Computers & Industrial Engineering 66 (2013) 643–652

Contents lists available at ScienceDirect

Computers & Industrial Engineering

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

A queuing model on supply chain with the form postponement strategy ☆

Wenhui Zhou [a], Renqian Zhang [b,*], Yongwu Zhou [a]

[a] School of Business Administration, South China University of Technology, Guangzhou 510640, China
[b] School of Economics and Management, Beihang University, Beijing 100191, China

ARTICLE INFO

Article history:
Received 6 January 2012
Received in revised form 5 August 2013
Accepted 21 September 2013
Available online 30 September 2013

Keywords:
Form postponement
Customer order decoupling point
Mass customization
Queuing network

ABSTRACT

The form postponement (FP) strategy is an important strategy for manufacturing firms to utilize to achieve a quick response to customer needs while keeping low inventory levels of finished products. It is an important and difficult task to design a supply chain that uses FP strategy to mitigate the conflict between inventory level and service level. To this end, we develop a two-stage tandem queuing network to model the supply chain. The first stage is the manufacturing process of the undifferentiated semi-finished product, which is produced on a Make-To-Stock basis: the inventory is controlled by base-stock policy. The second stage is the customization process based on customers' specified requirements. There are two types of order: ordinary order and special order. The former can be met by customizing from semi-finished product, while the latter must be entirely customized beginning from the first stage. The customer orders arrive according to a Poisson process. We first derive the inventory level and fill rate, and then present a total cost model. It turns out that the model is intractable due to the Poisson distribution in the objective function. To analytically solve the problem, we use normal distribution as an approximation of the Poisson distribution, which works well when the parameter of the Poisson distribution is quite large. Finally, some numerical experiments are conducted and managerial insights are offered based on the numerical results.

1. Introduction

Nowadays, more and more companies are enlarging product varieties in order to fulfill demand from increasingly different types of customers. Favorably, information techniques make the diversification of product feasible by providing companies with low cost platforms to interact with their customers and realize mass customization. However, product variety has a significant impact on inventory level and service performance (Lee & Tang, 1997). To offer a large variety of products in highly efficient ways, various supply chain structures have been previously explored. Most of them can be divided into two strategies (Zinn & Bowersox, 1988): One is the time postponement (TP) strategy which delays delivery until customer orders arrive. The other is the form postponement (FP) strategy which delays the differentiation of the product until the detailed specification is confirmed.

Form postponement is one of the most popular and successful strategies in mass-customizing supply chains (Lampel & Mintzberg, 1996;Ahlstrom & Westbrook, 1999). In practice, many companies have successfully implemented the FP strategy, e.g., Dell computer, Toyota's "Build your Toyota", Amazon's "Built your own ring", and Nike's "Design your shoes", etc. For maximizing efficiency of the FP strategy, companies are showing increasing interest in incorporating the customer order decoupling point (CODP) as an important input to the strategic design of manufacturing operations as well as supply chains. CODP is defined as the point in the value-adding chain that separates the decision based on forecast from the decision based on the detailed product specification of the order. In other words, CODP divides the material flow that is forecast-driven (upstream of the CODP) from the flow that is customer order-driven (downstream of the CODP). It is also referred to as "the point of differentiation" (Lee & Tang, 1997).

Since Buclin (1965) first introduced the term "postponement", there have been a large number of researches on the postponement strategy. We do not attempt to cite and discuss every significant contribution in this area. Instead, we refer readers to van Hoek (2001), Swaminathan and Lee (2003), Yang and Burns (2003) for a comprehensive review. More recently, Leung and Ng (2007) use a goal programming model to optimize production planning in a perishable supply chain with postponement. Kumar, Nottestad, and Murphy (2009) investigate the effect of product postponement on distribution network supply chains by using simulation models. Trentin, Salvador, Forza, and Rungtusanatham (2011) develop an operational procedure to identify and quantify the opportunities

☆ The manuscript was handled by the Area Editor Qiuhong Zhao.
* Corresponding author. Tel.: +86 10 82338412.
E-mail address: zhangrenqian@buaa.edu.cn (R. Zhang).

0360-8352/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cie.2013.09.022

for applying the FP strategy to a given product family. Wong, Potterb, and Naimb (2011) show that the postponement strategy can improve the performance of the soluble coffee supply chain. Sharda and Akiya (2011) investigate the inventory management policy for a specific chemical plant by using a postponement strategy simulation.

Here we focus on a few studies that are the most pertinent to our own work, i.e., the joint optimization of CODP and the inventory level in a mass-customizing supply chain. Aviv and Federgruen (2001a, 2001b) investigate the tradeoff between the inventory level and redesigning cost in a form postponement supply chain, but they do not consider the problems of congestion and order delay. Conversely, Su, Chang, and Feiguson (2005), Gupta and Benjaafar (2004) and Jewkes and Alfa (2009) all capture the impact of congestion on the FP strategy by using queuing models. Su et al. (2005) compare the TP strategy with the FP strategy based on total operational cost. In their paper, the FP supply chain is actually modeled as a two-stage Make-To-Stock (MTS) queuing network with exogenous CODP position. They assume that there are n categories of customizing processes in the downstream stage, which are also controlled by the base-stock policy. Both Gupta and Benjaafar (2004) and Jewkes and Alfa (2009) model the customizing process as an Make-To-Order (MTO) queue that incorporates CODP position optimization. The former assumes that the potential CODP position in a multi-stage supply chain is a discrete number. The latter constructs a two-stage tandem queuing network in which the CODP position is relaxed to be continuous number on the interval of $(0,1)$.

In this paper, we address the same basic question as Gupta and Benjaafar (2004) and Jewkes and Alfa (2009): How to optimize the CODP position and inventory level to minimize operational cost? Here, we develop a two-stage tandem queuing network to model the supply chain using an FP strategy. The first stage is the manufacturing process of the undifferentiated semi-finished product, which is produced on a Make-To-Stock (MTS) basis and the inventory is controlled by the base-stock policy. The second stage is the customization process based on customers' specific requirements. However, our model differs from Gupta and Benjaafar (2004) and Jewkes and Alfa (2009) in the following ways: First of all, we assume that the processing time (both replenishment process and customizing process) are constant, instead of exponential distributed in Gupta and Benjaafar (2004) and Jewkes and Alfa (2009). This assumption is practicable in some cases, e.g., in automatic production lines. It is shown that the performance evaluation of two stage tandem queuing network with mixed MTS and MTO is very difficult, even in case of the exponential distributed process time. In our work, we derived the closed-form performance measures based on the results of Zipkin (2000) and Sherbrooke (1975), such as inventory level and unfill rate. Secondly, we consider the effect of CODP position on the capability of customization. It is clear that the further downward the CODP position, the more customer orders cannot be met based on semi-finished product. We model this situation with two categories of order: ordinary order and special order. The former can be met by using semi-finished product, while the latter must be entirely customized beginning from the first stage. Furthermore, we assume that the fraction of ordinary customer orders γ is a decreasing function of CODP position θ. Third, we involve the lead-time quotation policy and the penalty cost of tardiness for being more practical.

The rest of the paper is organized as follows. In Section 2, we present the model description. Section 3 presents the optimization problem. The approximation of the cost function by normal distribution and the solution of the approximate model are given in Sections 4 and 5, respectively. Section 6 conducts numerical experiments to demonstrate the impact of the parameters on the optimal policy. Section 7 concludes the paper.

