final demand realization $D_n$ but also the updated demand forecasts at the remaining decision epochs. Even when $D_n > Q_i$, the marginal revenue may be zero if the manufacturer decides to reduce the order quantity in any of the subsequent production stages. The remaining terms of the right-hand side of Equation (11) give the expected value of the marginal cost of ordering one additional unit when $(Q_i - 1)$ units are already ordered. When the ${Q_i}^{th}$ unit is ordered, the manufacturer incurs the cost $c_i$. If the demand forecast at the next decision epoch exceeds the critical value (i.e., $D_{i+1} \ge \overline{D}_{i+1}$), the manufacturer orders $Q_i$ units at $t_{i+1}$ and incurs an additional cost of $c_{i+1}$ per unit and so forth.

PROPOSITION 1. *Optimal order quantity in an upstream echelon is always higher than the expected (optimal) order quantity in a downstream echelon such that $q_i > E[q_{i+j}|D_i]$, $i = 0, 1, \cdots, n-2$, and $j = 1, 2, \cdots, n-1-i$.*

Proposition 1 states that the interdependency between order quantities (due to supply constraints) and the accumulating cost structure induce the manufacturer to order in large quantities for the upstream operations even though the manufacturer expects the final order quantity to be lower. Next, we present the impact of cost parameters on optimal order quantities and the expected profit.

PROPOSITION 2. *A- Let $\boldsymbol{q} = \{q_0, q_1, \cdots, q_{n-1}\}$ be the vector of optimal order quantities at each decision epoch. If $c_j$ for $j \in \{0, \cdots, n-1\}$ increases, the optimal order quantities are updated such that $\boldsymbol{q'} = \{q'_0, q'_1, \cdots, q'_{n-1}\}$, where $q'_i$ is statistically smaller than $q_i$ (i.e., $q'_i \prec q_i$) $\forall i \in \{0, \cdots, n-1\}$.*

*B- Let* $c_0 = c_1 = \cdots = c_{i-1} = c_{i+1} = \cdots = c_{n-1} = c_{fixed}$ *and* $c_i > c_{fixed}$*. Swapping the operation* $i$ *with a downstream operation* $j \in \{i+1, \cdots, n-1\}$ *increases the total expected profit if the two operations swapped have the same lead time.*

Part A of Proposition 2 describes how the order quantities are affected by an increase in the cost of any operation. If the cost of an operation increases, order quantities at all decision epochs decrease. Part B shows how the sequence of the operations should be redesigned depending on the operational costs. The manufacturer increases its profits by swapping a high-cost and upstream operation with a low-cost and downstream operation if the operations swapped have the same lead time. Swapping the high-cost operation with a downstream lower-cost operation increases the upstream order quantity and hence the available supply (upper bounds) for the downstream operation. An increase in the upper bounds for the downstream quantities provides the manufacturer with additional flexibility to adjust order quantities according to updated demand forecasts, leading to higher profits. Proposition 2.B holds if the lead times of the swapped operations are identical. The following Proposition elaborates on the sensitivity of the profits in lead times.

PROPOSITION 3. *A- Reducing the lead time of operation* $i$ *for* $i \in \{1, \cdots, n-1\}$ *by an amount of* $\Delta t \le t_{i+1} - t_i$ *increases expected profit more than what can be achieved by reducing the lead time of operation* $j < i$ *by the same amount of* $\Delta t$*.*

*B- Let* $c_0 = c_1 = \cdots = c_i = \cdots = c_{n-1}$*,* $t_1 - t_0 = t_2 - t_1 = \cdots = t_{i-1} - t_{i-2} = t_{i+1} - t_i = \cdots = t_n - t_{n-1} = \Delta t_{fixed}$*, and* $t_i - t_{i-1} < \Delta t_{fixed}$*. Then, swapping operation* $i$ *with a downstream operation* $j \in \{i+1, \cdots, n-1\}$ *increases the total expected profit.*

Part A of Proposition 3 states that reducing the lead time of a downstream operation is more beneficial to the manufacturer than reducing the lead time of an upstream operation by the same amount. This result is in line with de Treville et al. (2004), who argue that the benefits of lead-time reduction highly depend on the demand management activities, so any effort to reduce the lead times should focus on downstream demand management activities. Proposition 3.A formally proves that lead time reduction in a downstream echelon is more beneficial than that of an upstream echelon. Part B of Proposition 3 states that if all operations have the same cost value, swapping a short-lead-time operation with a downstream operation that has a longer lead time helps increase the profit.

Propositions 2.B and 3.B combined are in line with Lee and Tang (1997) and Cao and So (2016). Lee and Tang (1997) state that redesigning the production processes so that high value-added and short operations take place later than low value-added and long operations leads to higher profits. Cao and So (2016) find that a manufacturer can generate high profits if a supplier with a long lead time supplies a low-value component, whereas another supplier with a short lead time supplies a high-value component. Propositions 2.B and 3.B effectively establish the same result for a more general setting.

The analytical results given by Propositions 1–3 provide useful insights and clear guidance on how a manufacturer should implement process-redesign and lead-time-reduction practices. Even in the absence of product proliferation at any echelon, manufacturers can still increase the profits by redesigning their processes to postpone high-cost operations. When a manufacturer aims to reduce its operational costs, it should first focus on the upstream operations and then move sequentially downstream. However, the manufacturer should start from the downstream operations and then move upstream if the objective is to reduce the lead time.

## 5. Product Proliferation Model

We now extend the single-product model to the multi-product case where the raw materials or semi-finished products are transformed into a variety of products and product proliferation is allowed at any decision epoch. Clearly, some of the insights from Section 4 can be generalized to the product proliferation model. For example, Proposition 3.A also holds in the presence of product proliferation (see the proof of Proposition 3.A in the online appendix for details). The objective of this section is to better understand the impact of key modeling parameters on profits when product proliferation occurs along the supply chain.

To facilitate model development, in Figure 5, we present an example where product proliferation occurs at two epochs: $t_1$ and $t_{n-2}$. We use $Q_i^j$ to denote the order quantity placed for component $j$ at time $t_i$. We use a unique code to label the component $j$ at $t_i$. The code is a sequence of single digits, and the length of the code shows how often product proliferation occurs from $t_0$ until $t_i$. In our example in Figure 5, three different components are ordered at $t_1$, each taking a different digit number. The second proliferation occurs at $t_{n-2}$, where the inventory of each component is allocated to produce three differentiated products, amounting to nine stock-keeping units (SKUs) available in the market. Thus, a new digit is added to the product code at $t_{n-2}$. Suppose, for example, a fashion-apparel manufacturer selling a product line to different markets uses a three-digit product code (e.g., 361). The first digit represents the size (e.g., small, medium, or large), the second digit represents the color, and the third digit denotes the market. The three-digit code means that product proliferation occurs three times along the supply chain (one for size, one for color, and the last for different markets).

The primary challenge in solving the product proliferation problem lies with the need to link the demand dynamics to the ordering constraints. For each ordering decision, it is necessary to consolidate the demand updates of different end products and then allocate the limited supply available from the previous operation to process different semi-finished or end products. We define

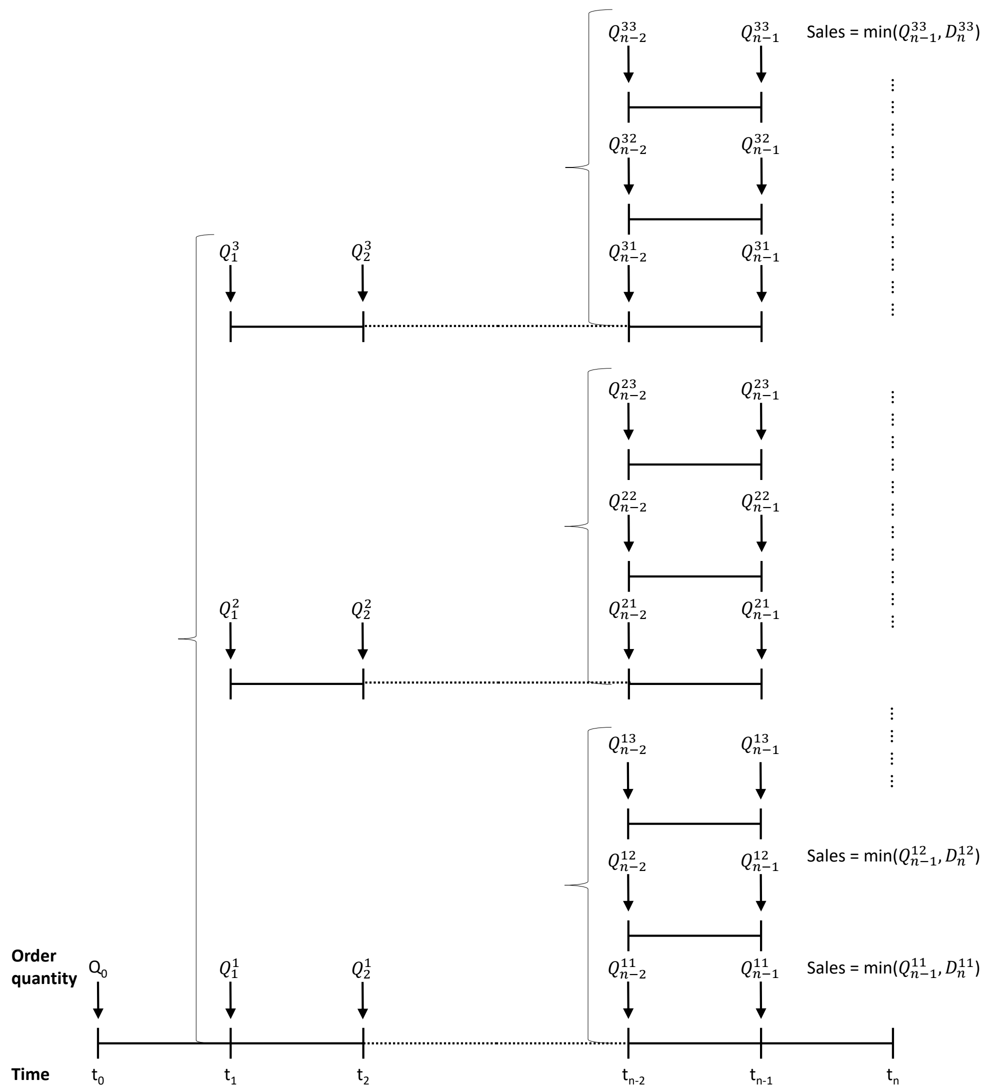


**Figure 5** **Ordering decisions in a supply chain with the product proliferation at the decision epochs $t_1$ and $t_{n-2}$**

*two different sets and their subsets* to formalize the problem. To capture the resource constraints, we use $\Theta_i$ to denote the set of all components produced at echelon $i \in \{1, \cdots, n\}$ at time $t_i$. We further partition the set $\Theta_i$ into $k$ pairwise disjoint subsets such as $\Theta_i^j$ for $j \in \{1, \cdots, k\}$ and $k = |\Theta_{i-1}|$. We define $\Theta_i^j$ as the set that contains all components that use the same upstream resource as their input. We then have, by definition:

$$\Theta_i = \bigcup_{j \in \Theta_{i-1}} \Theta_i^j \quad \text{and} \quad \emptyset = \bigcap_{j \in \Theta_{i-1}} \Theta_i^j. \tag{13}$$

Recalling our example in Figure 5, $\Theta_{n-1} = \{11, 12, 13, 21, 22, 23, 31, 32, 33\}$. There are nine ordering decisions in the previous period (i.e., $t = t_{n-2}$); therefore, the set $\Theta_{n-1}$ is partitioned into nine subsets such that:

$$\Theta_{n-1} = \Theta^{11}_{n-1} \cup \Theta^{12}_{n-1} \cup \Theta^{13}_{n-1} \cup \Theta^{21}_{n-1} \cup \Theta^{22}_{n-1} \cup \Theta^{23}_{n-1} \cup \Theta^{31}_{n-1} \cup \Theta^{32}_{n-1} \cup \Theta^{33}_{n-1},$$

where $\Theta^{11}_{n-1} = \{11\}$, $\Theta^{12}_{n-1} = \{12\}$, $\Theta^{13}_{n-1} = \{13\}$, $\Theta^{21}_{n-1} = \{21\}$, $\Theta^{22}_{n-1} = \{22\}$, $\Theta^{23}_{n-1} = \{23\}$, $\Theta^{31}_{n-1} = \{31\}$, $\Theta^{32}_{n-1} = \{32\}$, and $\Theta^{33}_{n-1} = \{33\}$. Likewise, at $t = t_{n-2}$, $\Theta_{n-2} = \{11, 12, 13, 21, 22, 23, 31, 32, 33\}$. There are three ordering decisions in the previous period (i.e., $t = t_{n-3}$) so $\Theta_{n-2}$ is partitioned into three subsets: $\Theta^1_{n-2} = \{11, 12, 13\}$, $\Theta^2_{n-2} = \{21, 22, 23\}$, and $\Theta^3_{n-2} = \{31, 32, 33\}$.

With these sets defined, we can write down the ordering constraints between echelons. That is, the sum of the order quantities for the products that use the same input cannot be larger than the order quantity of the input at the immediate upstream echelon. In mathematical terms:

$$\sum_{j \in \Theta^k_i} Q^j_i \leq Q^k_{i-1}. \tag{14}$$

Returning back to Figure 5, the order quantity constraints at $t_{n-1}$ are $Q^j_{n-1} \leq Q^j_{n-2}$ for each $j \in \Theta_{n-1}$. At $t_{n-2}$, we have three ordering constraints:

$$Q^{11}_{n-2} + Q^{12}_{n-2} + Q^{13}_{n-2} \leq Q^1_{n-3},$$
$$Q^{21}_{n-2} + Q^{22}_{n-2} + Q^{23}_{n-2} \leq Q^2_{n-3},$$
$$Q^{31}_{n-2} + Q^{32}_{n-2} + Q^{33}_{n-2} \leq Q^3_{n-3}.$$

We can then formalize the other order quantity constraints at $t_i$ as $Q^j_i \leq Q^j_{i-1}$ for $i \in \{2, \cdots, n-3\}$ and $j \in \Theta_i$. Finally, at $t = t_1$—that is, when the first proliferation occurs—we have $Q^1_1 + Q^2_1 + Q^3_1 \leq Q_0$.

We also define another set $\Upsilon^k_i$, which represents the set of end products produced by using component $k$ at echelon $i$. Therefore, $\Upsilon^k_i$ includes the end products (sold in the markets), whose availability depends on the order quantity decision of $Q^k_i$. In Figure 5, for example, the quantity $Q^1_1$ has a direct influence on the ordering decisions of the end products: $Q^{11}_{n-1}$, $Q^{12}_{n-1}$, and $Q^{13}_{n-1}$. Therefore, $\Upsilon^1_1 = \{11, 12, 13\}$. The set $\Upsilon_0 = \Theta_{n-1}$ since the quantity $Q_0$ has a direct influence on the final inventory of all end products. Let $p_j$ for $j \in \Theta_n$ denote the price of the products sold in the market.