2. Model description

We consider a mass-customizing supply chain that adopts the FP strategy. The entire manufacturing process is constant (say L) and additively separable, where "additively separable" means that the process can be interrupted at any time and continued just like without interruption. For instance, the manufacturing process is interrupted at θL ($0 \leqslant \theta \leqslant 1$), then when the process is continued, it just takes $(1-\theta)L$ to complete the entire process. Here, we refer to θ as the CODP position. In other words, the manufacturing process of the product is composed of two sub-processes: One is the manufacturing process for the undifferentiated semi-finished product, the time of which is equal to θL. In this stage, the semi-finished product is produced on a Make-To-Stock (MTS) basis and the inventory is controlled by the base-stock policy, with base-stock level S. The other sub-process is the manufacturing process of customization based on the semi-product, which is started after the order arrives and the detailed product specification is confirmed. Hence, the customization process runs based on a Make-To-Order (MTO) basis, and the customization processing time is equal to $(1-\theta)L$. We can easily envision two extreme cases of the policy. In the first case, $\theta = 0$ implies that the company adopts the pure MTO strategy. In the second case, $\theta = 1$, implies that the company adopts the pure MTS strategy. Additionally, the MTS process leads to inventory holding costs: denote $C(\theta)$ as the unit holding cost for the semi-finished product. It is well known that the later the CODP position, the larger the unit inventory holding cost will be, which means that $C(\theta)$ is an increasing function of θ. Denote C_h as the unit holding cost of finished product. It is clear that $C(\theta)$ must satisfy the following condition:

$$\begin{cases} C(1) = C_h, \\ C(0) = 0. \end{cases} \tag{1}$$

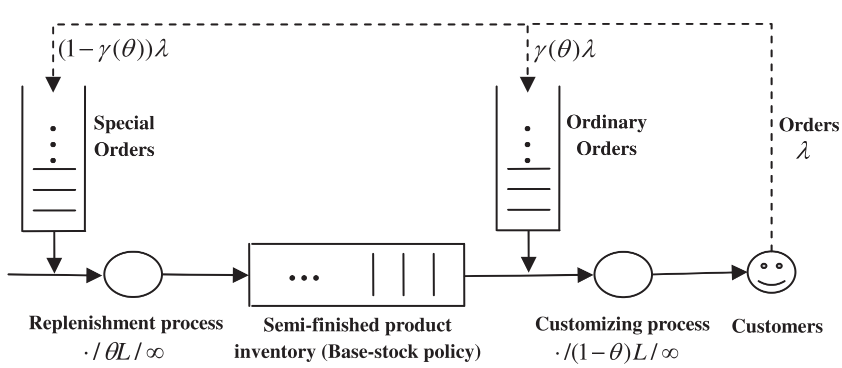

Fig. 1. The structure of supply chain with FP strategy.

We further assume that the replenishment and the customization are both uncapacitated. That means there are infinite servers for replenishment and customization in the system. Therefore, the supply chain acts as a two-stage tandem queuing network $\cdot/\theta L/\infty \to \cdot/(1-\theta)L/\infty$, as shown in Fig. 1.

The customer orders arrive according to a Poisson process with an arriving rate λ. Denote ρ as the utilization of the supply chain. It is easy to see that $\rho = \lambda L$. There are two categories of orders. One is the *ordinary order*, which can be met by using semi-finished product to customize if the semi-finished products are available. The other is the *special order* that must be entirely customized through the whole manufacturing process and cannot be satisfied by customization based on semi-finished product, even if the semi-finished product is not stocked out.

Let γ denote the probability that an arbitrary arriving customer order is ordinary. In other words, γ is the fraction of ordinary customer order among all the customer orders. In fact, γ reflects the likelihood of customer orders being satisfied by the customized products based on the semi-finished products. Intuitively, the later the CODP position, the more likely that an arrival order is of special type, and the less chance that an order can be met by customization based on semi-finished product. Thus, γ depends on the CODP position (denoted by θ). Hence, we reasonably assume that γ is a decreasing function with respect to θ, which is denoted by $\gamma(\theta)$. In practice, $\gamma(\theta)$ can be statistically approximated using historical data. Denote p_c as the probability that an arbitrary customer will be satisfied with the stock of finished product. We also denote p_c as the demand commonality. It is clear that $\gamma(\theta)$ must satisfy the following condition:

$$\begin{cases} \gamma(0) = 1, \\ \gamma(1) = p_c. \end{cases} \tag{2}$$

Note that when $\theta = 0$, the system is purely MTO and there is no semi-finished product. However, all the customer must be met so that γ must be equal to 1. Thus, we should write Eq. (2) as $\lim_{\theta\to 0^-}\gamma(\theta) = 1$. For concise, we write it as $\gamma(0) = 1$, which is just used for calibrating function $\gamma(.)$.

Most of mass-customizing companies implement a lead-time quotation policy. The lead-time quotation is the promised delivery time to the customer. Usually, the market competitive equilibrium could lead all companies to offer the same lead-time quotation. Hence, we introduce into our model a uniform lead-time quotation, T, which is an exogenous variable. We also consider a penalty cost per order (denoted by C_p) for the delayed orders if the actual delivery time is larger than the quotation, T. We further rule out a trivial case of $T \geqslant L$, because in this case the order will never be delayed, even though the firm adopts the pure MTO strategy. Consequently, we shall assume $T < L$ throughout the paper.

3. The decision model

One of the decision problems for a mass-customizing company is to select an appropriate CODP position θ and a base-stock level S to minimize the total cost. The total cost contains two parts: One is the inventory holding cost and the other is the penalty cost due to tardiness. Hence, the objective function of the decision problem can be mathematically stated as

$$\min_{\theta,S}\ \pi(\theta,S) = C(\theta)I + C_p\lambda\overline{F}, \tag{P1}$$

where I is the expected inventory level of the semi-finished product and $\overline{F}$ is the expected unfill rate, i.e., the proportion of the customers that are not satisfied within lead-time quotation.

For the sake of convenience, we define $G_p(\cdot|\lambda)$ as the cumulative distribution function (CDF) of Poisson distribution with λ, and denote $\overline{G_p}(\cdot|\lambda) = 1 - G_p(\cdot|\lambda)$.

3.1. Expected semi-finished product inventory level I

Since the special orders do not consume the semi-finished products, then only the arrivals of the ordinary orders trigger the inventory replenishment. Notice that the ordinary order arriving process is a random partition of Poisson process. Hence, it is also a Poisson process with rate $\gamma(\theta)\lambda$ (see Wolff, 1989, pp:74-76). Further according to Zipkin (2000, pp:181-186), it follows that for the base-stock policy, the expected inventory in system is basically the expected surplus inventory over the lead time, i.e.,

$$I = \sum_{j=0}^{S-1} G_p(j|\gamma(\theta)\lambda\theta L). \tag{3}$$

3.2. Unfill rate $\overline{F}$

Unfill rate is the proportion of the customers whose orders cannot be delivered within lead-time quotation. For $T < L$, we know that the unfill rate of the special orders always equals 1, so we just need to derive the unfill rate of the ordinary orders. Denote W the waiting time of an ordinary customer order. It is easy to see $W = W_1 + (1-\theta)L$, where W_1 is the waiting time for the replenishment. Recall that the arriving process of the ordinary order is a Poisson process with rate $\gamma(\theta)\lambda$, and then for the base-stock policy the CDF of W_1 (denoted by $H(x)$) is given by

$$H(x) = G_p(S-1|\gamma(\theta)\lambda(\theta L - x)). \tag{4}$$

Thus the unfill rate is given by the following proposition (see Sherbrooke (1975)).

Proposition 1.

$$\overline{F} = \begin{cases} 1, & \theta < 1 - \frac{T}{L} \\ 1 - \gamma(\theta)G_p(S-1|\gamma(\theta)\lambda(L-T)), & \theta \geqslant 1 - \frac{T}{L}. \end{cases}$$

Proof. See Appendix A. □

Proposition 1 shows that if $\theta < 1 - \frac{T}{L}$, then $\pi(\theta,S) = C(\theta)I + C_p\lambda$. The first term is always increasing in S and the second term is independent of S, so the optimal inventory level $S^*(\theta) = 0$ with given θ in case of $\theta < 1 - \frac{T}{L}$. Therefore, for $\theta < 1 - \frac{T}{L}$, the optimal strategy is MTO, i.e., $\theta^* = 0$.

When $\theta \geqslant 1 - \frac{T}{L}$, then the objective function is recast as

$$\pi(\theta,S) = C(\theta)\sum_{j=0}^{S-1} G_p(j|A(\theta)) + C_p\lambda[1 - \gamma(\theta)G_p(S-1|B(\theta))], \tag{5}$$

where $A(\theta) = \gamma(\theta)\theta L\lambda$ and $B(\theta) = \gamma(\theta)\lambda(L-T)$ are the parameters of the corresponding Poisson distributions.

The right hand side of Eq. (5) involves the CDFs of S Poisson distributions. We cannot get any convexity of objective function with respect to S or θ.

4. Model approximation by normal distribution

As mentioned above, the exact model (5) is difficult to be solved accurately. In order to reveal how the behavior of the optimal solution changes with parameters, we use normal distribution to approximate Poisson distribution. It is well known that when the parameter of Poisson distribution is large enough, the normal distribution approximates Poisson distribution very well. In practice, λ is usually quite large in mass-customizing situations. Furthermore, by $L > T$ and $\theta \geqslant 1 - \frac{T}{L}$, it follows that $A(\theta)$ and $B(\theta)$ are large, and thus our approximation is relevant. According to Zipkin (2000, pp. 206–209), the mean inventory on hand determined by Eq. (3) has the approximation as

$$I \approx \Phi^1\left(\frac{-S+A(\theta)}{\sqrt{A(\theta)}}\right)\sqrt{A(\theta)} = (S-A(\theta))\Phi_0\left(\frac{S-A(\theta)}{\sqrt{A(\theta)}}\right) + \sqrt{A(\theta)}\phi_0\left(\frac{S-A(\theta)}{\sqrt{A(\theta)}}\right), \quad (6)$$

where $\Phi^1(.)$ is the standard normal loss function, $\Phi_0(.)$ and $\phi_0(.)$ are the CDF and the probability density function (PDF) of standard normal distribution, respectively.

Similarly, the unfill rate has the approximation

$$\overline{F} \approx 1 - \gamma(\theta)\Phi_0\left(\frac{S-B(\theta)}{\sqrt{B(\theta)}}\right). \quad (7)$$

Then the objective function recasts approximately as

$$\pi(\theta,S) \approx \tilde{\pi}(\theta,S) = C(\theta)\left[(S-A(\theta))\Phi_0(z_1(\theta,S)) + \sqrt{A(\theta)}\phi_0(z_1(\theta,S))\right] + C_p\lambda[1-\gamma(\theta)\Phi_0(z_2(\theta,S))],$$

where $z_1(\theta,S) = \frac{S-A(\theta)}{\sqrt{A(\theta)}}$, $z_2(\theta,S) = \frac{S-B(\theta)}{\sqrt{B(\theta)}}$. Consequently, we consider the approximate optimization problem as follows.

$$\min_{\theta,S} \quad \tilde{\pi}(\theta,S). \quad (P2)$$

5. Solution of the approximate model

In this section, we concentrate on model ($P2$). Here, the base-stock level S is taken as a continuous variable, instead of as discrete variable. In this case, we can analyze the concavity and convexity of $\tilde{\pi}$ with respect to S. For the sake of convenience, we first give the following notations:

$$Q(S) = \frac{\phi_0(z_1(\theta,S))}{(B(\theta)-S)\phi_0(z_2(\theta,S))},$$

$$D(\theta) = \frac{C_p\lambda\gamma(\theta)\sqrt{A(\theta)}}{C(\theta)[B(\theta)]^{3/2}}.$$

The following theorem reports the concavity and convexity of $\tilde{\pi}(\theta,S)$ with respect to S.

Theorem 1.

(1) *When* $Q(0) \geqslant D(\theta)$, $\tilde{\pi}(\theta,S)$ *is strictly convex with respect to S on* $(0,\infty)$;
(2) *When* $Q(0) < D(\theta)$, $\tilde{\pi}(\theta,S)$ *is strictly convex with respect to S on* $(S_1(\theta),\infty)$, *and is strictly concave with respect to S on* $(0,S_1(\theta))$, *where* $S_1(\theta)$ *is the unique root of the following equation on* $(0,B(\theta))$,

$$Q(S) = D(\theta).$$

Proof. See Appendix A. □

The above theorem shows that the concavity and convexity of $\tilde{\pi}(\theta,S)$ sensitively depends on the parameters. Theorem 1 also gives the clear structure of $\tilde{\pi}(\theta,S)$ on base-stock level S. It gives us the pathway to find the optimal base-stock level, with the given CODP position. The following theorem reports the method. To simplify the presentation of the following theorem, we first give four following conditions directly.

$$\frac{\partial\tilde{\pi}(\theta,S)}{\partial S}\Big|_{S=0} \geqslant 0 \iff \frac{1}{C_p} \geqslant \frac{\lambda\gamma(\theta)\phi_0(\sqrt{B(\theta)})}{C(\theta)\sqrt{B(\theta)}\Phi_0(-\sqrt{A(\theta)})}, \quad (C1)$$

$$\frac{\partial\tilde{\pi}(\theta,S)}{\partial S}\Big|_{S=0} < 0 \iff \frac{1}{C_p} < \frac{\lambda\gamma(\theta)\phi_0(\sqrt{B(\theta)})}{C(\theta)\sqrt{B(\theta)}\Phi_0(-\sqrt{A(\theta)})}, \quad (C2)$$

$$\frac{\partial\tilde{\pi}(\theta,S)}{\partial S}\Big|_{S=S_1(\theta)} \geqslant 0 \iff \frac{\Phi_0\left(\frac{S_1-A(\theta)}{\sqrt{A(\theta)}}\right)}{\phi_0\left(\frac{S_1-B(\theta)}{\sqrt{B(\theta)}}\right)} \geqslant \frac{C_p\lambda\gamma(\theta)}{C(\theta)\sqrt{B(\theta)}}, \quad (C3)$$

$$\frac{\partial\tilde{\pi}(\theta,S)}{\partial S}\Big|_{S=S_1(\theta)} < 0 \iff \frac{\Phi_0\left(\frac{S_1-A(\theta)}{\sqrt{A(\theta)}}\right)}{\phi_0\left(\frac{S_1-B(\theta)}{\sqrt{B(\theta)}}\right)} < \frac{C_p\lambda\gamma(\theta)}{C(\theta)\sqrt{B(\theta)}}. \quad (C4)$$

Theorem 2. *Given* $\theta > 1-\frac{T}{L}$, *the optimal inventory level* $S^*(\theta)$ *can be determined by investigating all the following cases:*

(1) *When* $Q(0) \geqslant D(\theta)$ *and condition (C1) holds, then* $S^*(\theta) = 0$.
(2) *When* $Q(0) \geqslant D(\theta)$ *and condition (C2) holds, then* $S^*(\theta)$ *is the unique root of the following equation on* $(0,\infty)$.

$$\frac{\Phi_0(z_1(\theta,S))}{\phi_0(z_2(\theta,S))} = \frac{C_p\lambda\gamma(\theta)}{C(\theta)\sqrt{B(\theta)}}. \quad (8)$$

(3) *When* $Q(0) < D(\theta)$ *and condition (C3) holds, then* $S^*(\theta) = 0$.
(4) *When* $Q(0) < D(\theta)$, *and conditions (C2) and (C4) hold, then* $S^*(\theta)$ *is the unique root of Eq. (8) on* $(S_1(\theta),\infty)$.
(5) *When* $Q(0) < D(\theta)$, *and conditions (C1) and (C4) hold, let* $S_2(\theta)$ *denote the unique root of Eq. (8) on* $(S_1(\theta),\infty)$. *Then* $S^*(\theta)$ *is determined by*

$$S^*(\theta) = arg \min_{S\in\{0,S_2(\theta)\}} \tilde{\pi}(\theta,S).$$

Proof. See Appendix A. □

Theorem 2 offers a quick algorithm for finding optimal base-stock level S^*, given that $\theta \in [1-\frac{T}{L},1]$. Recall that the optimal base-stock level is always zero when $\theta \in [0,1-\frac{T}{L}]$, so that the optimal base-stock level is completely solved with the given CODP. Unfortunately, there are no concavity or convexity properties of $\tilde{\pi}$ with respect to θ. Note that $\theta \in [0,1]$, so we can use grid search on θ for the optimal solution. Additionally, $S_1(\theta)$ can be computed by using a bisection method, and $S^*(\theta)$ and $S_2(\theta)$ can be found by using a golden section search.