To determine the maximum expected profit at $t_{n-1}$, we write the following stochastic programming (SP) model (Shapiro et al. 2009, Ch. 1):

$$\underset{Q^j_{n-1}, \forall j \in \Theta_{n-1}}{\text{Maximize}} \quad z = \sum_{j \in \Theta_{n-1}} p_j \mathbb{E}\Big(\mathcal{W}_j(Q^j_{n-1}, D^j_n)\Big) - c^j_{n-1} Q^j_{n-1} \tag{15}$$

subject to:

$$\sum_{j \in \Theta^k_{n-1}} Q^j_{n-1} \leq Q^k_{n-2}, \qquad \forall k \in \Theta_{n-2}, \tag{16}$$

$$Q^j_{n-1} \geq 0, \qquad \forall j \in \Theta_{n-1}, \tag{17}$$

where $\mathcal{W}_j(Q^j_{n-1}, D^j_n) = \min\{Q^j_{n-1}, D^j_n\}$ denotes the sales and $D^j_n$ is a random variable. Constraint (16) guarantees that the sum of the order quantities of the items in a set $\Theta^k_{n-1}$ is less than the amount of their parent item $k$. In the appendix, we provide the solution for the mathematical problem (15)–(17). Specifically, we transform the SP model into a linear programming (LP) model as demonstrated by Shapiro et al. (2009, Ch. 1–3). By analyzing the LP model and its dual, we partition the demand space and determine the shadow prices (see Van Mieghem (1998) for a similar method to solve an SP problem). We then proceed backwards in a similar fashion, using induction, and determine the optimal ordering policy for upstream echelons. The optimal policy is satisfied when all products in a set $\Theta^k_i$ (for all $k$ and $i$ values) have the same marginal value of ordering one additional unit. If the quantity $Q^k_{i-1}$ is highly restrictive, the marginal value for all products in the set $\Theta^k_i$ would have a positive value. If the quantity $Q^k_{i-1}$ is excessive, the marginal value would become zero. This analysis reveals the structure of the optimal policy as well.

Theorem 2. *The optimal ordering policy for all the items in each decision epoch is resource-constrained and state-dependent such that the optimal order quantity is found by:*

$$q^j_i = \begin{cases} Q^{j^*}_i & \text{if } \sum\limits_{j\in\Theta^k_i} Q^{j^*}_i < Q^k_{i-1}, \\ \widehat{Q}^j_i & \text{if } \sum\limits_{j\in\Theta^k_i} Q^{j^*}_i \geq Q^k_{i-1}, \end{cases} \tag{18}$$

*where $Q^{j^*}_i$ satisfies the following expression:*

$$g^{\Upsilon^j_i}_i(Q^j_i, \mathbf{D}^{\Upsilon^j_i}_i) = \int\limits_{\underline{D}^{\Upsilon^j_i}_{i+1}}^{+\infty} g^{\Upsilon^j_i}_{i+1}(Q^j_{i+1}, D^{\Upsilon^j_i}_{i+1}) f(D^{\Upsilon^j_i}_{i+1} | D^{\Upsilon^j_i}_i) \partial D^{\Upsilon^j_i}_{i+1} - c^j_i = 0 \qquad \forall j \in \Theta^k_i. \tag{19}$$

Equations (18) and (19) imply that the solution for the product proliferation model reduces to that of the single-product model (given by Theorem 1) when the quantity constraint is not binding (i.e., $\sum\limits_{j\in\Theta^k_i} Q^{j^*}_i < Q^k_{i-1}$). If the constraint is binding (i.e., $\sum\limits_{j\in\Theta^k_i} Q^{j^*}_i \geq Q^k_{i-1}$), the limited amount of inputs is allocated to the products by comparing their marginal profits, which are given by Equation (19).

With the characterization of the optimal policy at hand, we can use our framework to analyze the impact of the point of proliferation, costs, and lead times on profits. If everything else remains the same, delaying differentiation (moving any point $t_i$ with proliferation forward) is beneficial to the firm. The following proposition shows how the value of delaying differentiation is affected by the costs of downstream operations.

Proposition 4. *The value of delaying the point of product proliferation increases as the costs of downstream operations that take place after the point of proliferation increase.*

Proposition 4 underpins delaying differentiation by swapping costly operations that cause proliferation with less costly downstream operations. As documented by the Benetton's case, delaying the point of proliferation and the costly dyeing operation were both achieved by only swapping two operations. If such an improvement achieved by a single change is not possible, redesigning processes such that costly operations scheduled before the proliferation are swapped with less costly post-proliferation operations should precede any attempt to delay the proliferation point.

THEOREM 3. *Deferring the proliferation by swapping the point of proliferation with the adjacent downstream operation results in a decrease (increase) in profits if the cost of the downstream operation is higher (lower) than a threshold value:*

$$\kappa = c_p \mathscr{P}, \tag{20}$$

*where $c_p$ is the cost of the operation that causes the proliferation and $\mathscr{P} > 0$ is a pooling factor. The pooling factor is constrained to being greater than one (i.e., $\mathscr{P} > 1$) if the lead time of the operation that causes the proliferation is shorter than or equal to the lead time of the adjacent downstream operation.*

This theorem is built on the trade-off between delaying product proliferation and delaying a high-cost operation. Manufacturers may be exposed to such a trade-off when they redesign their operations. For the single-product model, Proposition 2.B shows that swapping a costly upstream operation with a less costly downstream operation increases the profits if the operations swapped have the same lead time. Therefore, postponing an operation helps increase the profits if its cost is higher than the cost of the downstream operation given that both operations have the same lead time. For the multi-product model, however, postponing the point of proliferation (by swapping it with a downstream operation) increases the profits if the cost of the downstream operation is lower than the threshold. In the online appendix, we explicitly develop an expression for this threshold value (Equation (86)). The value of $\kappa$ determines the value of postponing the proliferation. The higher the $\kappa$ value, the more appealing it is to postpone the proliferation.

When $\mathscr{P} > 1$, the cost of the downstream operation should be higher than the cost of the operation that causes the proliferation to prevent any attempt to defer the proliferation. However, $\mathscr{P}$ value can be less than one, which is possible only when the lead time of the operation that causes the proliferation is strictly longer than the lead time of the adjacent downstream operation. If $\mathscr{P} < 1$, delaying the proliferation may cause a profit loss even when the costs of the two operations are the same. In order to illustrate the intuition behind this finding, we consider a hypothetical case in which the lead time of the downstream operation is equal to zero and the lead time of the adjacent upstream operation is very long. Product proliferation also occurs during the upstream operation.

Swapping these two operations does not change the starting time of the product proliferation. However, it does expedite the costly downstream operation, which in turn reduces the expected profit.

Theorem 3 provides useful insights regarding the limits of deferring the product proliferation in a sequential production system. Historically the operations management literature on delayed differentiation has advocated for the postponement of product proliferation (see Zinn (2019) for the review of the literature). Despite the widespread benefits of the postponement strategy, manufacturers may still differentiate products in the early stages of the production. Some manufacturers even coordinate with their suppliers for joint product development, and the suppliers supply differentiated components to be used in the initial stages of production (Petersen et al. 2005). This strategy would be effective when non-differentiating operations are costlier and/or shorter than the operation causing the differentiation.

The explicit analytical form of the cost threshold makes it possible to analyze the impact of demand correlation, demand uncertainty, and the number of products on the value and cost of postponing product proliferation. In the following corollary, we demonstrate the impact of demand correlation on the cost threshold value.

COROLLARY 1. *Suppose that the correlation between any two products in the product portfolio is fixed and equal to $\rho$. The product proliferation takes place at a single echelon. The value of $\kappa$ decreases with the value of $\rho$.*

The corollary indicates that the value of postponing the proliferation increases as the products become more negatively correlated. This result is consistent with the existing literature on centralized inventory management (Eppen 1979) such that the value of inventory pooling increases as the correlation between the demand for the products decreases. Postponing the proliferation means that inventory is pooled for a longer time period; thus, its value is expected to increase as the correlation decreases. Next, we elaborate on the impact of demand uncertainty on the $\kappa$ value.

COROLLARY 2. *Suppose that the demand for each product is independent and identically distributed. The volatility parameter $\sigma$ (given by Equation (2) for the additive and Equation (4) for the multiplicative model) is the same for each product. Then, the value of $\kappa$ increases with $\sigma$.*

Corollary 2 is also consistent with Eppen (1979) such that the value of postponing the proliferation (i.e., inventory is pooled for a longer time period) increases with demand uncertainty. In the following corollary, we present the impact of the number of products in the portfolio on the value of $\kappa$.

COROLLARY 3. *Suppose that the demand for each product is independent and the cost of the operation that causes the proliferation is independent of the number of products. Then, the $\kappa$ value increases with the number of products.*

Similar to the first two corollaries, Corollary 3 is consistent with Eppen (1979) such that the value of postponing the proliferation increases with the number of products in the portfolio. In Corollary 3, we assume that the cost of operation at each echelon is fixed. This corollary can also be extended to the case in which the cost of operation that causes proliferation depends on the number of products.

Corollary 4. *The value of postponing the proliferation is amplified if the cost of the operation that causes the proliferation increases with the number of products.*

In the following section, we complement our analytical results with a case analysis based on the data from Kordsa to compare our approach with a theoretical benchmark of the repetitive newsvendor model. We also develop some strategies to establish operational responsiveness and cost efficiency and assess their performance under different circumstances.

## 6. Case Analysis

We calibrate our model using data from Kordsa Inc. (refs blinded to respect the peer review). Kordsa is a global manufacturer of advanced composites and reinforcement materials used in the automotive, aerospace, construction, and infrastructure industries. The company operates in the US, Brazil, Turkey, Indonesia, and Thailand with 11 facilities and around 4,500 employees. Kordsa is the global leader in the tire cord market such that one third of automotive tires and two thirds of aircraft tires are made from the cords manufactured by Kordsa. Over the last decade, Kordsa has been exposed to an increase in product variety as its customers demand more differentiated cords along different dimensions such as elasticity, strength, and thermal resistance.

Cord production involves four stages: (1) polymer yarn production, (2) twisting, (3) weaving, and (4) dipping operations. In the first stage, some chemicals are processed with polypropylene to produce polymer yarns. The polymer yarns are of a single type so there is no product proliferation in the first stage. In the second stage, polymer threads are twisted. The number of twists per meter has a direct impact on the strength and the elasticity of the cords. Kordsa differentiates twisted cords along the density of twists, measured as the number of twists per meter. Therefore, product proliferation occurs in the second stage. In the third stage, twisted yarns are woven to form the cords. The number of twisted cords used per meter square may differ depending on the technical features of the cord that are specified by Kordsa's customers. For that reason, product proliferation also occurs in the third stage. In the fourth stage, the cords are treated with a chemical blending to bring the products to the right level of thermal resistance and elasticity. The same cords can be treated differently, resulting in different SKUs. Thus, product proliferation can occur in the last stage. Although the polymer yarns produced in the first stage are of a single type, there is a large variety of end products due to product proliferation in the last three stages.

We use anonymized sales and order data of a product portfolio of 10 SKUs. Our dataset covers the time period starting from January 2017 and ending in December 2019, and it consists of 1,548 data rows. Each row specifies a unique sales transaction, including information such as product code, the date the order was received, the delivery date requested, and the quantity ordered. Total (pooled) demand that was requested to be delivered in each month from January 2017 to December 2019 is presented in Figure 6. To preserve the anonymity of the data, we randomly scale the true values. We split the data into two sets (i.e., the training and test sets). The training set covers the first two-year time period. We use this data to train our model and estimate the model parameters. Then, we use the last year's data to assess the performance of our model compared to a theoretical benchmark of the repetitive newsvendor model.

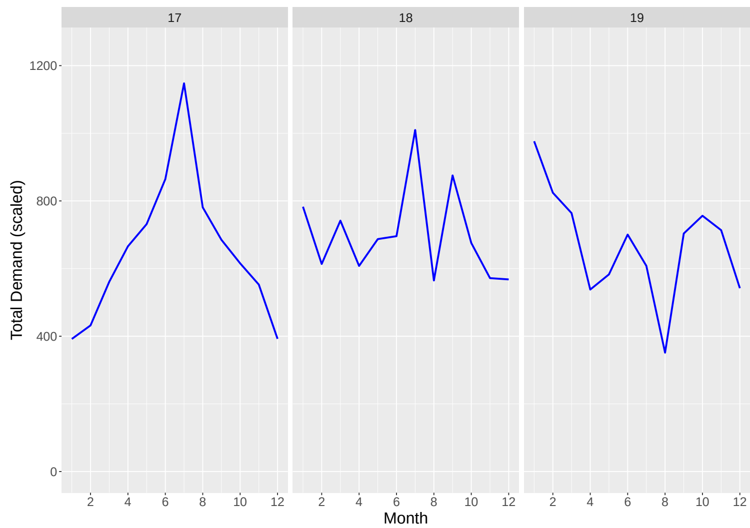


**Figure 6 Total demand requested to be delivered in each month**

Following the supply chain representation used in Figure 5, we present the product proliferation model of Kordsa in Figure 7. Polymer yarn is produced in the first stage, which is processed into five different components in the second stage. Four of them are not exposed to product proliferation in the following stages because each of them is transformed into a fast-selling SKU. The last component is further exposed to product proliferation in both the third and last stages. It is first processed into four sub-products in the third stage and then into six SKUs in the fourth stage.

We assume that it takes two months to complete the production such that each operation lasts a half month. We normalize the selling price of each SKU to \$1 per unit. We also assume that the total cost to produce one unit of each SKU is equal to \$0.5. We assign higher cost values to the first and last operations than the intermediate operations such that $c_0 = c_3 = \$0.15$ and $c_1 = c_2 = \$0.1$. This cost allocation makes it possible to compare alternative strategies that are discussed below in Subsection 6.3.

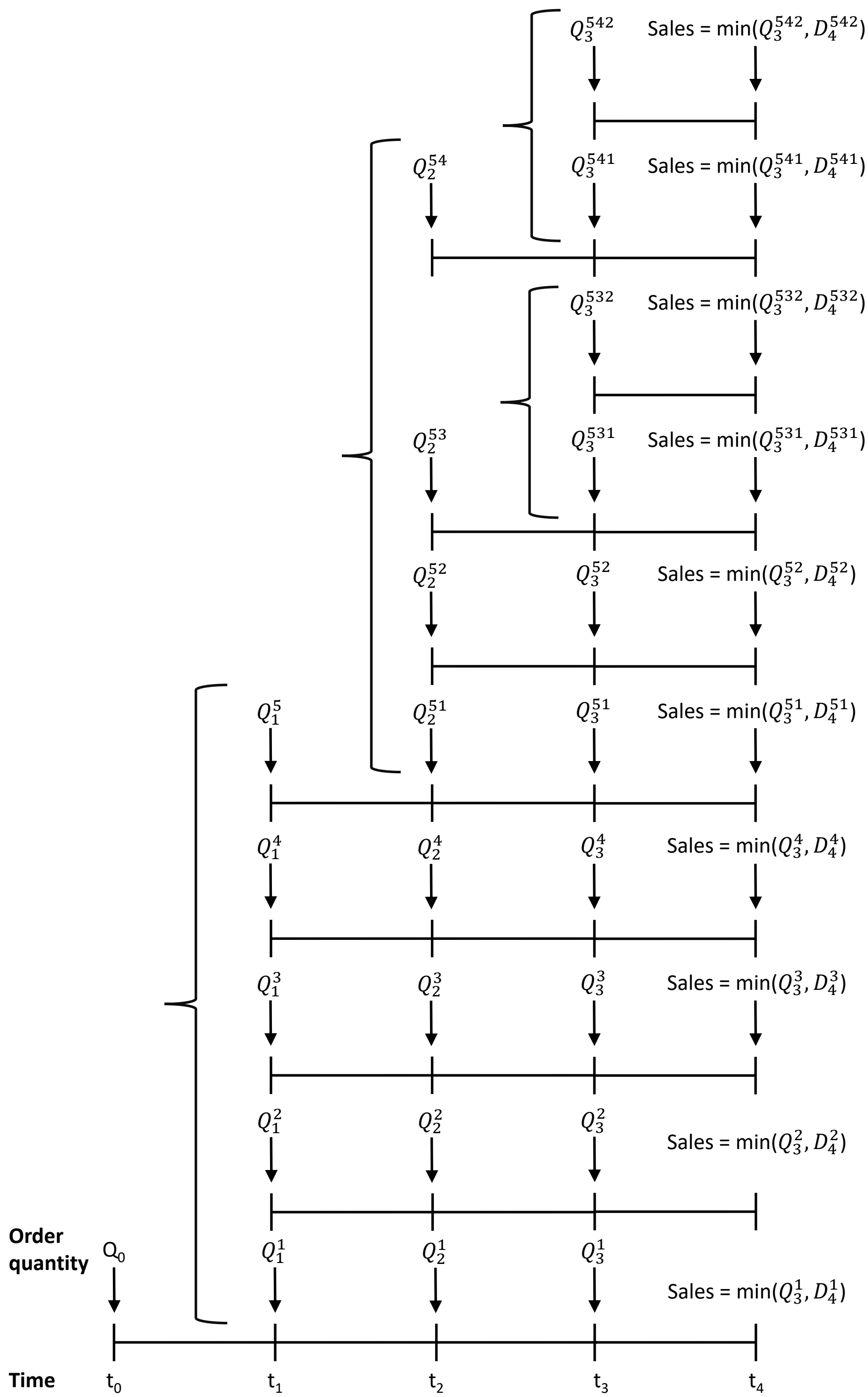


**Figure 7** **Supply chain structure for the case analysis**

In the following subsections, we first describe Kordsa's demand model, which is based on advance order information. Then, we present the results by comparing our model with a repetitive newsvendor model. Next, we change the model parameters and assess different strategies. Based on this assessment, we discuss the practical insights and develop a decision typology that shows effective operational strategies. Finally, we use some key modeling parameters to analyze the sensitivity of the model.