6. Numerical experiments

In this section, we first investigate the effect of parameters on the optimal policy based on a large number of numerical experiments. As will be seen, patterns that emerge from these experiments are reported below as observations, where we focus, respectively, on: (i) the effect of the demand commonality; (ii) the effect of finished product's unit holding cost; (iii) the effect of the penalty cost; and (iv) the effect of utilization. We also comment that the insightful observations are important to operations managers. Finally, we conduct a numerical experiment to test the sensitivity of the optimal policy to the parameters, as well as the accuracy of the approximation model. Throughout the section, we set the values of the parameters as follows, unless otherwise stated: $\lambda = 1$, $L = 1$, $T = 0.6$, and $C_p = 5$, with which $\gamma(1-T/L)\lambda(L-T) = 0.4\gamma(0.4)$, and for $\theta \in (0.4,1]$, $A(\theta) \geqslant 0.4\gamma(0.4)$ and

$B(\theta) \geqslant 0.4\gamma(0.4)$. With the parameters assigned these values, the normal approximation works well.

6.1. The effect of the demand commonality p_c

We assume that $C(\theta)$ is a linear function, specifically, $C(\theta) = 2\theta$. However, we construct three types of $\gamma(\theta)$ as follows:

$$\begin{cases} \gamma_1(\theta) = 1-(1-p_c)\theta, & \textit{Linear}, \\ \gamma_2(\theta) = (1-p_c)(\theta-1)^2 + p_c, & \textit{Convex}, \\ \gamma_3(\theta) = (p_c-1)\theta^2 + 1, & \textit{Concave}. \end{cases}$$

It is easy to be verified that all the $\gamma_i(\theta)$ (i = 1, 2, 3) satisfy Condition (2). Furthermore, $\gamma_1(\theta)$ is a linear function, $\gamma_2(\theta)$ is a convex function, and $\gamma_3(\theta)$ is a concave function.

Fig. 7 shows that the optimal policy (θ^*,S^*) varies with p_c under the different shape of $\gamma(\theta)$. It is easy to observe that there are different changing behaviors of θ^* and S^* for the three types of $\gamma(\theta)$. Specifically, both θ^* and S^* are always increasing as the demand commonality p_c increases under $\gamma_3(\theta)$. However, the shapes of (θ^*,S^*) under $\gamma_1(\theta)$ and $\gamma_2(\theta)$ are not as same as that under $\gamma_3(\theta)$. There are two structures under $\gamma_1(\theta)$: when $p_c < 0.201$, the optimal CODP is a constant, i.e., $\theta^* = 1 - \frac{T}{L} = 0.4$, and in this case, S^* is increasing with respect to p_c; when $p_c \geqslant 0.201$, both θ^* and S^* are increasing in p_c. It is worth to note that $p_c = 0.201$ is a discontinuous point for both curves of θ^* and S^*. For $\gamma_2(\theta)$, when $p_c \leqslant 0.819$, the shapes of θ^* and S^* are similar to those under $\gamma_1(\theta)$, which shows that a threshold of p_c, 0.397, divides the line into two increasing curves. However, when $p_c > 0.819$, the optimal strategy of the supply chain is pure Make-To-Stock, i.e., $\theta^* = 1$. Furthermore, S^* is increasing with respect to p_c. We conclude these phenomena as Observation 1.

Observation 1. There exist two thresholds $\underline{p}_c$ and $\bar{p}_c$ ($\underline{p}_c < \bar{p}_c$ and $\underline{p}_c, \bar{p}_c \in [0,1]$), such that when $p_c < \underline{p}_c, \theta^* = 1 - \frac{T}{L}$ and S^* is increasing with respect to p_c; when $\underline{p}_c \leqslant p_c \leqslant \bar{p}_c$, both θ^* and S^* are increasing in p_c; when $p_c > \bar{p}_c, \theta^* = 1$ (the optimal policy is pure Make-To-Stock) and S^* is increasing in p_c.

The observation has important implications: when demand commonality is small, which means that the customers' differentiation becomes large, the decision maker should set CODP as small as possible, but it will not be smaller than $1 - \frac{T}{L}$, otherwise, all of the order will be tardy. When demand commonality is quite large, the decision maker of the supply chain favors later differentiation (larger θ^*) and keep higher inventory levels. It obviously makes the inventory holding cost increase, but at the same time, the response to the customers becomes quicker, thus avoiding higher penalty costs for delayed orders.

It is also observed in Fig. 7 that when the demand commonality becomes smaller ($p_c < 0.48$ in this numerical example), the optimal CODP position θ^* under a concave function is larger than that under a linear function, and the one under a linear function is larger than that under a convex function. While the relationships are just opposite when the demand commonality is larger ($p_c > 0.48$ here). The relationship between p_c and S^* is similar to that between p_c and θ^*. It is interesting to note that when $p_c = 1$, the optimal policies under different type of $\gamma(\theta)$ are the same: a Make-To-Stock (MTS) strategy with an optimal base-stock level 53 in these numerical examples.

6.2. Effect of finished product's unit holding cost C_h

In this subsection, we offer several numerical examples to illustrate the effect of the finished product's unit holding cost C_h on the optimal policy (θ^*,S^*). Here, we assume that $\gamma(\theta)$ is a linear function and set $p_c = 0.3$, that is, $\gamma(\theta) = 1 - 0.7\theta$. Similarly, we consider three types of $C(\theta)$ as follows:

$$\begin{cases} C_1(\theta) = C_h\theta, & \textit{Linear}, \\ C_2(\theta) = C_h\theta^2, & \textit{Convex}, \\ C_3(\theta) = 2C_h\theta - C_h\theta^2, & \textit{Concave}. \end{cases}$$

It is also easy to verify that all $C_i(\theta)$ (i = 1, 2, 3) satisfy Condition (1). Furthermore, $C_1(\theta)$ is a linear function, $C_2(\theta)$ is a strictly convex function and $C_3(\theta)$ is a strictly concave function.

Fig. 8 reports that the optimal policy (θ^*,S^*) changes with C_h under different types of $C(\theta)$. It is similar to Fig. 7 that the threshold of C_h (0.37 for $C_1(\theta)$, 1.08 for $C_2(\theta)$, 0.23 for $C_3(\theta)$) is the discontinuous point for both curves of θ^* and S^*. We observe that the behaviors of the optimal policy under three types of $C(\theta)$ are similar. To be more specific, when C_h is sufficiently small (i.e., $C_h < 0.37$ for $C_1(\theta)$,$C_h < 1.08$ for $C_2(\theta)$, $C_h < 0.23$ for $C_3(\theta)$), the optimal CODP position is a constant with value $1 - \frac{T}{L} = 0.4$ and S^* is decreasing with respect to C_h. On the other hand, when C_h becomes larger, θ^* is increasing and S^* is still decreasing in C_h. The result is reported by the following observation.

Observation 2. There exists a threshold $\overline{C}_h$, such that when $C_h \leqslant \overline{C}_h, \theta^*$ is a constant with value $1 - \frac{T}{L}$; when $C_h > \overline{C}_h, \theta^*$ is increasing in C_h. However, S^* is always decreasing with respect to C_h.

Our result is similar to Observation 5 of Gupta and Benjaafar (2004), where our optimal policy (θ^*,S^*) is equivalent to (k^*,b^*) in the work of Gupta and Benjaafar (2004). The interpretation could be: when the finished product's unit holding cost increases, more valuable semi-finished products are stocked, which induces higher unit holding cost. However, the optimal base-stock level is set to be a lower value, which results in lower inventory level. Therefore, the total inventory cost decreases.

6.3. Effect of penalty cost C_p

In the third numerical experiment, we show how the penalty cost affects the optimal policy (θ^*,S^*). To this end, we assume $C(\theta)$ is a linear function, specifically, $C(\theta) = 2\theta$. However, we consider the same three types of $\gamma(\theta)$ just as in the first numerical experiment. Furthermore, we set $p_c = 0.3$. According to Fig. 9, it is easy to obtain the following result.

Observation 3. There exists a threshold $\overline{C}_p$, such that when $C_p \leqslant \overline{C}_p, \theta^*$ is increasing in C_p and when $C_p > \overline{C}_p, \theta^*$ is decreasing in C_p. S^* is always increasing in C_p.

We recognize that the observation that when $C_p > \overline{C}_p$ then θ^* is decreasing in C_p, is a little "counter intuitive". The reason could be that the penalty cost of the tardy order is a constant and is independent of the actual tardy time. If this is the case, the effect of reducing total penalty cost by increasing θ^* is limited. Thus, as penalty cost C_p increases, the optimal strategy of the supply chain is to reduce the CODP position θ^* and keep higher inventory level, which can reduce unit holding cost, and finally result in lower total inventory cost.

6.4. Effect of utilization ρ

In what following, we focus on the effect of utilization ρ. Here we assume that both $C(\theta)$ and $\gamma(\theta)$ are linear function, i.e., $C(\theta) = 2\theta$ and $\gamma(\theta) = 1 - (1 - p_c)\theta$. Fig. 10 shows the effect of utilization ρ on the optimal policy. It is easy to obtain the following observation based on Fig. 10.

Observation 4. There exist two thresholds $\underline{\rho}$ and $\bar{\rho}$ ($\underline{\rho} < \bar{\rho}$), such that when $\rho < \underline{\rho}$, $\theta^* = 0$ and $S^* = 0$, that is, the optimal strategy of

Table 1
Sensitivity of optimal policy to λ, C_p, C_h, p_c.

λ	C_p	C_h	p_c	θ^*	θ^* (%)	S^*	S^* (%)
1.00	5.00	4.0	0.300	0.680	–	44.3	–
0.95	5.00	4.0	0.300	0.689	+1.3	42.2	−4.7
1.05	5.00	4.0	0.300	0.675	−0.7	46.4	+4.7
1.00	5.00	4.0	0.300	0.680	–	44.3	–
1.00	4.75	4.0	0.300	0.685	+0.7	44.3	0.0
1.00	5.25	4.0	0.300	0.679	−0.1	44.4	+0.2
1.00	5.00	4.0	0.300	0.680	–	44.3	–
1.00	5.00	3.8	0.300	0.679	−0.1	44.4	+0.2
1.00	5.00	4.2	0.300	0.685	+0.7	44.3	0.0
1.00	5.00	4.0	0.300	0.680	–	44.3	–
1.00	5.00	4.0	0.295	0.684	+0.6	44.5	+0.5
1.00	5.00	4.0	0.305	0.680	0.0	44.1	−0.5

the supply chain is pure Make-To-Order; when $\underline{\rho} \leqslant \rho \leqslant \bar{\rho}$, both θ^* and S^* are increasing in ρ; when $\rho > \bar{\rho}$, θ^* is decreasing while S^* is increasing in ρ.

The implication of this result is: when the supply chain is quite less congested, the decision maker of the supply chain should adopt MTO policy and not need to worry about the penalty cost due to the short waiting time. When the utilization increases, the base-stock level is always increasing, while the CODP position is increasing for a moderate utilization and decreasing for a large utilization. It is worthwhile to notice that our result is quite different from Table 2 in Jewkes and Alfa (2009). In their work, θ^* is always decreasing and K^* (corresponds to S^* in our work) is always increasing in λ. This pattern is the same to our result when the utilization is large. Although the optimal model in Jewkes and Alfa's work is quite different from ours, the pattern should be similar. Why is there so big difference between Jewkes and Alfa's work and ours? The truth would be that Jewkes and Alfa (2009) ignored the effect of utilization when it is small. From Table 2 of Jewkes and Alfa (2009), they only considered $\lambda \in (0.1, 0.9)$, equivalently, $\rho \in (0.1, 0.9)$ that is just a subset of $\rho \in [0,1]$ in our paper. Based on our observation, the behavior of the optimal policy is quite different just when the utilization is small.

6.5. Sensitivity analysis and approximation performance

In practice, λ, C_p, C_h, and p_c are usually estimated by using historical data. A high sensitivity of the optimal policy to parameters implies potentially serious inaccuracies for the decision. We therefore investigate the sensitivity of the optimal policy to the variations of these parameters. Basically, we assume that $C(\theta)$ and $\gamma(\theta)$ are linear functions of θ, say $C(\theta) = C_h\theta$ and $\gamma(\theta) = 1 - p_c\theta$.

Table 1 reports the changes of the optimal policy with ±5% variations of the parameters λ, C_p, C_h, and p_c. Overall, (θ^*, S^*) is not sensitive to C_p, C_h, and p_c (the changes of θ^* and S^* are all less than 1% when the parameters change by ±5% individually). Although S^* is more sensitive to the change of λ, the variations are not larger than 5%.

As mentioned above, in this paper, we optimize the approximate objective function $\tilde{\pi}(\theta, S)$, instead of the original cost function $\pi(\theta, S)$. This may lead to the deviation from the exact optimal policy. In the following, we investigate the performance of the approximation by comparing the optimal policy of $\pi(\theta, S)$ to that of $\tilde{\pi}(\theta, S)$. Let (θ_0^*, S_0^*) denote the optimal policy of $\pi(\theta, S)$ and we implement grid search to determine (θ_0^*, S_0^*). Obviously, θ is not larger than 1 so we will search θ_0^* on interval [0, 1] with a step of 0.001.

However, the upper bound of S_0^* is unknown. Considering that S^* is an approximation of S_0^*, we can estimate an upper bound of S_0^* based on S^*. Thus, we first solve the approximate model of $\tilde{\pi}(\theta, S)$ yielding S^*, and then select a number that is sufficiently larger than S^* as the upper bound of S_0^*. Specifically, for the two groups of examples in the following Table 2, we set the upper bound of S_0^* equal to 100 for the first group and 200 for the second group. Then, we search the optimal value of S from 1 to the upper bound with a step of 1.

For the sake of convenience, we define the absolute relative errors of θ^* and S^* as $\theta\% = 100 \times \frac{|\theta^* - \theta_0^*|}{\theta_0^*}\%$ and $S\% = 100 \times \frac{|S^* - S_0^*|}{S_0^*}\%$, respectively. From Table 2, we observe that the approximate policy on θ performs very well in average for both groups, but the approximation is not so good for S in the first group. However, we can see that when the optimal value of S^* becomes larger, the relative error becomes smaller. For the second group, the average relative error drops to 3.44%. The reason is that the example in the second group has a larger λ such that the Poisson parameters $A(\theta)$ and $B(\theta)$ is larger, where the normal distribution can approximate the Poisson distribution better.