### 6.1. Demand Model at Kordsa Inc.

Kordsa receives orders from its customers in advance of the delivery dates. The time period that elapses from receiving a customer order until its delivery is referred to as the *demand lead time*. The demand lead time varies from a few weeks to a few months. Advance demand helps Kordsa better forecast actual demand, which is consistent with the literature. For instance, Fisher and Raman (1996) showed that even a small quantity of advance orders can help increase forecast accuracy substantially. The demand planning team considers advance demand as an important factor that highly correlates with actual demand.

Advance demand is collected in the form of firm commitments. Customers guarantee that the ordered amount is purchased in full on the pre-specified delivery day. The volume of advance demand is equal to $D_0$ at the beginning of the forecasting horizon—that is, at time $t_0$. Kordsa receives orders during the forecasting horizon such that advance demand accumulates until $t_n$. Therefore, $D_i$ denotes accumulated advance demand, which amounts to the actual demand at time $t_n$. The demand process is characterized by a submartingale model such that $\mathbb{E}(D_i|D_j) \geq D_j$ for $t_i \geq t_j$. We use the multiplicative model with a positive drift rate given by Equation (4) to capture these dynamics.

Equation (4) implies that the volume of demand that accumulates from $t_i$ to $t_n$ depends on the amount of firm orders already received from the customers (i.e., $D_i$). The dependency between the firm orders and future demand is not captured in an additive model. However, we observe such a dependency structure in both literature and practice when the forecast horizon is relatively long. Fisher and Raman (1996) report that the amount of firm orders highly correlates with future demand. The same dynamics are also observed at Kordsa. For that reason, we use the multiplicative process given by Equation (4) to model the forecast evolution process at Kordsa.

We estimate the $\mu$ and $\sigma$ parameters of each SKU by fitting a normal distribution to $\ln(D_4/D_0)$ values in the training set. The mean value of the fitted distribution is set equal to $\mu - \sigma^2/2$, and the standard deviation is set equal to $\sigma$. We drop the time element for parameter estimation because the length of the planning horizon is normalized to one. The advance orders are collected from the customers homogeneously over the planning horizon. Therefore, the $\mu$ and $\sigma$ values for each SKU are fixed. We note that if advance orders are clustered within a specific time duration (e.g., $t_4 - t_3$), the model parameters would be different for each echelon. In such a case, the parameters for each echelon can be estimated by fitting a normal distribution to $\ln(D_i/D_{i-1})$ for $i \in \{1, \cdots, 4\}$.

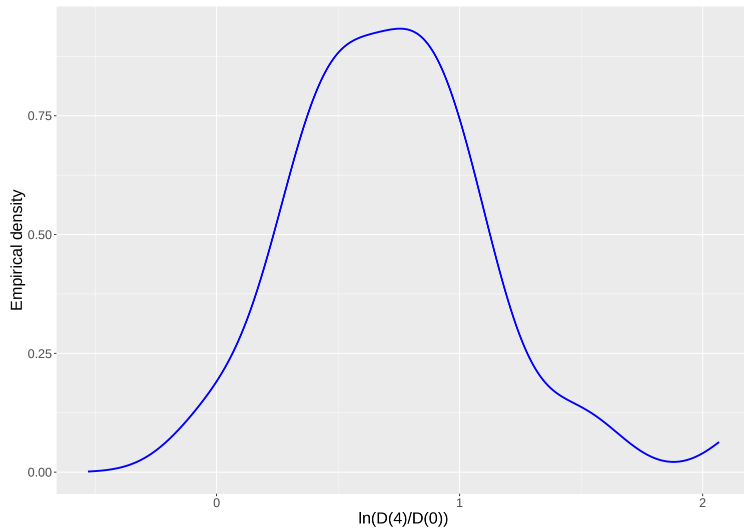


**Figure 8 The empirical density of $\ln(D_4/D_0)$ values for the selected SKU**

For each SKU, we test for normality of $\ln(D_4/D_0)$ values according to the Kolmogorov-Smirnov test. The null hypothesis that the empirical values follow a normal distribution is statistically true for all the SKUs. We randomly select an SKU and show its empirical distribution in Figure 8, which resembles the bell-shape curve of the normal density function.

### 6.2. Results

After estimating the $\mu$ and $\sigma$ values for all the SKUs, we use Theorem 2 to calculate the order quantities at the beginning of each echelon. At the beginning of each month over the test time window, we obtain the pooled advance demand that has to be delivered within the next month, which is equal to $D_0$. The first production order (i.e., $Q_0$) is also placed to fulfil the demand requested to be delivered in the next month. A half month after placing the first production order, the first process is completed and the production orders for the second operation are placed. At that time, the advance demand values for the subcomponents are calculated from the test dataset. The second process is completed after another half month, and then the production orders for the third operation are placed. Final products are produced after completion of the fourth process, which occurs two months after starting the first process. If the first process is started at the beginning of month $k$, the whole production is completed and the final products are obtained at the end of month $k+1$. Then, the demand for each SKU requested to be delivered within month $k+1$ is fulfilled from inventory. We assume that excess inventory at the end of each month is salvaged and unmet demand is lost.

We compare our approach with a theoretical benchmark of a repetitive newsvendor model. In the benchmark model, the newsvendor quantities for the SKU $j$ are calculated at time $t_i$ for $i \in \{0,1,2,3\}$ by the critical-fractile formula:

$$Q_i^j = D_i^j \exp(\mu_j(t_4 - t_i) + z_i \times \sigma_j \sqrt{t_4 - t_i}), \tag{21}$$

where $z_i$ is found by the inverse of the standard normal distribution:

$$z_i = \Phi^{-1}\Big[\Big(p - \sum_{j=i}^{3} c_j\Big)/p\Big]. \tag{22}$$

If the quantity found by Equation (21) is less than the order quantity in the previous echelon (i.e., $t_{i-1}$), the order quantity at time $t_i$ becomes equal to that of Equation (21). Otherwise, the order quantity is constrained by the order quantity in the previous echelon. In a mathematical form,

$$q_i^j = \begin{cases} Q_i^j, & \text{if } Q_i^j \leq q_{i-1}^j, \\ q_{i-1}^j, & \text{otherwise.} \end{cases} \tag{23}$$

The profit values for both our approach and the benchmark model are calculated by subtracting total production cost from total revenue (i.e.,$p \sum_{j \in \Theta_4} \min(D_4^j, q_3^j)$). Figure 9 presents the results over the test time window. The solid curve gives the profit values for our approach, whereas the dashed curve shows the results for the benchmark model. Our product proliferation model outperforms the benchmark model for ten months. However, the benchmark model performs slightly better than our approach for two months: the tenth and eleventh months. The average profit is equal to \$126.5K based on our approach and \$53.7K for the benchmark model. Therefore, our approach results in 135% higher profits than the benchmark model.

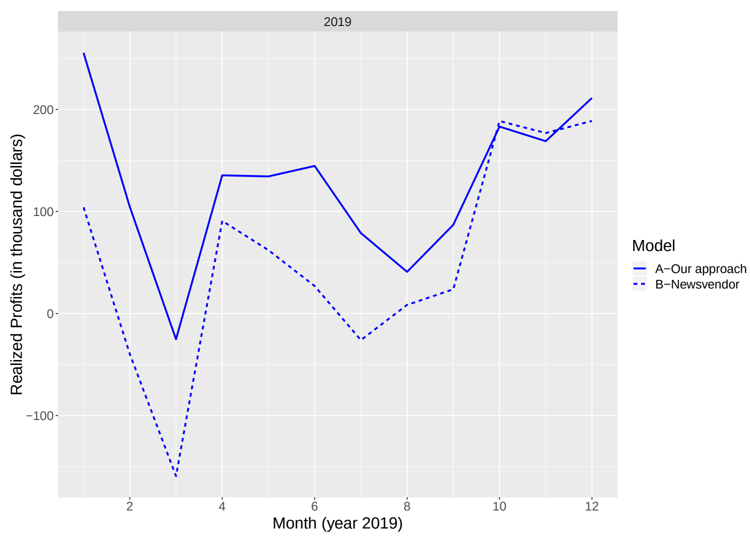


**Figure 9 Comparison of profits between our approach and the repetitive newsvendor model over the test period**

The repetitive newsvendor model calculates the order quantities at each decision epoch using the critical-fractile solution given by Equation (22). Hence, the newsvendor approach does not take into account the possibility of reducing the order quantity in the subsequent stages. Our approach corrects for this misalignment and determines the order quantities at each decision epoch by considering the possibility that order quantities are updated backward in the following decision epochs. For that reason, our approach outperforms the repetitive newsvendor model.

The newsvendor model takes a more static approach than ours, which makes demand forecasting more important than modelling the evolutionary dynamics of demand forecasts. In a dynamic approach, similar to ours, modelling the evolutionary dynamics becomes more important than increasing the forecast accuracy at the very beginning (Elmachtoub and Grigas 2021). Suppose, for example, a manufacturer places monthly production orders for a production period of two months in a single-echelon setting. The monthly capacity is equal to 300 units. Demand for the product during a short selling season is between 100 and 500 units. If demand uncertainty is resolved one month after starting the production and the profit margin is high, the manufacturer would order 200 units for the first month and determine the order quantity for the second month after observing the updated demand forecast. Unless the demand turns out to be less than 200 units, the manufacturer perfectly matches supply with demand. The manufacturer also bears the risk of ending up with excess inventory (when demand becomes less than 200) given the high profit margin.

If demand uncertainty is resolved one month after starting the production and the profit margin is low, the manufacturer would order 100 units for the first month and determine the order quantity for the second month after observing the demand forecast. Unless the demand turns out to be more than 400 units, the manufacturer perfectly matches supply with demand. The manufacturer also takes the risk of ending up with lost sales (when demand becomes more than 400) given the low profit margin.

In this example, high demand uncertainty at the very beginning does not have any severe impact on the first ordering decision due to the availability of the second ordering opportunity. Thus, the resolution of demand uncertainty until the second ordering epoch would be more important than marginally improving the demand forecast at the very beginning. The demand uncertainty resolution can be achieved by collecting information from the customers (e.g., advance orders). If this is not possible, demand uncertainty at the very beginning has a huge impact on the ordering decisions. In this case, a dynamic approach does not offer any substantial improvement over the newsvendor approach. Therefore, the manufacturer would use sophisticated statistical models (e.g., time series) with historical data to improve the forecast accuracy at the very beginning.

### 6.3. Analysis of Alternative Strategies

Our analytical results offer useful insights regarding the operational strategies that can be employed to improve the bottom-line performance of supply chains with product proliferation. We now synthesize these strategies and compare their relative performance.

We consider two types of supply chains: (1) inflexible supply chains where operations cannot be swapped and (2) flexible supply chains where operations can be swapped. We begin our analysis with the non-flexible supply chains. Propositions 2 and 4 state that reducing the cost is more beneficial at an upstream echelon than at a downstream echelon. Thus, manufacturers can improve profits by systematically reducing costs, starting from upstream operations and then moving downstream. We call this strategy *systematic cost reduction*. Since upstream operations are often related to procurement of raw materials or subassemblies, systematic cost reduction calls for prioritizing the improvement of the procurement efficiency via consolidating purchasing orders, creating purchase bundles, using low-cost substitutes, or other policies (Paranikas et al. 2015). For inflexible supply chains, delaying the proliferation is only possible via the lead-time reduction. Proposition 3 corroborates that manufacturers should try to reduce the lead time of downstream operations first and then move upstream in the supply chain. We call this strategy *systematic lead-time reduction*.

When there is some flexibility to adjust the processes, manufacturers redesign the operations by changing their sequence. Such process redesigning practices can also be complemented with the lead-time and cost-reduction approaches to maximize the profit. We propose two different process-redesign strategies: (1) performance-based process redesign and (2) mixed strategy. To describe these strategies, we define six different conditions regarding any two adjacent operations.

**C.1 :** Differentiation occurs during the upstream operation and no differentiation occurs during the downstream operation.

**C.2 :** No differentiation occurs during both operations.

**C.3a:** The cost of the upstream operation is greater than or equal to the cost of the downstream operation.

**C.3b:** The cost of the upstream operation is strictly greater than the cost of the downstream operation.

**C.4a:** The lead time of the upstream operation is shorter than or equal to the lead time of the downstream operation.

**C.4b:** The lead time of the upstream operation is strictly shorter than the lead time of the downstream operation.

Theorem 3 affirms that manufacturers can improve profits through a *performance-based process redesign* strategy, which iteratively examines adjacent operations and swaps the orders if one of the followings is satisfied:

- [C.1 and C.3a and C.4a]
- or [C.2 and C.3a and C.4b]
- or [C.2 and C.3b and C.4a]

If, for example, there are two adjacent operations (i.e., $i$ and $i+1$) with the same lead time such that $c_i \geq c_{i+1}$ and the differentiation occurs only during operation $i$, swapping these two operations to postpone operation $i$ later than operation $i+1$ leads to an increase in profits. The proof of Theorem 3 shows that $\mathscr{P} > 1$ in this case. Because $c_i \geq c_{i+1}$, the cost of the downstream operation (i.e., $c_{i+1}$) does not exceed the threshold value $\kappa$, which is even greater than $c_i$ for $\mathscr{P} > 1$. For that reason, swapping the operations to postpone the proliferation results in a profit increase. If no differentiation occurs, postponing a high-cost and short-lead-time operation later than a low-cost and long-lead-time operation leads to a profit increase. No swapping should occur if none of the three conditions are met. Performance-based process redesign strategy is practical in the sense that it does not require the calculation of the cost threshold value. Instead, it focuses on certain cases that render swapping the operations profitable.

If differentiation occurs during a low-cost or long-lead-time operation, performance-based process redesign may induce decision makers to differentiate products in the early stages of supply chains. In such cases, we augment process redesign with lead-time reduction to postpone high-cost and high-differentiation processes at the same time. We refer to this strategy as the *mixed* strategy, which first sorts operations according to their costs in an ascending order. Therefore, the low-cost operations precede those with higher costs. After sequencing the operations in this way, the lead times are reduced starting from the downstream operations.

We now limit our attention to two fast-selling products of Kordsa. The structure of the chain is shown in Figure 10 such that product proliferation occurs during the second operation. We refer

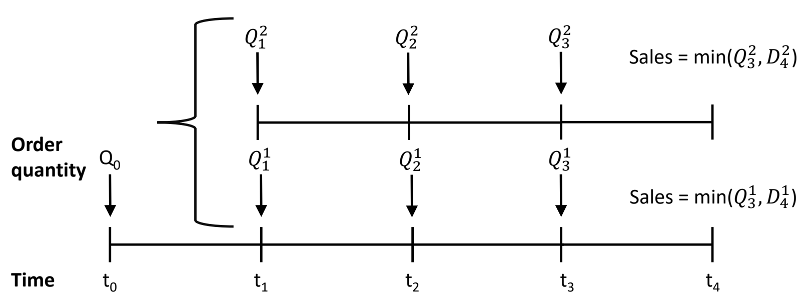


**Figure 10** **Base case: Supply chain structure for two fast-selling products**

to this structure as the *base case*. We recall that the second and third operations have the same cost values. If the performance-based process redesign is implemented over the base case, the only change that needs to be made is to swap the second and third operations. Its structure is given in Figure 11. This leads to an increase in profits. We use $\Pi_b$ and $\Pi_p$ to denote the profit values under the base case and the performance-based process redesign, respectively. Thus, $\Pi_p > \Pi_b$.