Additionally, in Table 2, T_a and T_0 are the computational time of solving the approximate model $\tilde{\pi}(\theta, S)$ and solving the original model $\pi(\theta, S)$ by grid search, respectively. The computational time shows that the grid search method is time-consuming when the value of S becomes larger, but the approximate approach is very efficient and therefore might be more practical for solving real problems.

7. Conclusion

The form postponement strategy is an efficient tool to balance the tradeoff between high customization and quick response. In this paper, we developed a two-stage tandem queuing network with constant process time to evaluate the operational performance of a form postponement supply chain. Based on Zipkin' result and Shebrooke's result, we derived the closed-form performance measures, such as inventory level and unfill rate. By using normal approximation of Poisson distribution, we optimized

Table 2
Performance of the normal approximation.

λ	C_p	C_h	p_c	θ^*	θ_0^*	$\theta\%$	S^*	S_0^*	S %	T_a (s)	T_0 (s)
1	5	2	0.4	0.536	0.541	0.92	29.7	33	10.00	0.39	77.53
1	5	2	0.9	0.783	0.776	1.56	44.3	42	5.48	0.36	83.97
1	5	1	0.3	0.638	0.621	2.74	49.8	52	4.23	0.37	85.65
1	5	3	0.3	0.703	0.711	1.13	48.6	51	4.71	0.38	79.66
1	1	2	0.3	0.662	0.653	1.38	19.3	21	8.10	0.38	88.52
Avg.	–	–	–	–	–	1.55	–	–	6.50	0.38	83.07
10	10	0.2	0.4	0.687	0.675	1.78	122.7	127	3.39	0.39	325.8
10	10	0.2	0.9	0.923	0.941	1.91	152.3	148	2.91	0.38	342.5
10	10	0.1	0.3	0.621	0.602	3.16	128.5	123	4.47	0.37	322.8
10	10	0.3	0.3	0.635	0.643	1.24	119.2	124	3.87	0.39	332.1
10	10	0.2	0.3	0.626	0.641	2.34	121.8	125	2.56	0.39	319.2
Avg.	–	–	–	–	–	2.09	–	–	3.44	0.38	328.5

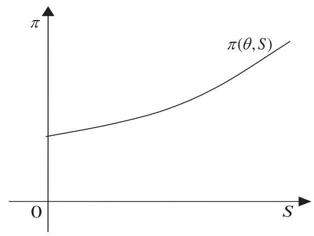

Fig. 2. Case 1.

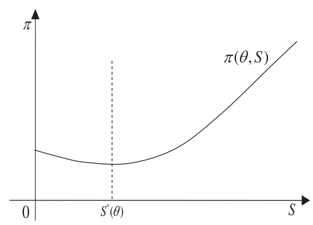

Fig. 3. Case 2.

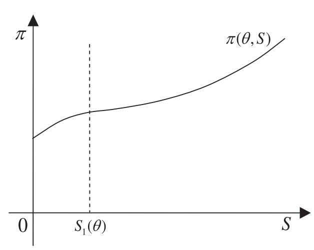

Fig. 4. Case 3.

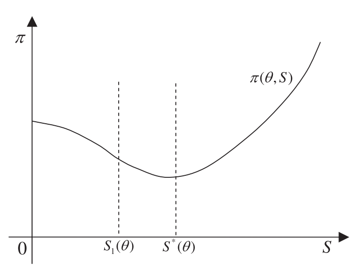

Fig. 5. Case 4.

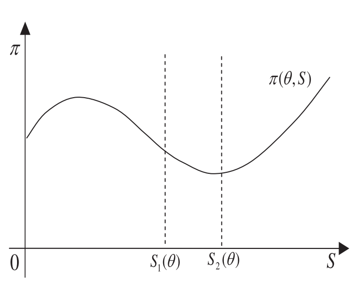

Fig. 6. Case 5.

the CODP position θ and the base-stock level S to minimize the total cost. Furthermore, we developed an efficient algorithm for finding the optimal policy. Our numerical examples show that the optimal policy is not sensitive to most of the decision parameters, except the demand rate. Based on the numerical results, we can gain managerial insights: (1) As the demand commonality increases, the optimal policy delays the CODP position and keeps higher base-stock levels; (2) As the finished product's unit holding cost increases, it is better for the supply chain to set the CODP position more closely to the finished product node of the supply chain and reduce the base-stock level; (3) For the policy of constant penalty cost, larger penalty costs causes a more forward CODP position with larger inventory levels; (4) When the system is quite less congested, the optimal policy for the supply chain is Make-To-Order, and when the utilization is moderate, the later differentiation is favored for larger load. However, when the system becomes quite congested, the CODP position decreases as the utilization increases.

Our research can be further extended along the following three lines: (1) To consider that the semi-finished product inventory is controlled by (s,S) policy; (2) To let the processing time be subject to an arbitrary distribution; (3) To relax the assumption of independent Poisson demand and allow for more complex demand structure, e.g. the demand rate depends on the price and lead-time quotation, which may be more practical.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant Nos. 71271089, 71271010, 70871040, and 71131003. The authors sincerely thank the reviewer and the editor for the valuable suggestions and comments, which have helped to improve the quality and presentation of the paper substantially.

Appendix A. Proofs

A.1. Proof of Proposition 1

Note that the unfill rate of the special orders is always equal to 1. Furthermore, we know that if $\theta \leqslant 1 - \frac{T}{L}$, then all the ordinary rders will be delayed such that $\overline{F} = 1$.

When $\theta > 1 - \frac{T}{L}$, we have

$$\begin{aligned}\overline{F} &= 1 - \gamma(\theta) + \gamma(\theta)Pr\{W > T\} \\ &= 1 - \gamma(\theta) + \gamma(\theta)(1 - Pr\{W_1 \leqslant T - (1-\theta)L\}) \\ &= 1 - \gamma(\theta)H(T - (1-\theta)L) \\ &= 1 - \gamma(\theta)G_p(S - 1|\gamma(\theta)\lambda(L - T)).\end{aligned}$$

The proof is now completed. □

To prove Theorem 2, the following lemma is necessary.

Lemma 1. *When $\theta > 1 - \frac{T}{L}$, $Q(S)$ is strictly increasing on the interval $(0, B(\theta))$ with respect to S.*

Proof. At first, we define function

$$K(x) \triangleq \frac{S - x}{\sqrt{x}}.$$

It is easy to see that when $S > 0$ and $x > 0$, we have

$$K'(x) = -\frac{S + x}{2x^{3/2}} < 0.$$

Note that when $\theta > \frac{L-T}{L}$, then $A(\theta) > B(\theta) > 0$. By the monotonicity of $K(\cdot)$ and $S < B(\theta)$, it follows that

$$z_1(\theta, S) = K(A(\theta)) < K(B(\theta)) = z_2(\theta, S) < 0.$$

Therefore,

$$Q'(S) = \frac{Se^{\frac{1}{2}\left(z_2^2 - z_1^2\right)} \frac{A(\theta) - B(\theta)}{A(\theta)B(\theta)}(B(\theta) - S) + e^{\frac{1}{2}\left(z_2^2 - z_1^2\right)}}{(B(\theta) - S)^2} > 0,$$