If the mixed strategy is implemented, the processes are ordered in the sequence given in Figure 12, and the lead time of the last operation is reduced by $\Delta t_4$. The value of $\Delta t_4$ is determined in such a way that the profit under the mixed strategy, denoted by $\Pi_m$, becomes equal to $\Pi_p$. Since we aim to compare four different strategies (i.e., performance-based process redesign, mixed strategy, systematic lead-time reduction, and systematic cost reduction), we design the structures of the supply chains for these four strategies such that the realized profits are the same with the original

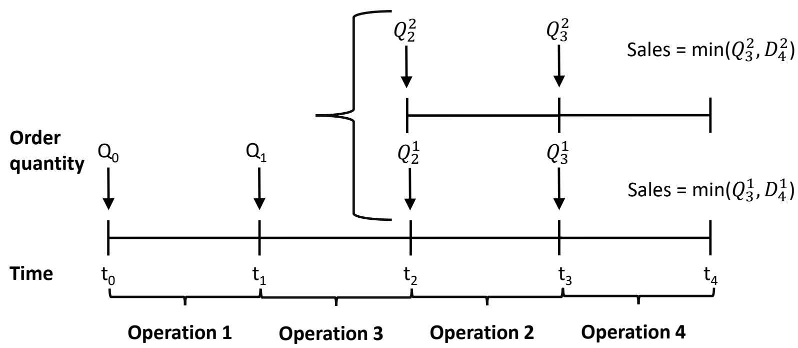


**Figure 11** **Performance-based process redesign: The second and the third operations are swapped**

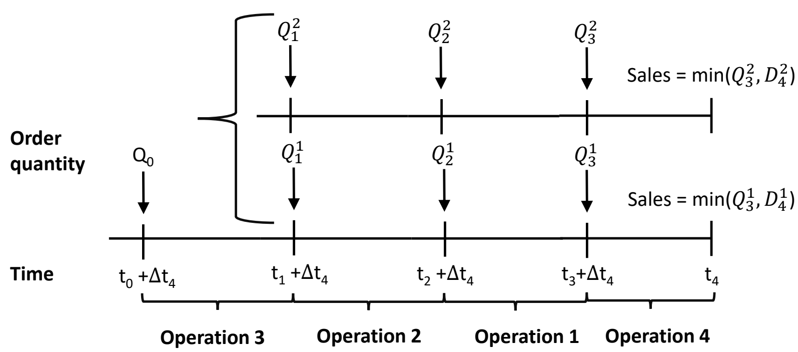


**Figure 12** **Mixed strategy: Operations are sequenced from the lowest cost to the highest and the lead time of the last operation is reduced by $\Delta t_4$**

parameter values. We then change the price and volatility parameters to assess the performance of each strategy for the innovative and standard products.

Under the systematic lead-time reduction strategy, the sequence of the operations is not changed. We only reduce the lead time of the last operation by $\overline{\Delta} t_4$ such that the profit under this strategy, denoted by $\Pi_l$, becomes equal to $\Pi_l = \Pi_p$. Under systematic cost reduction, the cost of the first operation is reduced by $\Delta c_1$ such that the profit under this strategy, denoted by $\Pi_c$, becomes equal to $\Pi_c = \Pi_p$.

| Strategy | Expected Profit | Key Parameter Values |
|---|---|---|
| Performance-based process redesign | $\Pi_p = \$104.5K$ | Cost and lead time values are the same as the base case. The only difference is that proliferation takes place at time $t_2$. |
| Mixed strategy | $\Pi_m = \$104.5K$ | $c_1 = c_2 = 0.1$, $c_3 = c_4 = 0.15$, and $\Delta t_4 = 0.0906$ such that $t_4 = 0.1594$. |
| Systematic lead-time reduction | $\Pi_l = \$104.5K$ | $\overline{\Delta} t_4 = 0.09502$ such that $t_4 = 0.15498$ |
| Systematic cost reduction | $\Pi_c = \$104.5K$ | $\Delta c_1 = 0.02671$ such that $c_1 = 0.12329$ |

**Table 1** **Break-even structures of all four strategies**

We generate 1,000 sample paths of the evolution of advance demand values and calculate the expected profit by taking the average of the realized profits for these sample paths. Therefore, we do not use the test dataset because we compare the strategies based on the randomly generated sample paths. For the base case in Figure 10, the average profit obtained is \$96.9K (i.e, $\Pi_b = \$96.6K$). When the structure of the supply chain is switched to that seen in Figure 11 (i.e., performance-based process redesign), the average profit increases by 8.17% to \$104.5K. The mixed strategy with $\Delta t_4 = 0.0906$ yields the same expected profit as the performance-based strategy such that $\Pi_m = \Pi_p = \$104.5K$. Likewise, a systematic cost-reduction strategy with $\Delta c_1 = 0.02671$ results in the same expected profit. Finally, we determine the equivalent systematic lead-time reduction strategy such that reducing the lead time of the last operation by $\overline{\Delta} t_4 = 0.09502$ leads to the same profit. The break-even structures of all four strategies are given in Table 1.

After determining the break-even parameters for each strategy, we first increase the price by 30% and the volatility by 50% for each product. Following the Fisher's product matrix (Fisher 1997), we use these new parameter values to investigate the impact of each strategy on an innovative product portfolio. For each sample path, we dynamically find the order quantities at each echelon and calculate the realized profit depending on the evolution of demand. Then, the average profit values are found for each strategy as follows:

$$\Pi_m = \$172111 > \Pi_l = \$170914 > \Pi_p = \$168978 > \Pi_c = \$166921. \tag{24}$$

The difference between any two values of profits is statistically significant based on the two-sample t test.

We then reduce the price by 30% and the volatility by 50% for each product to analyze the impact of each strategy on a standard product portfolio. Following the same approach above, we generate 1,000 sample paths of the evolution of advance demand values and find the average profit values. For the standard product portfolio, the average profits are as follows:

$$\Pi_p = \$49021 > \Pi_c = \$47488 > \Pi_m = \$43831 > \Pi_l = \$43701. \tag{25}$$

The values of $\Pi_m$ and $\Pi_l$ are very close such that their difference is not statistically significant based on the two-sample t test. However, the other pairs are statistically different.

The strategic value that can be harvested by manufacturers from adopting these four strategies depends naturally on characteristics of both the industry in which they operate as well as the markets their products serve. When manufacturers do not have the process flexibility to conduct re-sequencing or major process changes, it indicates that they are in relatively mature industries where the processes are standardized and widely adopted. Such manufacturers have to rely on established templates for producing their products. To improve profits, they can conduct systematic cost reduction and/or systematic lead-time reduction because these two strategies do not necessitate re-sequencing processes. Naturally, performance-based process redesign and mixed strategy are not possible for them due to process inflexibility. Manufacturers using propriety processes to produce their products, however, may have the flexibility to change the sequence of their operations and redesign their processes based on cost and lead-time parameters. Therefore, such manufacturers can implement any of the four strategies to improve profits.

We develop a decision typology in Figure 13 that shows effective operational strategies depending on product characteristics and process flexibility. We categorize manufacturers along two dimen-

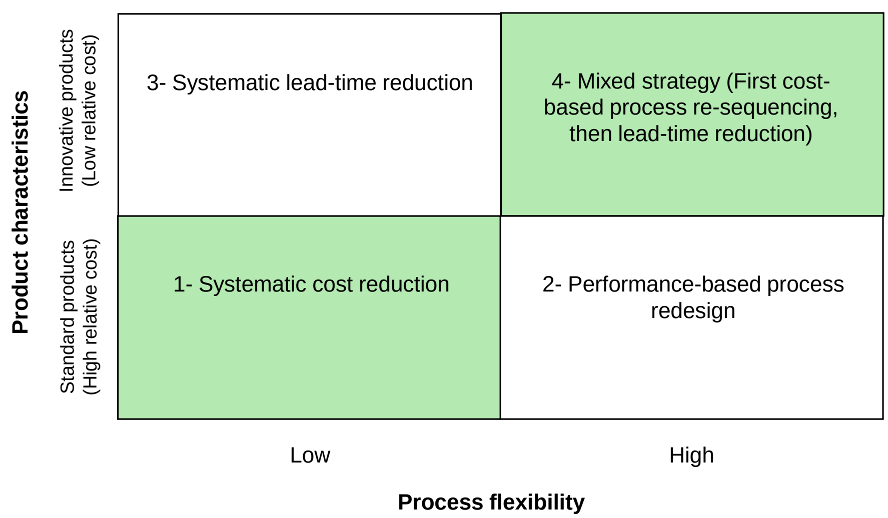


**Figure 13 Decision typology based on product characteristics and process flexibility**

sions. The first dimension is related to product characteristics. Analytical expressions (24) and (25) show that the performance of each strategy depends on the product type. Therefore, we consider two product types: (1) standard and (2) innovative. Manufacturers' flexibility to implement these strategies depends on the process flexibility. For that reason, the second dimension of our categorization is process flexibility. Our categorization of manufacturers is similar to that developed by Ferdows et al. (2016), who categorize manufacturers based on their product characteristics and process flexibility.

*1-Bottom-left quadrant (Systematic cost reduction):* These manufacturers produce commodity-like products using industry-standard production methods. Some business units of chemical companies (e.g., DuPont, BASF, etc.) producing commodity-like products fall into this category. Such manufacturers cannot change the sequence of operations; hence, it is not possible for them to conduct a performance-based process redesign or mixed strategy. Given that the products are sold at a low margin, costs represent a significant portion of revenues and the cost of raw materials may constitute up to 80% of total revenue. Analytical expression (25) indicates that systematic cost reduction is the most effective way to improve the bottom line. Therefore, it is not uncommon in such industries that manufacturers try to reduce upstream costs by pressuring their suppliers.[3]

*2-Bottom-right quadrant (Performance-based process redesign):* Some manufacturers excel in process flexibility while producing standard products. Ferdows et al. (2016) give an example of a US-based steel manufacturer producing steel rolls that established process flexibility through some advanced processes. Analytical expression (25) shows that performance-based process redesign leads to the highest increase in profits for standard products. Manufacturers in this category can adopt this strategy to cope with product proliferation owing to their process flexibility. This was certainly the case for Benetton when it resequenced its operations to postpone the costly dyeing operations for its highly standardized sweaters. As discussed in the introduction, this strategy helped the company increase its profits and market share significantly.

*3-Top-left quadrant (Systematic lead-time reduction):* Manufacturers with strong brands, such as some fashion-apparel manufacturers and pharmaceutical companies fall into this group (Ferdows et al. 2016). Although standard processes are used in production, the innovative/fashionable nature of the product and the brand value allow premium pricing and generate higher margins. As standardized processes leave little room for restructuring, lead time is the only lever for managing proliferation. Like Zara, manufacturers in this category should systematically reduce lead times to delay differentiation and thereby improve responsiveness and profits.

*4-Top-right quadrant (Mixed strategy):* Manufacturers with proprietary products and processes, such as Intel, can differentiate themselves through both product design and process technology (Ferdows et al. 2016). Their products are sold in the market at a high margin, which makes lead-time reduction appealing. With process flexibility, redesigning the processes to postpone high-cost activities may also be possible, which amplifies the value generated by reducing lead times. Analytical expression (24) indicates that manufacturers in this category are ideally suited for following the mixed strategy of coupling cost-based process re-sequencing with lead-time reduction. ASML, a Dutch company producing modular lithography systems for semiconductor manufacturers, has implemented this strategy as part of its value-sourcing initiative (van Rooy 2010). The company postponed the operations that required expensive components to a later stage in production, and reduced their sourcing lead times by paying the suppliers premiums.[4] This mixed strategy enables ASML to delay both the point of proliferation and high-cost operations.

Kordsa does not have the flexibility to change the sequence of the operations. For example, the first step of cord manufacturing is yarn production, which cannot be swapped with any other operation.

[3] https://www.mckinsey.com/industries/chemicals/our-insights/pursuing-purchasing-excellence-in-chemicals

[4] See also http://www.co-makers.com/gsls09/Buyers%20Breakfast%20GSLS09%20ASML.pdf

The company also has a large product portfolio so there are both standard and innovative products. For that reason, the principles of systematic cost reduction and systematic lead-time reduction have been used to improve the performance. To achieve cost reduction starting from the upstream operations, Kordsa internalized the yarn production through an efficient and continuous process. To reduce the lead time starting from the downstream operations, Kordsa increased the speed of the dipping and weaving operations over years. The variety of chemical blends has been increased over years to reduce the complexity in the earlier stages and postpone the proliferation to the last stage.

### 6.4. Sensitivity Analysis

The alignment of the operational strategies with the product and process characteristics given by Figure 13 is affected by key parameters such as volatility, correlation between the demand values for the products, and the number of products in the product portfolio. We now turn our attention to the performance assessment of the four operational strategies when we vary the values of the key parameters. We take the break-even structure in Table 1 as the basis of our analysis because the expected profit realized by employing each strategy is the same in that case. We then vary one of the parameters and evaluate the profit values for different values of the parameter. To calculate the expected profit, we generate 1,000 sample paths and take the average of the profits.

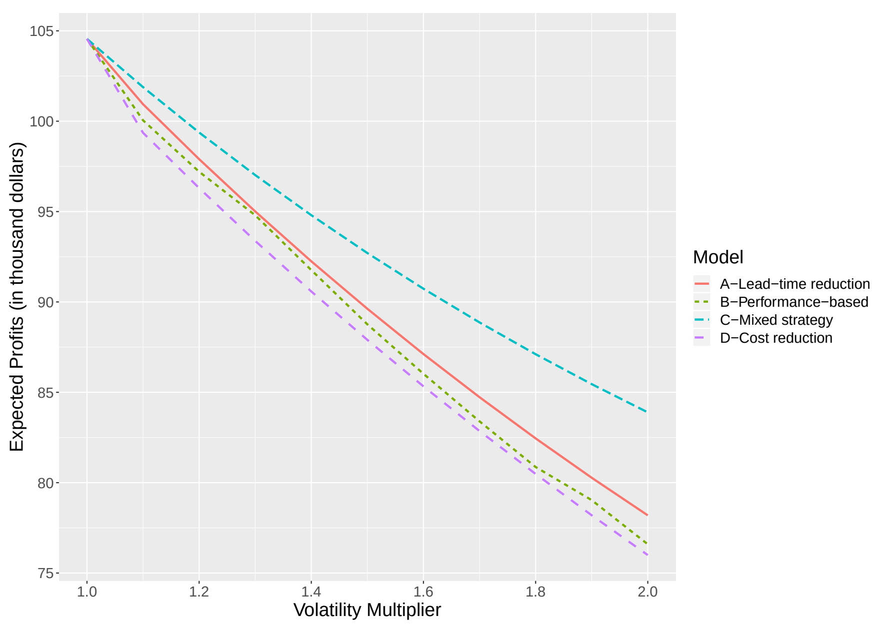


**Figure 14 Assessment of the strategies for varying volatility**

Figure 14 depicts the profit values for each strategy for the varying volatility multiplier. In particular, we multiply the volatility parameter $\sigma$ with a multiplier ranging from one to two. The x-axis represents the value of the multiplier, and the y-axis represents the expected profit. The red solid curve represents the lead-time reduction strategy, the dark green dashed line the performance-based strategy, the blue dashed line the mixed strategy, and the purple dashed line the systematic cost reduction. The multiplier value being equal to one corresponds to the break-even structure such that profit values are the same for each strategy. As the volatility increases, the expected profit decreases for all the strategies. The steepest decline is observed for the systematic cost reduction, and the mixed strategy outperforms the other three strategies for high volatility values.