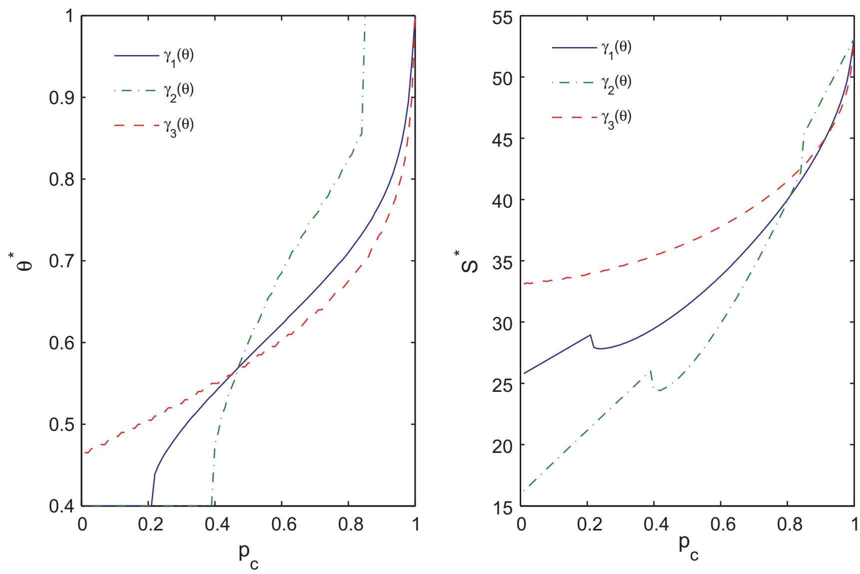

Fig. 7. The effect of the demand commonality p_c on the optimal policy.

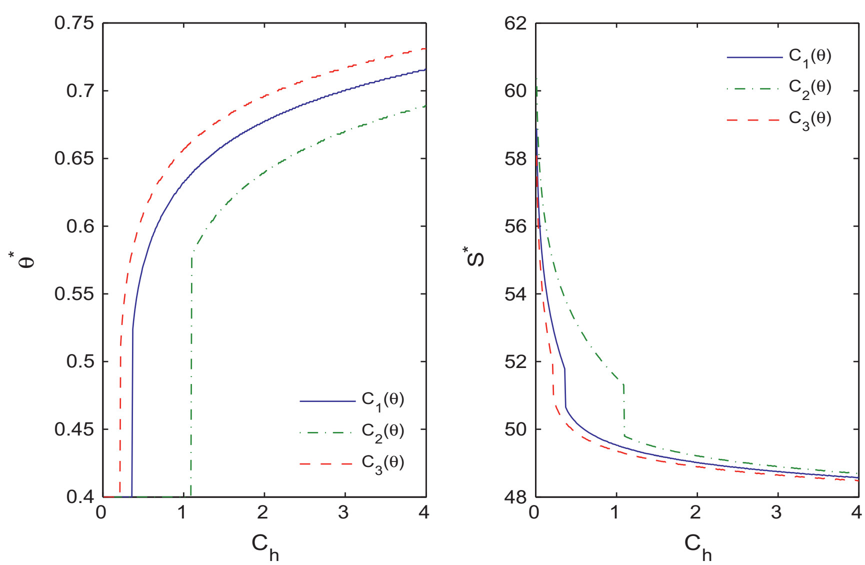

Fig. 8. The effect of finished product unit holding cost C_h on the optimal policy.

which completes the proof. □

A.2. Proof of Theorem 1

Taking the first and the second derivatives of $\tilde{\pi}$ with respect to S gives

$$\frac{\partial\tilde{\pi}}{\partial S}=C(\theta)\Phi_0\left(\frac{S-A(\theta)}{\sqrt{A(\theta)}}\right)-\frac{C_p\lambda\gamma(\theta)}{\sqrt{B(\theta)}}\phi_0\left(\frac{S-B(\theta)}{\sqrt{B(\theta)}}\right),$$

$$\frac{\partial^2\tilde{\pi}}{\partial^2 S}=\frac{C(\theta)}{\sqrt{A(\theta)}}\phi_0(z_1(\theta,S))+\frac{C_p\lambda\gamma(\theta)(S-B(\theta))}{[B(\theta)]^{3/2}}\phi_0(z_2(\theta,S)). \tag{9}$$

(1) It is easy to see that $\frac{\partial^2\tilde{\pi}}{\partial^2 S}>0$ for any given $S\geqslant B(\theta)$, which follows that $\tilde{\pi}(\theta,S)$ is strictly convex with respect to S on $[B(\theta),\infty)$.
In what following, we consider the case of $S\in[0,B(\theta))$. With some algebraic manipulation on Eq. (9), it follows that $Q(S)>D(\theta)$ is equivalent to $\frac{\partial^2\tilde{\pi}}{\partial^2 S}>0$, and $Q(S)<D(\theta)$ is equivalent to $\frac{\partial^2\tilde{\pi}}{\partial^2 S}<0$. Note that $Q(S)\to\infty$ as $S\to B(\theta)$. According to Lemma 1, it follows that if $Q(0)\geqslant D(\theta)$, then for any $S\in(0,\infty)$, we have $Q(s)>D(\theta)$, equivalently, $\frac{\partial^2\tilde{\pi}}{\partial^2 S}|_{S=0}\geqslant 0$. Hence, $\tilde{\pi}$ is strictly convex with respect to S on $(0,\infty)$.

(2) Note that $Q(0)<D(\theta)$ and $\lim_{S\to B(\theta)}Q(S)=\infty$. According to Lemma 1, it follows that $Q(S)=D(\theta)$ has a unique root $S_1(\theta)$ on $(0,B(\theta))$. Furthermore, we know that for $S\in(S_1(\theta),\infty)$, $Q(s)>D(\theta)$, and for $S\in(0,S_1(\theta))$, $Q(s)<D(\theta)$. Now the proof is completed. □

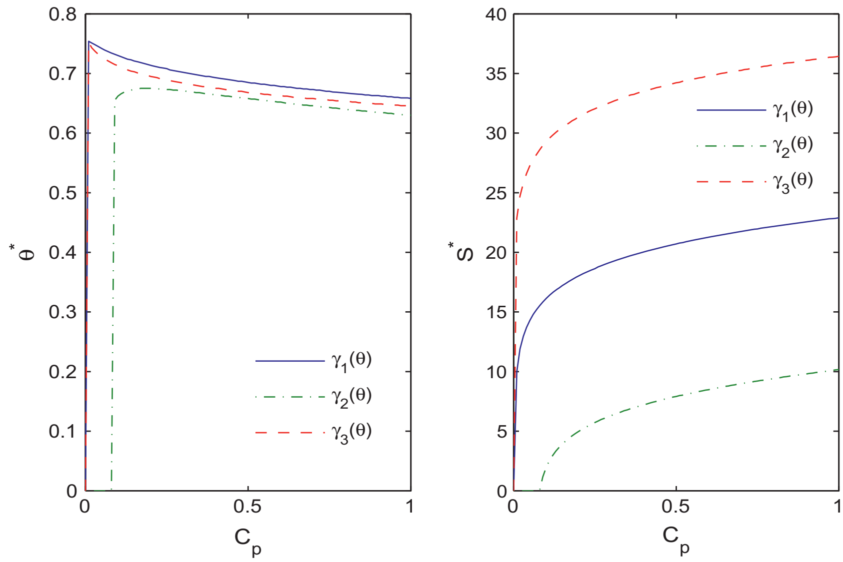

Fig. 9. The effect of penalty cost C_p on the optimal policy.