Corollary 2 indicates that the value of postponing proliferation increases with volatility. Therefore, we expect the systematic cost reduction to underperform the other three strategies for high volatility values because it is the only strategy in which proliferation is not delayed. That the lowest profit is observed for the systematic cost reduction is thus consistent with Corollary 2.

The mixed strategy takes a balanced approach to manage product variety and demand uncertainty. Although the point of proliferation is not delayed as much as the performance-based strategy, it sequences the operations according to the cost values and postpones not only the proliferation point but also the other ordering decisions. Owing to this balanced approach, the mixed strategy outperforms other strategies for high demand uncertainty.

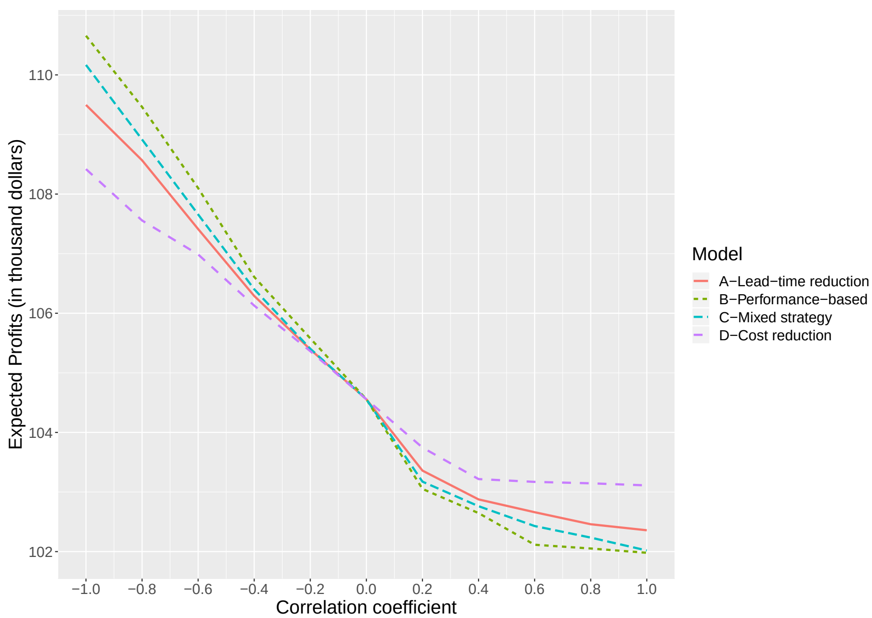


**Figure 15** **Assessment of the strategies for varying correlation**

Figure 15 shows the profit values for varying correlation values. The x-axis represents the correlation between the demand values of the two products, and the y-axis represents the expected profit values. We use the correlation value to generate the sample paths and calculate the order quantities and profit values accordingly. The case with a correlation value of zero corresponds to the break-even case in Table 1. As the correlation decreases, the expected profit for each strategy increases. For negative correlation values, the lowest profits are observed for the systematic cost reduction. This results from the fact that the proliferation is not delayed for the systematic cost reduction and the value of postponing the proliferation decreases with the correlation. This result is also consistent with Corollary 1.

As given in Table 1, the proliferation is delayed by 25% of total duration for the performance-based strategy compared to the base case, whereas this value reduces to 9.06% and 9.52% for the mixed and lead-time reduction strategies, respectively. Therefore, the performance-based strategy yields the highest profits for negative correlation values.

We remark that the analysis in Figure 15 is based on the assumption that demand correlation is exogenously given. However, the correlation values may be affected endogenously by product availability such that a customer who is exposed to a stock-out would switch to another product. In this case, a strong negative correlation is observed between the products that are out of stock frequently (see Honhon et al. (2010) and Netessine and Rudi (2003)). While we do not analyze endogenous factors, we conjecture that the benefits of product substitution is analogous to the benefits of shrinking the product portfolio while keeping the sales volume the same. Following Corollary 3 and the next sensitivity analysis, we expect that the benefits of postponing the proliferation decrease as the products become more substitutable.

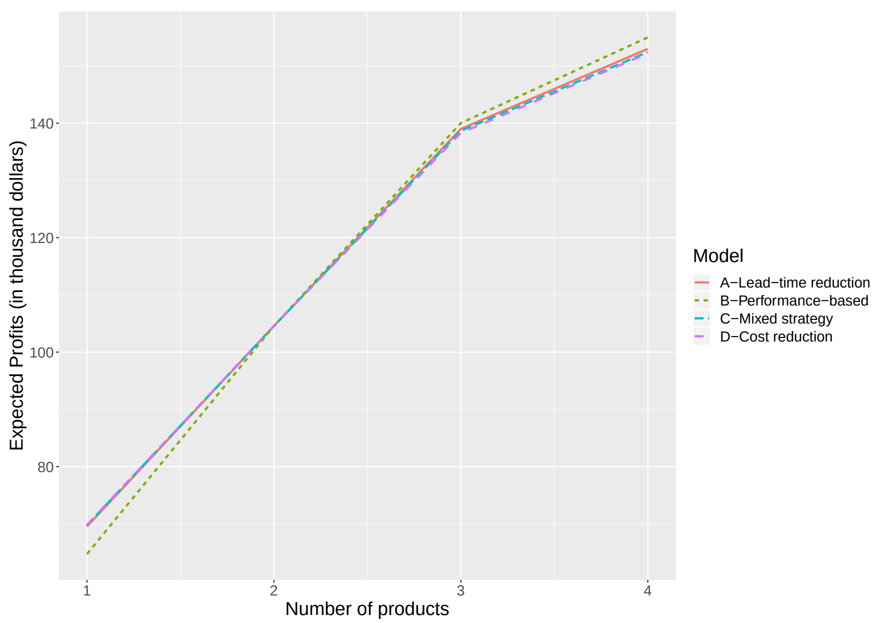


**Figure 16** **Assessment of the strategies for different numbers of products**

In Figure 16, we present profit values for different numbers of products. The products are selected from Kordsa's portfolio based on their sales volume. We select the first four products in Table 7 and sort them in descending order according to their total sales during the training period. Note that all products have the same product margin. We then add the products in sequence to the set of our analysis and calculate the expected profits. The case with two products corresponds to the break-even structure. As we increase the number of products, the performance-based strategy outperforms the other strategies because the proliferation takes place at a later time epoch under the performance-based strategy than those of the other strategies.

The results in this section provide useful insights regarding the boundaries of the decision typology in Figure 13. The typology uses a basic categorization of products and processes, and it aligns each category with an operational strategy. However, the performance of each strategy is affected by the key modeling parameters. Systematic cost reduction focuses on reducing the cost in the early production stages, and it would be effective for low volatility, positively correlated, and small product portfolios. The performance-based strategy defers the proliferation later than any other strategy because it primarily focuses on delaying the proliferation point. This strategy would be more effective than the others when the number of products in the portfolio increases or the demand values become more negatively correlated. The mixed strategy aims to delay the proliferation to some extent, but it also focuses on delaying the other decision epochs and redesigning the processes based on the cost values. Therefore, it takes a more holistic approach to improve the profits. When the volatility is increased, the benefits of this holistic approach exceed the benefits of solely focusing on postponing the proliferation such as occurs in the performance-based strategy.

## 7. Concluding Remarks

In this paper, we develop an analytical model to quantify the impact of supply chain structure along with the cost, demand, and lead-time parameters on profits in a multi-echelon and multi-product newsvendor model with product proliferation occurring at pre-specified echelons. In such a setting, decision makers can improve profits by establishing the responsiveness and/or cost efficiency. Delaying proliferation helps decision makers establish the responsiveness, which can be implemented through lead-time reduction or process redesign. Establishing the cost efficiency is also possible through systematic cost reduction or by postponing high-cost operations until there is partial or full resolution of demand uncertainty. Utilizing our analytical framework, we develop a decision typology that shows effective strategies depending on the product and the process characteristics.

Our model inherently assumes a make-to-stock supply chain with positive lead times for production stages but zero promised lead time for customers (i.e., maximum length of time in which a customer order is guaranteed to be delivered). When a customer is willing to wait, the manufacturer can quote a positive promised lead time at a discounted price and follow a combination of make-to-order and make-to-stock policies—that is, creating a decoupling point in the supply chain. Reducing lead time in this context could possibly help companies delay differentiation after the decoupling point, so product proliferation would take place after getting firm customer orders, completely eliminating inventory risk at downstream echelons. We believe that the trade-off between completely eliminating the downstream inventory risk and profit losses due to offering price discounts for longer promised lead times would be an interesting avenue of future research that requires incorporation of lead-time quotation and product proliferation models.

## References

Atalı A, Özer Ö (2012) Stochastic multi-item inventory systems with markov-modulated demands and production quantity requirements. *Probability in the Engineering and Informational Sciences* 26(2):263–293.

Aviv Y, Federgruen A (2001a) Design for postponement: A comprehensive characterization of its benefits under unknown demand distributions. *Operations Research* 49(4):578–598.

Aviv Y, Federgruen A (2001b) Capacitated multi-item inventory systems with random and seasonally fluctuating demands: Implications for postponement strategies. *Management Science* 47(4):512–531.

Biçer I, Hagspiel V, de Treville S (2018) Valuing supply-chain responsiveness under demand jumps. *Journal of Operations Management* 61:46–67.

Biçer I, Seifert RW (2017) Optimal dynamic order scheduling under capacity constraints given demand-forecast evolution. *Production and Operations Management* 26(12):2266–2286.

Brynjolfsson E, Hu Y, Simester D (2011) Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales. *Management Science* 57(8):1373–1386.

Cao J, So KC (2016) The value of demand forecast updates in managing component procurement for assembly systems. *IIE Transactions* 48(12):1198–1216.

Caro F, Martínez-de Albéniz V (2015) Fast fashion: Business model overview and research opportunities. Agrawal N, Smith SA, eds., *Retail Supply Chain Management: Quantitative Models and Empirical Studies*, 237–264 (Boston, MA, US: Springer).

de Treville S, Shapiro RD, Hameri AP (2004) From supply chain to demand chain: the role of lead time reduction in improving demand chain performance. *Journal of Operations Management* 21(6):613 – 627, ISSN 0272-6963, URL `http://dx.doi.org/https://doi.org/10.1016/j.jom.2003.10.001`.

Diermann C, Huchzermeier A (2017) Case–canyon bicycles: Judgmental demand forecasting in direct sales. *INFORMS Transactions on Education* 17(2):63–74.

E2Open (2018) Forecasting and inventory benchmark study. *Corporate Report* Available online: `https://www.e2open.com/demand-sensing-forecasting-and-inventory-benchmark-study-2018/`.

Elmachtoub AN, Grigas P (2021) Smart "predict, then optimize". *Management Science* URL `http://dx.doi.org/10.1287/mnsc.2020.3922`.

Eppen GD (1979) Note–effects of centralization on expected costs in a multi–location newsboy problem. *Management Science* 25(5):498–501.

Ferdows K, Vereecke A, De Meyer A (2016) Delayering the global production network into congruent subnetworks. *Journal of Operations Management* 41:63–74.

Fisher M, Raman A (1996) Reducing the cost of demand uncertainty through accurate response to early sales. *Operations Research* 44(1):87–99.

Fisher ML (1997) What is the right supply chain for your product? *Harvard Business Review* 75:105–117.

Gallego G, Topaloglu H (2019) *Revenue Management and Pricing Analytics* (Springer, New York, US).

Ghemawat P, Nueno JL (2006) Zara: Fast Fashion, Harvard Business School Case: 9-703-497.

Hausman WH (1969) Sequential decision problems: A model to exploit existing forecasters. *Management Science* 16(2):B93–B111.

Heath DC, Jackson PL (1994) Modeling the evolution of demand forecasts with application to safety stock analysis in production/distribution systems. *IIE Transactions* 26(3):17–30.

Heskett JL, Signorelli S (1989) Benetton (A), Harvard Business School Case: 9-685-014.

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

Johnson ME, Anderson E (2000) Postponement strategies for channel derivatives. *The International Journal of Logistics Management* 11(1):19–36.

Kouvelis P, Tian Z (2014) Flexible capacity investments and product mix: Optimal decisions and value of postponement options. *Production and Operations Management* 23(5):861–876.

Lee HL, Tang CS (1997) Modelling the costs and benefits of delayed product differentiation. *Management Science* 43(1):40–53.

Lee HL, Tang CS (1998) Variability reduction through operations reversal. *Management Science* 44(2):162–172.

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

Oh S, Özer Ö (2013) Mechanism design for capacity planning under dynamic evolutions of asymmetric demand forecasts. *Management Science* 59(4):987–1007.

Paranikas P, Whiteford GP, Tevelson B, Belz D (2015) How to negotiate with powerful suppliers. *Harvard Business Review* 93(7):90–96.

Parsons JCW, Graves SC (2005) Reebok NFL replica jerseys: A case for postponement, MIT Sloan School of Management, SCM Case.

Paul A, Tan YR, Vakharia AJ (2015) Inventory planning for a modular product family. *Production and Operations Management* 24(7):1033–1053.

Petersen KJ, Handfield RB, Ragatz GL (2005) Supplier integration into new product development: coordinating product, process and supply chain design. *Journal of operations management* 23(3-4):371–388.

Rajagopalan S (2013) Impact of variety and distribution system characteristics on inventory levels at US retailers. *Manufacturing & Service Operations Management* 15(2):191–204.

Shapiro A, Dentcheva D, Ruszczyński A (2009) *Lectures on Stochastic Programming: Modeling and Theory* (SIAM:Society for Industrial and Applied Mathematics, Philadelphia, PA 19104-2688, USA).

Song JS, Zipkin PH (2012) Newsvendor problems with sequentially revealed demand information. *Naval Research Logistics* 59(8):601–612.

Swaminathan JM, Tayur SR (1998) Managing broader product lines through delayed differentiation using Vanilla Boxes. *Management Science* 44(12-part-2):S161–S172.

Van Mieghem JA (1998) Investment strategies for flexible resources. *Management Science* 44(8):1071–1078.

van Rooy JC (2010) *Logistical information sharing and collaboration between* ASML *and its supply chain*. Master's thesis, Eindhoven University of Technology, URL `https://pure.tue.nl/ws/portalfiles/portal/46993635/691329-1.pdf`.

Viswanathan S, Allampalli DG (2012) Four star industries Singapore–matching supply with demand, Nanyang Business School Case: ABCC-2009-005.

Wang T, Atasu A, Kurtulus M (2012) Multiordering newsvendor with dynamic forecast evolution. *Manufacturing and Service Operations Management* 14(3):472–484.

Zinn W (2019) A historical review of postponement research. *Journal of Business Logistics* 40(1):66–72.

# Online Appendix

## Proof of Theorem 1

At $t = t_{n-1}$, the expected profit can be formalized as a newsvendor problem:

$$G_{n-1}(Q_{n-1}, D_{n-1}) = \mathbb{E}_{D_n|D_{n-1}}\Big(p\min(D_n, Q_{n-1})\Big) - c_{n-1}Q_{n-1}, \tag{26}$$

$$= (p - c_{n-1})Q_{n-1} - p\int_0^{Q_{n-1}} (Q_{n-1} - D_n)f(D_n|D_{n-1})\partial D_n, \tag{27}$$

where $f(\cdot|\cdot)$ and $F(\cdot|\cdot)$ denote conditional demand density and distribution functions, respectively. The optimal order quantity is obtained by:

$$\frac{\partial G_{n-1}}{\partial Q_{n-1}} = p(1 - F(Q_{n-1}|D_{n-1})) - c_{n-1} = 0. \tag{28}$$

With $p > c_{n-1}$, $G_{n-1}(\cdot, D_{n-1})$ is a concave function for any given $D_{n-1}$. Then,

$$V_{n-1}(Q_{n-2}, D_{n-1}) = \max_{Q_{n-1}\le Q_{n-2}}\Big\{G_{n-1}(Q_{n-1}, D_{n-1})\Big\}. \tag{29}$$

For $Q^*_{n-1} = \{Q_{n-1}|\partial G_{n-1}/\partial Q_{n-1} = 0\}$,

$$V_{n-1}(Q_{n-2}, D_{n-1}) = \begin{cases} G_{n-1}(Q_{n-2}, D_{n-1}) & \text{if } Q^*_{n-1} > Q_{n-2}, \\ G_{n-1}(Q^*_{n-1}, D_{n-1}) & \text{if } Q^*_{n-1} \le Q_{n-2}. \end{cases} \tag{30}$$

$V_{n-1}(\cdot, D_{n-1})$ is a non-decreasing concave function due to the concavity of $G_{n-1}(\cdot, D_{n-1})$. Then,

$$G_{n-2}(Q_{n-2}, D_{n-2}) = \mathbb{E}_{D_{n-1}|D_{n-2}}\Big(V_{n-1}(Q_{n-2}, D_{n-1})\Big) - c_{n-2}Q_{n-2}, \tag{31}$$

which is also concave because $V_{n-1}(\cdot, D_{n-1})$ is concave. Then, $G_i(\cdot, D_i)$ is a concave function (by induction) for $i \in \{0, 1, \cdots, n-2\}$, and the optimal policy is:

$$Q^*_i = \arg\max_{Q_i}\{G_i(Q_i, D_i)\}, \qquad \forall i \in \{0, 1, \cdots, n\}. \tag{32}$$

Suppose in period $i+1$ for $i \in \{0, 1, \cdots, n-2\}$,

$$V_{i+1}(Q_i, D_{i+1}) = \begin{cases} G_{i+1}(Q_i, D_{i+1}) & \text{if } Q^*_{i+1} > Q_i, \\ G^*_{i+1}(D_{i+1}) & \text{if } Q^*_{i+1} \le Q_i, \end{cases} \tag{33}$$

where $G^*_{i+1}(D_{i+1}) = G_{i+1}(Q^*_{i+1}, D_{i+1})$. Then,

$$G_i(Q_i, D_i) = \mathbb{E}_{D_{i+1}|D_i}[V_{i+1}(Q_i, D_{i+1})] - c_iQ_i, \tag{34}$$

$$= \int_{\overline{D}_{i+1}}^{+\infty} G_{i+1}(Q_i, D_{i+1})f(D_{i+1}|D_i)\partial D_{i+1} + \int_0^{\overline{D}_{i+1}} G^*_{i+1}(D_{i+1})f(D_{i+1}|D_i)\partial D_{i+1} - c_iQ_i, \tag{35}$$

where $\overline{D}_{i+1}$ is the value of demand forecast at time $i+1$ that makes the optimal order quantity equal to that of the previous period (i.e., $Q^*_{i+1} = Q_i$). Taking the first derivative, we obtain the following result:

$$\frac{\partial G_i}{\partial Q_i} = g_i(Q_i, D_i) = \int_{\overline{D}_{i+1}}^{+\infty} g_{i+1}(Q_i, D_{i+1})f(D_{i+1}|D_i)\partial D_{i+1} - c_i = 0. \tag{36}$$

Using Equation (36), the optimal value of $Q_i$ for $i \in \{1, \cdots, n-2\}$ can be found by backward induction.

The optimal value of $Q_{n-1}$ is given by Equation (28). Combining Equations (28) and (36), the optimal value of $Q_{n-2}$ can be calculated by:

$$g_{n-2}(Q_{n-2}, D_{n-2}) = \int_{\overline{D}_{n-1}}^{+\infty} (pPr(D_n > Q_{n-2}|D_{n-1}) - c_{n-1}) f_{n-1}(D_{n-1}|D_{n-2}) \partial D_{n-1} - c_{n-2},$$
$$= pPr(D_n > Q_{n-2}, D_{n-1} > \overline{D}_{n-1}) - c_{n-1} Pr(D_{n-1} > \overline{D}_{n-1}) - c_{n-2}. \quad (37)$$

By induction, we obtain for $i \in \{0, 1, \cdots, n-1\}$ the following result:

$$g_i(Q_i, D_i) = pPr(D_n > Q_i, \mathbf{D}_{\{i+1,n-1\}} > \overline{\mathbf{D}}_{\{i+1,n-1\}}) - c_{n-1} Pr(\mathbf{D}_{\{i+1,n-1\}} > \overline{\mathbf{D}}_{\{i+1,n-1\}})$$
$$- c_{n-2} Pr(\mathbf{D}_{\{i+1,n-2\}} > \overline{\mathbf{D}}_{\{i+1,n-2\}}) - \cdots - c_{i+1} Pr(D_{i+1} > \overline{D}_{i+1}) - c_i, \quad (38)$$

where $\mathbf{D}_{\{i+1,n-1\}}$ is a vector denoting demand forecasts from period $i+1$ to $n-1$. Then, the optimal order quantity in each period can be found by $q_i = \min(Q_{i-1}, Q_i^*)$ such that $Q_i^* = \{Q_i | g_i(Q_i, D_i) = 0\}$.

**Proof of Proposition 1**

The proof is straightforward from the final result of the proof of Theorem 1: $q_i = \min(Q_{i-1}, Q_i^*)$. If the demand forecast in period $i$ turns out to be high, the order quantity is constrained by $q_{i-1}$. Otherwise, $q_i < q_{i-1}$. Therefore, $q_0 > \mathbb{E}[q_1|D_0] > \cdots > \mathbb{E}[q_{n-1}|D_0]$.

**Proof of Proposition 2**

***Part A:*** For $j \in \{i, \cdots, n-1\}$, suppose $c_j$ is increased by $\Delta c_j$, making the cost of processing the $j^{th}$ operation equal to $c_j + \Delta c_j$. Suppose $Q_i' = Q_i$ and $g_i(Q_i, D_i) = 0$. Then,

$$g_i(Q_i', D_i) - g_i(Q_i, D_i) = g_i(Q_i', D_i) = -\Delta c_j Pr(\mathbf{D}_{\{i+1,j\}} > \overline{\mathbf{D}}_{\{i+1,j\}}). \quad (39)$$

If $Q_i' = Q_i$, the term $g_i(Q_i', D_i)$ fails to be equal to zero after the cost increase, meaning that setting $Q_i' = Q_i$ does not optimize the ordering decision anymore. It follows from Equation (39) that the ordering decision after the cost increase is optimized for $Q_i' < Q_i$ such that $g_i(Q_i', D_i) = 0$ for $Q_i' < Q_i$. Thus, $q_i' \prec q_i$ for $i \in \{0, \cdots, j\}$, meaning that an increase in the cost of an operation leads to a reduction of order quantities at upstream echelons. Because the downstream order quantities are constrained by the upstream ones, such that $Q_0 \geq Q_1 \geq \cdots \geq Q_{n-1}$, such a reduction of upstream order quantities also has a cascading impact of reducing the downstream order quantities. Thus, $q_i' \prec q_i$ for $i \in \{0, \cdots, n-1\}$.

***Part B:*** Given $c_0 = \cdots = c_{i-1} = c_{i+1} = \cdots = c_{n-1} = c_{fixed} < c_i$, the expected profit at $t = t_0$ is written as follows:

$$G_0(Q_0^*, D_0 | i = j) = \int_0^{Q_0^*} g_0(Q_0, D_0) dQ_0, \quad (40)$$

where $Q_0^*$ is the optimal order quantity at $t_0$ when $i = j$ for $j \in \{0, \cdots, n-2\}$. Thus, $G_0(Q_0^*, D_0 | i = j+1)$ gives a lower bound for the expected profit for $i = j+1$. Combining this expression with Equation (11) yields the following result:

$$G_0(Q_0^*, D_0|i=j+1) - G_0(Q_0^*, D_0|i=j) = -\int_0^{Q_0^*} (c_i - c_{fixed}) Pr(\mathbf{D}_{\{1,i+1\}} > \overline{\mathbf{D}}_{\{1,i+1\}}) dQ_0 + \int_0^{Q_0^*} (c_i - c_{fixed}) Pr(\mathbf{D}_{\{1,i\}} > \overline{\mathbf{D}}_{\{1,i\}}) dQ_0. \quad (41)$$

Equation (41) always has a non-negative value because $Pr(\mathbf{D}_{\{1,i\}} > \overline{\mathbf{D}}_{\{1,i\}}) \geq Pr(\mathbf{D}_{\{1,i+1\}} > \overline{\mathbf{D}}_{\{1,i+1\}})$. Thus, swapping a high-cost operation with the adjacent downstream one leads to higher profits. It is straightforward by induction that swapping operation $i$ with any operation from the set $\{i+1, \cdots, n-1\}$ increases the profit.

**Proof of Proposition 3**

***Part A:*** If the lead time of operation $i$ is reduced by $\Delta t$, the starting times for the first $i+1$ operations are updated as follows:

$$t_0 + \Delta t = t_1 + \Delta t = \cdots = t_j + \Delta t = \cdots = t_i + \Delta t. \quad (42)$$

Then, ordering decisions for the first $i+1$ operations are made after a delay of $\Delta t$. Delaying the ordering decisions leads to improved demand accuracy for the first $i+1$ decisions as given by Equation (4), which therefore increases the expected profit.

If the lead time of operation $j$ is reduced by $\Delta t$, the starting times for the operations until $j$ are likewise delayed for $\Delta t$. Reducing the lead-time of operation $i$, compared to that of $j$ for $i > j$, makes it possible to also delay the decision epochs for $j+1$, $j+2$, $\cdots$, $i$. This results in a higher profit than what can be achieved by reducing the lead time of $j$, which completes the proof of the proposition.

The proposition can also be extended to the multi-product case. Suppose product proliferation occurs once along the supply chain at time $t_i$. Following the same logic, we can conclude that reducing the lead time of operation $j$ for $j \in \{i, \cdots, n-1\}$ by an amount of $\Delta t_j \leq t_{j+1} - t_j$ increases the expected profit more than what can be achieved by reducing the lead time of operation $j$ for $j \in \{0, \cdots, i-1\}$. Therefore, reducing the lead times of operations post-proliferation yields higher profits than reducing the lead times of those before the proliferation.

***Part B:*** The decision epochs are $t_0$, $t_1$, ..., $t_{n-1}$ in the initial case. After swapping the operation $i$ with $j \in \{i+1, \cdots, n-1\}$, the decision epochs for the operations $\{i+1, \cdots, j\}$ are updated as $t_{i+1} + \Delta t$, $t_{i+2} + \Delta t$, ..., $t_j + \Delta t$, where $\Delta t = \Delta t_{fixed} - (t_i - t_{i-1})$. Therefore, the operations $\{i+1, \cdots, j\}$ are postponed by $\Delta t$. Because all operations have the same cost value, swapping the operations $i$ and $j$ leads to a profit increase.

**Proof of Theorem 2**

Let $D_{n,r}^j \geq 0$ denote a realization of $D_n^j$ such that $r \in \mathcal{S}$, where $\mathcal{S} = \{1, 2, \cdots\}$ is defined as a large finite set of positive integers. The values of $D_{n,r}^j$ $\forall r \in \mathcal{S}$ constitute the set of all possible demand realizations. We use $\mathcal{W}_j^r$ to denote the sales value for a demand realization of $D_{n,r}^j$ such that $\mathcal{W}_j^r = \mathcal{W}_j(Q_{n-1}^j, D_{n,r}^j)$. Then, the SP model (15)–(17) can be written as a large-scale LP model:

$$\underset{Q_{n-1}^j, \forall j \in \Theta_{n-1}}{\text{Maximize}} \quad z = \sum_{j \in \Theta_{n-1}} \left( p_j \sum_{r \in \mathcal{S}} Pr(\mathcal{W}_j^r) \mathcal{W}_j^r - c_{n-1}^j Q_{n-1}^j \right) \quad (43)$$

subject to:

$$\mathcal{W}_j^r - Q_{n-1}^j \le 0, \qquad \forall j \in \Theta_{n-1}, \quad \forall r \in \mathcal{S}, \tag{44}$$

$$\mathcal{W}_j^r \le D_{n,r}^j, \qquad \forall j \in \Theta_{n-1}, \quad \forall r \in \mathcal{S}, \tag{45}$$

$$\sum_{j \in \Theta_{n-1}^k} Q_{n-1}^j \le Q_{n-2}^k, \qquad \forall k \in \Theta_{n-2}, \tag{46}$$

$$Q_{n-1}^j \ge 0, \quad \mathcal{W}_j^r \ge 0, \qquad \forall j \in \Theta_{n-1}, \quad \forall r \in \mathcal{S}. \tag{47}$$

We remark that we add Constraints (44) and (45) to satisfy the condition $W_j = \min\{Q_{n-1}^j, D_n^j\}$ for the optimal solution. Then, the dual problem is:

$$\underset{\lambda_{j,r},\beta_{j,r},\gamma}{\text{Minimize}} \quad w = \sum_{j \in \Theta_{n-1}} \sum_{r \in \mathcal{S}} \beta_{j,r} D_{n,r}^j + \sum_{k \in \Theta_{n-2}} \gamma_k Q_{n-2}^k \tag{48}$$

subject to:

$$\lambda_{j,r} + \beta_{j,r} \ge Pr(\mathcal{W}_j^r) p_j, \qquad \forall j \in \Theta_{n-1}, \quad r \in \mathcal{S}, \tag{49}$$

$$\sum_{r \in \mathcal{S}} \lambda_{j,r} - \gamma_k \le c_{n-1}^j, \qquad \forall j \in \Theta_{n-1}^k, \quad k \in \Theta_{n-2}, \quad r \in \mathcal{S}, \tag{50}$$

$$\lambda_{j,r} \ge 0, \quad \beta_{j,r} \ge 0, \quad \gamma_k \ge 0, \qquad \forall j \in \Theta_{n-1}^k, \quad k \in \Theta_{n-2},, \quad r \in \mathcal{S}. \tag{51}$$

The values of $\lambda_{j,r}$ and $\beta_{j,r}$ for each $j \in \Theta_{n-1}$ are found by the parametric analysis:

1. $\lambda_{j,r} = 0$ and $\beta_{j,r} = Pr(\mathcal{W}_j^r) p_j$ for $j \in \Theta_{n-1}$ when $D_{n,r}^j \le Q_{n-1}^j$.
2. Likewise, $\lambda_{j,r} = Pr(\mathcal{W}_j^r) p_j$ and $\beta_{j,r} = 0$ for $j \in \Theta_{n-1}$ when $D_{n,r}^j > Q_{n-1}^j$.

Then, the constraint (50) is written as follows:

$$p_j Pr(D_n^j > Q_{n-1}^j) - \gamma_k \le c_{n-1}^j, \qquad \forall j \in \Theta_{n-1}. \tag{52}$$

We set a value for the dual variable $\gamma_k$ for $k \in \Theta_{n-2}$ such that:

$$\gamma_k = (p_j - c_{n-1}^j) - p_j Pr(D_n^j \le Q_{n-1}^j) = g_{n-1}^j(Q_{n-1}^j, D_{n-1}^j), \qquad \forall j \in \Theta_{n-1}^k. \tag{53}$$

Then, the objective function value for the dual problem becomes:

$$w = \sum_{j \in \Theta_{n-1}} \Big[ (p_j - c_{n-1}^j) Q_{n-1}^j - p_j \int_0^{Q_{n-1}^j} (Q_{n-1}^j - D_n^j) f^j(D_n^j) \partial D_n^j \Big]. \tag{54}$$

Equation (54) is equivalent to the solution of the primal problem for the feasible $Q_{n-1}^j$ values. It follows from the strong theorem of duality that the optimal solution satisfies Equation (53). Therefore, we have the following conditions of optimality:

$$\gamma_k = g_{n-1}^j(Q_{n-1}^j, D_{n-1}^j) \ge 0, \qquad \forall j \in \Theta_{n-1}^k, \quad k \in \Theta_{n-2} \tag{55}$$

$$\sum_{j \in \Theta_{n-1}^k} Q_{n-1}^j \le Q_{n-2}^k, \qquad \forall k \in \Theta_{n-2}. \tag{56}$$

If the constraint (46) is not binding for a given $k \in \Theta_{n-2}$, the dual variable $\gamma_k$ becomes zero. In this case, the optimal solution reduces to the solution of $|\Theta_{n-1}^k|$ independent newsvendor problems in the last period—that is, the order quantity for each product in the set $\Theta_{n-1}^k$ can be found by solving an unconstrained newsvendor problem. Otherwise, the optimal solution exists at the point where the marginal value of producing one unit is the same for all products in the set $\Theta_{n-1}^k$.