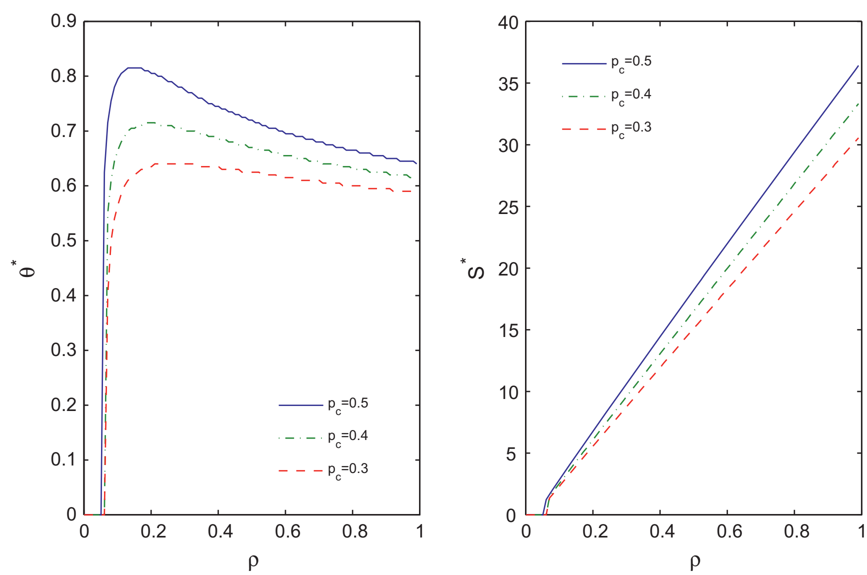

Fig. 10. The effect of utilization ρ on the optimal policy.

A.3. Proof of Theorem 2

Differentiating $\tilde{\pi}(\theta, s)$ with respect to S and setting it equal to 0, we have Eq. (8).

(Figs. 2–6 are supposed to be here.).

(1) See Fig. 2. According to Theorem 1, it follows that for any given $\theta(\theta > 1 - \frac{T}{L})$, when $Q(0) \geqslant D(\theta)$, then $\tilde{\pi}(\theta, S)$ is strictly convex with respect to S on $[0, \infty)$. Combining condition $(C1)$, i.e., $\frac{\partial \tilde{\pi}(\theta,S)}{\partial S}|_{S=0} > 0$, then $S^*(\theta) = 0$.

(2) See Fig. 3. This is the same as case (1) except condition $(C2)$, i.e., $\frac{\partial \tilde{\pi}(\theta,S)}{\partial S}|_{S=0} < 0$. Note that $\lim_{S\to\infty} \frac{\partial \tilde{\pi}(\theta,S)}{\partial S} = C(\theta) > 0$, so Eq. (8) has a unique root on $(0, \infty)$, which is exactly $S^*(\theta)$.

(3) See Fig. 4. According to Theorem 1, it follows that there exists a unique critical value $S_1(\theta) > 0$, such that $\tilde{\pi}(\theta, S)$ is strictly concave with respect to S on $[0, S_1(\theta))$ and strictly convex with respect to S on $(S_1(\theta), \infty)$. Combining this with condition $(C3)$, i.e., $\frac{\partial \tilde{\pi}(\theta,S)}{\partial S}|_{S=S_1(\theta)} \geqslant 0$, then $\tilde{\pi}(\theta, S)$ is not decreasing with respect to S on $[0, \infty]$, which yields $S^*(\theta) = 0$.

(4) See Fig. 5. Similar to case (3), we know that $\tilde{\pi}(\theta, S)$ is strictly convex with respect to S on $(S_1(\theta), \infty)$ and strictly concave with respect to S on $[0, S_1(\theta))$. Considering conditions $(C2)$ and $(C4)$, then $\tilde{\pi}(\theta, S)$ is strictly decreasing with respect to S on $[0, S_1(\theta))$. Note that $\lim_{S\to\infty} \frac{\partial \tilde{\pi}(\theta,S)}{\partial S} = C(\theta) > 0$, so Eq. (8) has a unique root on $(S_1(\theta), \infty)$, which is $S^*(\theta)$.

(5) See Fig. 6. Similar to the proof of case (4), we know that $\tilde{\pi}(\theta, S)$ is strictly convex with respect to S on $(S_1(\theta), \infty)$ and strictly concave with respect to S on $[0, S_1(\theta))$. From condition $(C4)$ and the fact $\lim_{S\to\infty} \frac{\partial \tilde{\pi}(\theta,S)}{\partial S} = C(\theta) > 0$, then $\tilde{\pi}(\theta, S)$ has a unique optimum on $(S_1(\theta), \infty)$, which is the unique root $S_2(\theta)$ of Eq. (8). By condition $(C1)$, it follows that the potential optima of $\tilde{\pi}(\theta, S)$ on $[0, S_1(\theta))$ are either 0 or $S_1(\theta)$. For the concavity of

$\tilde{\pi}(\theta, S)$, it is easy to see that $\tilde{\pi}(\theta, S_1(\theta)) > \tilde{\pi}(\theta, S^*(\theta))$. Hence, we just need to compare 0 with $S^*(\theta)$ to determine the optimal inventory level with given θ. □

(see Figs. 7–10)

References

Ahlstrom, P., & Westbrook, R. (1999). Implications of mass customisation for operations management: An exploratory survey. *International Journal of Operations and Production Management, 19*(3), 262–274.

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.

Buclin, L. (1965). Postponement, speculation and the structure of distribution channels. *Journal of Marketing Research, 2*(1), 26–31.

Gupta, D., & Benjaafar, S. (2004). Make-to-order, make-to-stock, or delay product differentiation? a common framework for modeling and analysis. *IIE Transactions, 36*(6), 529–546.

Jewkes, E. M., & Alfa, A. S. (2009). A queueing model of delayed product differentiation. *European Journal of Operational Research, 199*(3), 734–743.

Kumar, S., Nottestad, D. A., & Murphy, E. E. (2009). Effects of product postponement on the distribution network: A case study. *Journal of the Operational Research Society, 60*(4), 471–480.

Lampel, J., & Mintzberg, H. (1996). Customizing customization. *Sloan Management. Review, Fall*, 21–30.

Lee, H., & Tang, C. S. (1997). Modelling the costs and benefis of delayed product differntiation. *Management Science, 43*(1), 40–53.

Leung, S., & Ng, W. (2007). A goal programming model for production planning of perishable products with postponement. *Computers & Industrial Engineering, 53*, 531–541.

Sharda, B., & Akiya, N. (2011). Selecting make-to-stock and postponement policies for different products in a chemical plant: A case study using discrete event simulation. *International Journal of Production Economics, 23*(4), 819–820.

Sherbrooke, C. C. (1975). Waiting time in an $(s - 1, s)$ inventory system-constant service time case. *Operations Research, 23*(4), 819–820.

Su, J. C., Chang, Y. L., & Feiguson, M. (2005). Evaluation of postponement structures to accommodate mass customization. *Journal of Operations Management, 23*(3-4), 305–318.

Swaminathan, J., & Lee, H. (2003). *Handbooks in operations research and management science: Supply chain management: Design, coordination and operation., ch. Design for postponement* (pp. 199–228). Amsterdam: The Netherlands: Elsevier Publishers.

Trentin, A., Salvador, F., Forza, C., & Rungtusanatham, M. J. (2011). Operationalising form postponement from a decision-making perspective. *International Journal of Production Research, 49*(7), 1977–1999.

van Hoek, R. I. (2001). The rediscovery of postponement: A literature review and directions for research. *Journal of Operations Management, 19*(1), 161–184.

Wolff, R. (1989). *Stochastic Modeling and the Theory of Queues*. Englewood Cliffs, NJ: Prentice-Hall.

Wong, H., Potterb, A., & Naimb, M. (2011). Evaluation of postponement in the soluble coffee supply chain: A case study. *International Journal of Production Economics, 131*(1), 355–364.

Yang, B., & Burns, N. (2003). Implications of postponement for the supply chain. *International Journal of Production Research, 41*(9), 2075–2090.

Zinn, W., & Bowersox, D. (1988). Planning physical distibutiion with the priciple of postponement. *Journal of Business Logistics, 9*(2), 117–136.

Zipkin, P. H. (2000). *Foundations of Inventory Management*. New York: McGraw Hill.