In period $i \in \{1, 2, \cdots, n-2\}$, the optimization problem is written as follows:

$$\underset{Q_i^j, \forall j \in \Theta_i}{\text{Maximize}} \quad z = \sum_{j \in \Theta_i} G_i^{\Upsilon_i^j}(Q_i^j, \mathbf{D}_i^{\Upsilon_i^j}) \tag{57}$$

subject to:

$$\sum_{j \in \Theta_i^k} Q_i^j \leq Q_{i-1}^k, \quad 0 \leq Q_i^j, \qquad \forall k \in \Theta_{i-1}, \quad \forall j \in \Theta_i^k. \tag{58}$$

The term $G_i^{\Upsilon_i^j}(Q_i^j, \mathbf{D}_i^{\Upsilon_i^j})$ is the total expected profit generated from all the products in the set $\Upsilon_i^j$, and $\mathbf{D}_i^{\Upsilon_i^j}$ is the vector of demand forecasts of the products in $\Upsilon_i^j$ at time $t_i$. Note that $\Upsilon_i^j$ is the set of end products sold in the market whose availability depends on the order quantity decision of $Q_i^j$. We will discuss the derivation of $G_i^{\Upsilon_i^j}(Q_i^j, \mathbf{D}_i^{\Upsilon_i^j})$ in detail below.

The dual problem (57)–(58) is formulated as:

$$\underset{\lambda_j, \beta_j, \gamma_k}{\text{Minimize}} \quad w = \sum_{k \in \Theta_{i-1}} \gamma_k Q_{i-1}^k \tag{59}$$

subject to:

$$g_i^{\Upsilon_i^j}(Q_i^j, \mathbf{D}_i^{\Upsilon_i^j}) \leq \gamma_k, \qquad \forall j \in \Theta_i^k, \quad \forall k \in \Theta_{i-1}, \tag{60}$$

with $\partial G_i^{\Upsilon_i^j}(Q_i^j, \mathbf{D}_i^{\Upsilon_i^j}) / \partial Q_i^j = g_i^{\Upsilon_i^j}(Q_i^j, \mathbf{D}_i^{\Upsilon_i^j})$. Then, the optimal solution in each period $i \in \{1, \cdots, n-2\}$ satisfies the following equations:

$$\gamma_k = g_i^{\Upsilon_i^j}(Q_i^j, \mathbf{D}_i^{\Upsilon_i^j}) \geq 0, \qquad \forall j \in \Theta_i^k, \quad k \in \Theta_{i-1} \tag{61}$$

$$\sum_{j \in \Theta_i^k} Q_i^j \leq Q_{i-1}^k, \qquad \forall k \in \Theta_{i-1}. \tag{62}$$

Following the steps similar to the proof of Theorem 1, we obtain the following expression for $t = t_{n-2}$:

$$\begin{aligned} G_{n-2}^{\Upsilon_{n-2}^j}(Q_{n-2}^j, \mathbf{D}_{n-2}^{\Upsilon_{n-2}^j}) = & \int_{\overline{D}_{n-1}^{\Upsilon_{n-2}^j}}^{+\infty} G_{n-1}^{\Upsilon_{n-2}^j}(Q_{n-2}^j, D_{n-1}^{\Upsilon_{n-2}^j}) f(D_{n-1}^{\Upsilon_{n-2}^j} | D_{n-2}^{\Upsilon_{n-2}^j}) \partial D_{n-1}^{\Upsilon_{n-2}^j} \\ & + \int_0^{\overline{D}_{n-1}^{\Upsilon_{n-2}^j}} G_{n-1}^{*\Upsilon_{n-2}^j}(Q_{n-2}^{*j}) f(D_{n-1}^{\Upsilon_{n-2}^j} | D_{n-2}^{\Upsilon_{n-2}^j}) \partial D_{n-1}^{\Upsilon_{n-2}^j} - c_{n-2}^j Q_{n-2}^j, \end{aligned} \tag{63}$$

where $D_{n-1}^{\Upsilon_{n-2}^j}$ is a random variable denoting the sum of demand forecasts of the items in the set $\Upsilon_{n-2}^j$ at $t = t_{n-1}$ (i.e., $\sum_{k \in \Upsilon_{n-2}^j} D_{n-1}^k$). $\overline{D}_{n-1}^{\Upsilon_{n-2}^j}$ is the value of demand forecast at $t = t_{n-1}$ that makes the optimal order quantity at $t_{n-1}$ equal to that of the previous period (i.e., $Q_{n-2}^j$). Then,

$$g_{n-2}^{\Upsilon_{n-2}^j}(Q_{n-2}^j, \mathbf{D}_{n-2}^{\Upsilon_{n-2}^j}) = \int_{\overline{D}_{n-1}^{\Upsilon_{n-2}^j}}^{+\infty} g_{n-1}^{\Upsilon_{n-2}^j}(Q_{n-1}^j, D_{n-1}^{\Upsilon_{n-2}^j}) f(D_{n-1}^{\Upsilon_{n-2}^j} | D_{n-2}^{\Upsilon_{n-2}^j}) \partial D_{n-1}^{\Upsilon_{n-2}^j} - c_{n-2}^j, \tag{64}$$

where $g_{n-1}^{\Upsilon_{n-2}^j}(Q_{n-1}^j, D_{n-1}^{\Upsilon_{n-2}^j}) = g_{n-1}^k(Q_{n-1}^k, D_{n-1}^k)$ for any $k \in \Upsilon_{n-2}^j$ as given by Equation (55). Using the last expression, we obtain the following result by induction:

$$g_i^{\Upsilon_i^j}(Q_i^j, \mathbf{D}_i^{\Upsilon_i^j}) = \int_{\underline{D}_{i+1}^{\Upsilon_i^j}}^{+\infty} g_{i+1}^{\Upsilon_i^j}(Q_{i+1}^j, D_{i+1}^{\Upsilon_i^j}) f(D_{i+1}^{\Upsilon_i^j} | D_i^{\Upsilon_i^j}) \partial D_{i+1}^{\Upsilon_i^j} - c_i^j = \gamma_k, \qquad \forall j \in \Theta_i^k. \tag{65}$$

If the constraint (58) for a given $k$ is not binding, the dual variable $\gamma_k$ becomes zero. In this case, the optimal solution reduces to the solution of $|\Theta_i^k|$ unconstrained problems using Equation (11). Let $Q_i^{j^*}$ denote the order quantity for $j \in \Theta_i^k$ and $k \in \Theta_{i-1}$ satisfying Equation (11) and $\widehat{Q}_i^j$ denote the order quantity satisfying Equation (65). Then, the optimal order quantity is found as follows:

$$q_i^j = \begin{cases} Q_i^{j^*} & \text{if } \sum_{j \in \Theta_i^k} Q_i^{j^*} < Q_{i-1}^k, \\ \widehat{Q}_i^j & \text{if } \sum_{j \in \Theta_i^k} Q_i^{j^*} \geq Q_{i-1}^k. \end{cases} \tag{66}$$

**Proof of Proposition 4**

Suppose product proliferation occurs once at time $t_i$ along the supply chain. The mathematical model given by (57)–(58) is then rewritten as follows with an objective function of maximizing the expected profit at time $t_i$:

$$\underset{Q_i^j, \forall j \in \Theta_i}{\text{Maximize}} \quad z = \sum_{j \in \Theta_i} G_i^j(Q_i^j, \mathbf{D}_i^{\Theta_i}) \tag{67}$$

subject to:

$$\sum_{j \in \Theta_i^k} Q_i^j \leq Q_{i-1}^k, \qquad \forall k \in \Theta_{i-1}. \tag{68}$$

Then, the value of postponing the point of the proliferation is calculated by $\partial z / \partial t_i$.

As the next step, we fix $c_i^j = c_{i+1}^j = \cdots = c_{n-1}^j = 0 \ \forall j \in \Theta_i$ and analyze the impact of incrementally increasing $c_i^j$ on the expected profit. Since $c_i^j = c_{i+1}^j = \cdots = c_{n-1}^j = 0$, manufacturer orders the maximum amount in all the remaining periods (i.e., $\{i, i+1, \cdots, n-1\}$) such that:

$$Q_{i-1}^j = Q_i^j = \cdots = Q_{n-1}^j. \tag{69}$$

And,

$$G_i^j(Q_i^{j^*}, D_i^j) = G_i^j(Q_{i-1}^j, D_i^j) \qquad \forall j \in \Theta_i, \tag{70}$$

Given that the order quantity at $t_i$ is set equal to $Q_{i-1}$, the expected profit is not affected by the ordering decision and we have the following relationship:

$$\partial G_i^j(Q_i^j, D_i^j) / \partial t_i = 0 \quad \rightarrow \quad \partial z / \partial t_i = 0. \tag{71}$$

Having set $c_i^j = c_{i+1}^j = \cdots = c_{n-1}^j = 0 \ \forall j \in \Theta_i$ implies that the constraint (68) is binding. Now, we slightly increase $c_i^j$ such that $c_i^j > 0$ and $c_{i+1}^j = \cdots = c_{n-1}^j = 0 \ \forall j \in \Theta_i$. Then, we obtain:

$$G_i^j(Q_i^j, D_i^j) = (p_j - c_i^j) Q_i^j - p_j \int_0^{Q_i^j} (Q_i^j - D_n^j) f(D_n^j | D_i^j) \partial D_n^j, \tag{72}$$

$$= (p_j - c_i^j) Q^j - p_j Q^j \Phi\Big( \frac{\ln(Q_i^j / D_i^j)}{\sigma \sqrt{t_n - t_i}} + \sigma \sqrt{t_n - t_i} / 2 \Big) + p_j D_i^j \Phi\Big( \frac{\ln(Q_i^j / D_i^j)}{\sigma \sqrt{t_n - t_i}} - \sigma \sqrt{t_n - t_i} / 2 \Big). \tag{73}$$

From Equation (73), we obtain:

$$\partial G_i^j(Q_i^j, D_i^j)/\partial t_i > 0 \quad \rightarrow \quad \partial z/\partial t_i > 0. \tag{74}$$

Therefore, $\partial z/\partial t_i$ increases as $c_i$ increases, which completes the proof of Proposition 4.

**Proof of Theorem 3**

Suppose the point of proliferation is deferred to $t_{i+1}$ by swapping it with the adjacent high-cost operation. We consider a generalized model such that $|\Theta_{i+1}| \geq 2$. Then, $c_{i+1}$ is the cost of operation per unit that causes the proliferation. And, $c_i$ is the cost of the high-cost operation. In the first part, we show that postponing the proliferation causes a loss of profit if $c_i$ exceeds a certain threshold. To this end, we use marginal analysis. We refer the reader to Gallego and Topaloglu (2019) for the applications of the marginal analysis in the context of revenue management (especially the derivations of Lemma 1.4, Proposition 1.5, and Theorem 1.6 therein are useful illustrations of reconstructing the Dynamic Programming formulation with marginal analysis).

The Bellman equation at time $t_i$ is then given by Equation (6):

$$V_i(Q_{i-1}, \mathbf{D}_i^{\Upsilon_i}) = \max_{Q_i \leq Q_{i-1}} \Big\{ -c_iQ_i + \mathbb{E}_{D_{i+1}|D_i}\Big[V_{i+1}(Q_i, \mathbf{D}_{i+1}^{\Upsilon_i})\Big]\Big\}, \tag{75}$$

where the expression inside the maximization is equal to $G_i^{\Upsilon_i}(Q_{i-1}, \mathbf{D}_i^{\Upsilon_i})$, which is a concave function of $Q_{i-1}$ as stated above. Theorem 1 shows that the solution to problem (75) satisfies:

$$Q_i = \begin{cases} Q_{i-1} & \text{if } D_i^{\Upsilon_i} \geq \overline{D}_i^{\Upsilon_i}, \\ Q_i^* & \text{otherwise.} \end{cases} \tag{76}$$

As stated in the proof of Theorem 2, the condition $D_i^{\Upsilon_i} \geq \overline{D}_i^{\Upsilon_i}$ implies that the dual variable is larger than zero: $\gamma_i = g_i(Q_i, \mathbf{D}_i^{\Upsilon_i}) > 0$. Thus,

$$\begin{aligned} \mathbb{E}[V_i(Q_{i-1}, \mathbf{D}_i^{\Upsilon_i})] = &\Big(-c_iQ_{i-1} + \mathbb{E}[V_{i+1}(Q_{i-1}, \mathbf{D}_i^{\Upsilon_i})]\Big)Pr(\gamma_i > 0) \\ &+\Big(-c_iQ_i^* + \mathbb{E}[V_{i+1}(Q_i^*, \mathbf{D}_i^{\Upsilon_i})]\Big)Pr(\gamma_i = 0). \end{aligned} \tag{77}$$

Following the same steps, we develop an expression for $\mathbb{E}[V_i(Q_{i-1} - 1, \mathbf{D}_i^{\Upsilon_i})]$:

$$\begin{aligned} \mathbb{E}[V_i(Q_{i-1} - 1, \mathbf{D}_i^{\Upsilon_i})] = &\Big(-c_i(Q_{i-1} - 1) + \mathbb{E}[V_{i+1}(Q_{i-1} - 1, \mathbf{D}_i^{\Upsilon_i})]\Big)Pr(\gamma_i > 0) \\ &+\Big(-c_iQ_i^* + \mathbb{E}[V_{i+1}(Q_i^*, \mathbf{D}_i^{\Upsilon_i})]\Big)Pr(\gamma_i = 0). \end{aligned} \tag{78}$$

In what follows, we apply the marginal analysis to complete the proof of the theorem. We first calculate the expected marginal profit:

$$\begin{aligned} \mathbb{E}[\Delta V_i(Q_{i-1}, \mathbf{D}_i^{\Upsilon_i})] &= \mathbb{E}[V_i(Q_{i-1}, \mathbf{D}_i^{\Upsilon_i})] - \mathbb{E}[V_i(Q_{i-1} - 1, \mathbf{D}_i^{\Upsilon_i})], \\ &= \Big(-c_i + \mathbb{E}[\Delta V_{i+1}(Q_{i-1}, \mathbf{D}_i^{\Upsilon_i})]\Big)Pr(\gamma_i > 0). \end{aligned} \tag{79}$$

The "$\Delta V_i(Q_{i-1}, \mathbf{D}_i^{\Upsilon_i})$" term is equal to $g_i(Q_i, \mathbf{D}_i^{\Upsilon_i})$ at the optimality, which can take a non-zero value if the order quantity constraint is binding—that is, when $\gamma_i > 0$. Otherwise, $g_i(Q_i = Q_i^*, \mathbf{D}_i^{\Upsilon_i}) = 0$, which is the condition of optimality. Thus, the expectation term at the right-hand side of Equation (79) is equal to:

$$\mathbb{E}[\Delta V_{i+1}(Q_{i-1}, \mathbf{D}_i^{\Upsilon_i})] = \mathbb{E}[g_{i+1}(Q_{i-1}, \mathbf{D}_i^{\Upsilon_i})]Pr(\gamma_{i+1} > 0), \tag{80}$$

where $g_{i+1}(Q_{i-1},\mathbf{D}_i^{\Upsilon_i})=\gamma_{i+1}=g_{i+1}^j(Q_{i+1}^j,D_{i+1}^j)\ \forall j\in\Theta_{i+1}$ subject to $\sum_{j\in\Theta_{i+1}}Q_{i+1}^j=Q_{i-1}$. Then, Equation (79) is written as follows:

$$\mathbb{E}[\Delta V_i(Q_{i-1},\mathbf{D}_i^{\Upsilon_i})] = \Big(-c_i+\mathbb{E}[g_{i+1}(Q_{i-1},\mathbf{D}_i^{\Upsilon_i})]Pr(\gamma_{i+1}>0)\Big)Pr(\gamma_i>0) \tag{81}$$

Suppose now that the proliferation occurs at time $t_i$, and its cost is still denoted by $c_{i+1}$. We use the "$\widehat{\ }$" sign to denote the value function for this case. Then, the Bellman equation at time $t_i$ is written as follows:

$$\widehat{V}_i(Q_{i-1},\mathbf{D}_i^{\Upsilon_i}) = \max_{Q_i^j,\forall j\in\Theta_i,\sum Q_i^j\le Q_{i-1}}\Big\{-c_{i+1}\sum_{j\in\Theta_i}Q_i^j+\mathbb{E}_{D_{i+1}|D_i}\Big[\widehat{V}_{i+1}(\mathbf{Q}_i^{\Upsilon_i},\mathbf{D}_{i+1}^{\Upsilon_i})\Big]\Big\}. \tag{82}$$

Following the steps between Equations (75)–(81), we obtain the following result:

$$\mathbb{E}[\Delta\widehat{V}_i(Q_{i-1},\mathbf{D}_i^{\Upsilon_i})] = \Big(-c_{i+1}+\mathbb{E}\Big[\sum_{j\in\Theta_i}\omega_j\hat{g}_{i+1}^j(Q_{i+1}^j,D_{i+1}^j)\Big]Pr(\widehat{\gamma}_{i+1}>0)\Big)Pr(\widehat{\gamma}_i>0), \tag{83}$$

where $\omega_j$ is the allocation weight of product $j$, equal to $Q_i^j/Q_{i-1}$, so $\sum_{j\in\Theta_i}\omega_j=1$.

From the concavity of $G_i(Q_i,\mathbf{D}_i^{\Upsilon_i})$, both $\mathbb{E}[\Delta V_i(Q_{i-1},\mathbf{D}_i^{\Upsilon_i})]$ and $\mathbb{E}[\Delta\widehat{V}_i(Q_{i-1},\mathbf{D}_i^{\Upsilon_i})]$ are monotone decreasing functions of $Q_i$. And, $\mathbb{E}[\Delta V_i(0,\mathbf{D}_i^{\Upsilon_i})]=\mathbb{E}[\Delta\widehat{V}_i(0,\mathbf{D}_i^{\Upsilon_i})]=p-\sum_{j=i}^{n-1}c_j$ because the marginal value of producing one unit is equal to the profit at $Q_{i-1}=0$. Therefore, $\widehat{V}_i(Q_{i-1},\mathbf{D}_i^{\Upsilon_i})>V_i(Q_{i-1},\mathbf{D}_i^{\Upsilon_i})$ when $\mathbb{E}[\Delta\widehat{V}_i(Q_{i-1},\mathbf{D}_i^{\Upsilon_i})]>\mathbb{E}[\Delta V_i(Q_{i-1},\mathbf{D}_i^{\Upsilon_i})]$. This relationship is satisfied when

$$\Big(-c_{i+1}+\mathbb{E}\Big[\sum_{j\in\Theta_i}\omega_j\hat{g}_{i+1}^j(Q_{i+1}^j,D_{i+1}^j)\Big]Pr(\widehat{\gamma}_{i+1}>0)\Big)Pr(\widehat{\gamma}_i>0) > \Big(-c_i+\mathbb{E}[g_{i+1}(Q_{i-1},\mathbf{D}_i^{\Upsilon_i})]Pr(\gamma_{i+1}>0)\Big)\times Pr(\gamma_i>0). \tag{84}$$

The term "$\mathbb{E}\Big[\sum_{j\in\Theta_i}\omega_j\hat{g}_{i+1}^j(Q_{i+1}^j,D_{i+1}^j)\Big]Pr(\widehat{\gamma}_{i+1}>0)Pr(\widehat{\gamma}_i>0)$" is the expected marginal profit at $t_{i+1}$ conditional on that $Q_{i-1}$ units are fully utilized in both $i^{th}$ and $(i+1)^{th}$ operations based on an allocation policy determined at $t_i$. Likewise, the term "$\mathbb{E}[g_{i+1}(Q_{i-1},\mathbf{D}_i^{\Upsilon_i})]Pr(\gamma_{i+1}>0)Pr(\gamma_i>0)$" is the expected marginal profit at $t_{i+1}$ conditional on that $Q_{i-1}$ units are fully utilized in both $i^{th}$ and $(i+1)^{th}$ operations based on an *expected* allocation policy determined at $t_i$. Therefore,

$$\mathbb{E}\Big[\sum_{j\in\Theta_i}\omega_j\hat{g}_{i+1}^j(Q_{i+1}^j,D_{i+1}^j)\Big]Pr(\widehat{\gamma}_{i+1}>0)Pr(\widehat{\gamma}_i>0)-c_iPr(\widehat{\gamma}_{i+1}>0)Pr(\widehat{\gamma}_i>0) = \mathbb{E}[g_{i+1}(Q_{i-1},\mathbf{D}_i^{\Upsilon_i})]Pr(\gamma_{i+1}>0)Pr(\gamma_i>0)-c_{i+1}Pr(\gamma_{i+1}>0)\ Pr(\gamma_i>0). \tag{85}$$

Combining the expressions (84) and (85),

$$c_i > c_{i+1}\frac{Pr(\widehat{\gamma}_i>0)-Pr(\gamma_{i+1}>0)Pr(\gamma_i>0)}{Pr(\gamma_i>0)-Pr(\widehat{\gamma}_{i+1}>0)Pr(\widehat{\gamma}_i>0)}, \tag{86}$$

where the right-hand side of this expression gives us the threshold value of $\kappa$. Hence, postponing the proliferation causes a loss of profit if the cost of the high-cost downstream operation (i.e., $c_i$) exceeds the cost of the operation that causes the proliferation (i.e., $c_{i+1}$) multiplied by a pooling factor (i.e., $\mathscr{P}=[Pr(\widehat{\gamma}_i>0)-Pr(\gamma_{i+1}>0)Pr(\gamma_i>0)]/[Pr(\gamma_i>0)-Pr(\widehat{\gamma}_{i+1}>0)Pr(\widehat{\gamma}_i>0)]$).

We now apply asymptotic analysis to prove that the pooling factor $\mathscr{P}$ is greater than one when lead times for the two operations swapped are the same—that is, $t_{i+1}-t_i=t_{i+2}-t_{i+1}$. We recall that the proliferation occurs at time $t_{i+1}$ in the primary case. After switching the operations, it occurs at time $t_i$, which is the

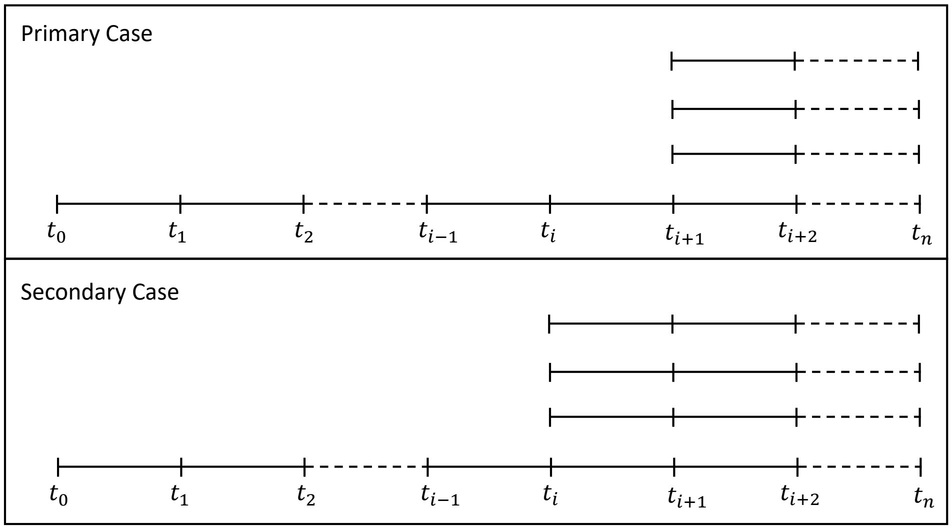


**Figure 17** **Product proliferation structure for both primary and secondary cases**

secondary case. Figure 17 depicts the proliferation structure for both cases. The probability terms in the pooling factor formula can be considered as the probability of observing an inventory shortage. For example, $Pr(\widehat{\gamma}_i > 0)$ is the probability of observing an inventory shortage at time $t_i$ such that the optimal order quantities at $t_i$ in the secondary case is higher than the allocated quantities.

Suppose that the lead times for the two operations swapped are set equal to zero. Thus, $t_i = t_{i+1}$ because both operations start and end immediately at $t_i$. When we increase the lead times slightly such that $t_{i+1} - t_i = t_{i+2} - t_{i+1} = K_\epsilon$ where $K_\epsilon$ is an infinitesimal positive number, we obtain:

$$\lim_{t_i \to^- t_{i+1}} Pr(\gamma_i > 0) < Pr(\widehat{\gamma}_i > 0), \tag{87}$$

because pooling demand slightly reduces the risk of inventory shortage for the primary case. And,

$$\lim_{t_i \to^- t_{i+1}} Pr(\gamma_{i+1} > 0) = Pr(\widehat{\gamma}_{i+1} > 0), \tag{88}$$

because the expected order quantities at $t_{i+1}$ projected at $t_0$ are the same for both primary and secondary cases. Plugging these results in the pooling factor formulation yields:

$$\lim_{t_i \to^- t_{i+1}} \mathscr{P} > 1. \tag{89}$$

The relationships given in Equations (87) and (88) are preserved when both operations have the same lead times. Therefore, $\mathscr{P} > 1$ when the lead times for the two operations swapped are the same.

We now consider another case in which the lead time of the operation that causes the proliferation is substantially longer than the adjacent upstream operation such that $t_{i+2} - t_{i+1} >>> t_{i+1} - t_i$ in the primary case. In this case, switching the two operations does not change the starting time of the operation that causes the proliferation, but it postpones the other operation. Therefore, expediting the proliferation would be viable even when $0 < \mathscr{P} < 1$. In an extreme case such that the lead time of the adjacent upstream operation is close to zero, $\mathscr{P}$ would be close to zero:

$$\lim_{t_{i+1} - t_i \to^+ 0} \mathscr{P} = 0. \tag{90}$$

This result indicates that expediting the proliferation by swapping it with the adjacent upstream operation may help increase the expected profit even when the cost of the upstream operation is very low, which happens when the lead time of the upstream operation is close to zero. Combining this result with Equation (89) yields that $\mathscr{P} < 1$ is only possible when the lead time of the adjacent upstream operation is shorter than the lead time of the operation that causes the proliferation.

Suppose that there are $n$ products in both cases and a new product is added to the portfolio. The order quantity at $t_0$ is fixed to $Q_0$ and the demand for the new product has a perfect negative correlation with one of the existing products. Due to the perfect negative correlation, adding the new product does not change $Pr(\gamma_i > 0)$. But, other probability values increase. Thus, $\mathscr{P}$ increases after a new product whose demand has a perfect negative correlation with one of the existing products is added to the portfolio. As the correlation increases so does the value of $Pr(\gamma_i > 0)$. However, the magnitude of the increase in $Pr(\gamma_i > 0)$ is lower than the magnitude of the increase for the other probability values. Thus, $\mathscr{P}$ also increases after adding the new product for the positive correlation case. However, the magnitude of the change is lower than that of the negative correlation case.

**Proof of Corollary 1**

Suppose that product proliferation occurs at time $t_{i+1}$. We consider a generalized model such that $|\Theta_{i+1}| \geq 2$, and the cost threshold is given by Equation (86):

$$\kappa = c_{i+1} \frac{Pr(\widehat{\gamma}_i > 0) - Pr(\gamma_{i+1} > 0)Pr(\gamma_i > 0)}{Pr(\gamma_i > 0) - Pr(\widehat{\gamma}_{i+1} > 0)Pr(\widehat{\gamma}_i > 0)}. \tag{91}$$

Demand correlation does not affect the values of $Pr(\widehat{\gamma}_i > 0)$, $Pr(\widehat{\gamma}_{i+1} > 0)$, and $Pr(\gamma_{i+1} > 0)$ because these values do not depend on the pooled demand. Before postponing the proliferation, order quantities are placed at the SKU level at $t_i$ and $t_{i+1}$. After postponing the proliferation, the quantities are placed at the SKU level at time $t_{i+1}$. But, order quantities are placed based on pooled demand at time $t_i$ after postponing the proliferation. Therefore, only $Pr(\gamma_i > 0)$ depends on demand correlation.

As indicated in the proof of Theorem 2, the condition $D_i^{\Upsilon_i} \geq \overline{D}_i^{\Upsilon_i}$ is equivalent to the condition that the dual variable is larger than zero: $\gamma_i = g_i(Q_i, \mathbf{D}_i^{\Upsilon_i}) > 0$. The probability of $Pr(\gamma_i > 0)$ increases with demand correlation. Equation (91) shows that the cost threshold value decreases as the value of $Pr(\gamma_i > 0)$ increases. Thus, the $\kappa$ value decreases with demand correlation, which completes the proof of the corollary.

**Proof of Corollary 2**

As the volatility increases so does the values of $Pr(\gamma_i > 0)$, $Pr(\gamma_{i+1} > 0)$, $Pr(\widehat{\gamma}_i > 0)$, $Pr(\widehat{\gamma}_{i+1} > 0)$. The increase in $Pr(\gamma_{i+1} > 0)$ is the same as the increase in $Pr(\widehat{\gamma}_{i+1} > 0)$. This results from the fact that order quantities placed at time $t_{i+1}$ are determined based on the same dynamics before and after switching the operations. However, $Pr(\gamma_i > 0)$ is less affected by demand uncertainty than $Pr(\widehat{\gamma}_i > 0)$. The decision at time $t_i$ is based on the pooled demand, not at the SKU level. For that reason, as the volatility increases, the increase in $Pr(\gamma_i > 0)$ becomes lower than the increase in $Pr(\widehat{\gamma}_i > 0)$. Plugging these dynamics into Equation (91), we conclude that the cost threshold value increases as the volatility increases.

**Proof of Corollary 3**

The proof follows directly from the proof of Corollary 2. Similar to Corollary 2, as the number of products increases, so do the values of $Pr(\gamma_i > 0)$, $Pr(\gamma_{i+1} > 0)$, $Pr(\widehat{\gamma}_i > 0)$, $Pr(\widehat{\gamma}_{i+1} > 0)$. The increase in $Pr(\gamma_{i+1} > 0)$ is the same as the increase in $Pr(\widehat{\gamma}_{i+1} > 0)$. However, $Pr(\gamma_i > 0)$ is less affected by the number of products than $Pr(\widehat{\gamma}_i > 0)$. Plugging these dynamics into Equation (91), we conclude that the cost threshold value increases as the number of products increases.

**Proof of Corollary 4**

Suppose the product proliferation occurs at echelon $j$. The number of components that can be produced at echelon $j$ is equal to $|\Theta_j|$. We now consider the case in which the value of $c_j$ increases with $|\Theta_j|$.

Corollary 3 demonstrates that the $\kappa$ value increases with the value of $|\Theta_j|$. Equation (86) indicates that the $\kappa$ value increases with the cost of the operation that causes the proliferation (i.e., $c_{i+1}$ in Equation (86)). Combining Corollary 3 with Equation (86), hence, the value of postponing the proliferation is amplified if $c_j$ increases as a result of increasing number of products (i.e., $|\Theta_j|$).

It follows from Equations (59) and (60) that the expected profit increases with $|\Theta_j|$ because the number of constraints increases with $|\Theta_j|$. This leads to an increase in the value of the dual parameter $\gamma_k$, which in turn increases the expected profit. However, the expected profit decreases with $c_j$. Therefore, there is a trade-off between increasing the number of products and increasing the cost value. This trade-off determines the boundaries of product portfolios such that adding too many products to a portfolio may hurt profits.