Ann Oper Res (2016) 241:431–456
DOI 10.1007/s10479-013-1311-x

Timing order fulfillment of capital goods under a constrained capacity

Qing Li · Qi-Ming He · Xiaoli Wu

Published online: 17 January 2013

Abstract This paper studies the order-fulfillment process of a supplier producing multiple customized capital goods. The times when orders are confirmed by customers are random. The supplier can only work on one product at any time due to capacity constraints. The supplier must determine the optimal time to start the process for each order so that the total expected cost of having the goods ready before or after their orders are confirmed is minimized. We formulate this problem as a discrete time Markov decision process. The optimal policy is complex in general. It has a threshold-type structure and can be fully characterized only for some special cases. Based on our formulation, we compute the optimal policy and quantify the value of jointly managing the order fulfillment processes of multiple orders and the value of taking into account demand arrival time uncertainty.

Keywords Supply chain management · Markov decision process · Dynamic programming · Multi-item production/inventory · Stopping time

1 Introduction

In this paper, we study the order-fulfillment process of a supplier producing customized capital goods. It is critical for firms in capital goods industries to determine the proper time to start order-fulfillment processes. On one hand, their customers expect them to be responsive and the time that they are prepared to wait for the product is usually much shorter than

Q. Li
School of Business and Management, Hong Kong University of Science and Technology, Kowloon, Hong Kong SAR
e-mail: imqli@ust.hk

Q.-M. He
Department of Management Sciences, University of Waterloo, Waterloo, Ontario, Canada
e-mail: q7he@uwaterloo.ca

X. Wu (✉)
School of Business Administration, South China University of Technology, Guangzhou, P.R. China
e-mail: pearvoo@hotmail.com

the time needed to produce and deliver it. On the other hand, it is risky for the firms to start production before the customers confirm their orders due to inventory holding costs and the possibility of order cancellation. It is challenging, however, for firms to determine the best time to start their order-fulfillment processes. First, it is highly uncertain when the demand will arrive. For example, in the semiconductor equipment supply chain studied by Cohen et al. (2003), of 143 initial orders, 43 orders were later canceled and 76 experienced changes with respect to the delivery date. Second, firms typically supply multiple products (ordered by the same or different customers), which have different priorities. For example, customers for the new A380 superjumbo jet include, among others, Quantas Airways, Singapore Airlines, Air France and Emirates Airlines.[1] The priority depends on the products and the related services that the customers purchase and the contracts that they sign with the supplier. It may also depend on the customers' impact on the firm's long-term profitability, which determines how the customers will be treated by the supplier. Since different orders share the same capacity, their fulfillment processes must be coordinated.

These two dimensions of order-fulfillment complication have been observed in various context. Since its acquisition of Canadian Airlines in 2000, Air Canada has been dealing with union infighting among the pilots over seniority rankings. Air Canada placed orders with Boeing in April 2005, but made the purchase contingent on the pilots' approval of the labor deal. The source of uncertainty in this case comes from the labor dispute. In mid-June of 2005, a proposed labor pact was rejected by the Air Canada Pilots Association (ACPA) and, as a result, the orders with Boeing were in tatters. Air Canada finally lost two delivery slots because Boeing reallocated its capacity to processing orders from other airlines (Jang 2005). In making this decision, Boeing had to take into account various factors, including production lead time, uncertainty and priority. Suppliers in defense industries face similar problems (Mora et al. 2011; MacCormack and Mason 2005). They produce highly customized products that are constantly evolving.

Cohen et al. (2003), to the best of our knowledge, is the first paper to explicitly model the order-fulfillment processes of capital goods. They adopt a newsvendor model with the time to start the process, as opposed to the usual quantity, as the decision variable. They estimate the cost parameters based on empirical data. Their data reveal that the supplier perceives the holding cost and the cancellation cost to be about three and two times higher, respectively, than the delay cost. Motivated by the same problem, Li (2007) uses a different model that allows the supplier to start the process if the demand has been confirmed before the time that has been preset. He examines the effects of the lead time gap, the lead time gap uncertainty and the risk aversion of the supplier on the optimal policy. He also provides alternative explanations for the data presented by Cohen et al. (2003). The basic model in Li (2007) is based upon the framework of Katircioglu and Atkins (1996), who study the unit demand inventory problem. All the papers mentioned above consider only one product, or, equivalently, unlimited capacity. The products can therefore be managed independently.

The problem can be viewed as a multi-item production/inventory systems where orders arrive randomly one at a time (e.g., Ha 1997). Due to the difficulty in handling large dimensions of state space, this literature typically relies on two important assumptions. First, all times (i.e., arrival and production times) are exponentially distributed.[2] Second, there is an infinite horizon. Porteus (2002, Chap. 14) provides a detailed discussion about this line of

[1] http://www.usatoday.com/money/biztravel/2005-06-01-airbus-usat_x.htm?csp=N009.

[2] Ha (2000) is an exception. He allows the production (or processing) time to have an Erlang distribution and shows that a single-state variable called the work storage level can be used to capture completely the information regarding the inventory level and the status of the current production.

research and nicely explains what has been done, what has not been done and where the challenges remain. See also Benjaafar et al. (2005) for later development and references. While similar in spirit, our model is neither a consequence of, nor more general than, any work in such literature. In our model, we consider only one demand arrival for each product, which is typical in capital goods industries. Consequently, we can avoid the dimensionality problem and consider different distributions than only the exponential distribution. In particular, in our model, the production lead times are deterministic and can possibly be extended to any positive random variables. The arrival time (i.e., the time when an order is confirmed in our model) is also allowed to be more general than the exponential distribution.

Our work is also, in spirit, related to the literature on scheduling with earliness and tardiness penalties. Jia (2001) considers uncertain due date, but all jobs share a common due date, which is exponentially distributed. In Elhafsi (2002), when setting planned lead times, one also faces the earliness and tardiness trade-off. But his focus is on meeting the (deterministic) due date of a single job in a production systems with multiple processing stages. There is a vast literature on scheduling that assumes no holding cost of completing a job before it is due (Chaps. 10 and 11, Pinedo 2002). There is also a vast literature on due date setting (see Duenyas and Hopp 1995; Duenyas 1995; Wein 1991, and the literature review therein). In both, the trade-off between producing too early and too late, which is crucial in our work, is not modeled.

Specifically, in this paper, we consider the production of multiple products, which share the same facility and are ordered by different customers. The supplier can only work on one product at any time due to capacity constraints. Production requires a lead time. Production of any product, once started, cannot be interrupted until it is complete. The problem facing the supplier is to decide when to start the production of each product so as to minimize the total expected holding and penalty costs.

We formulate the problem as a discrete time Markov decision process. The structure of the optimal policy in general is very complex. The optimal policy has a threshold-type structure only for the case with two orders, one of which is confirmed and the other is not. In fact, even the formulation itself is nontrivial. The complexity is due to, first, the capacity linkage across different orders, and second, the uncertainty of demand arrival times. The formulation enables us to search for the policy that minimizes the total expected cost over an infinite horizon. If we drop either the capacity linkage across different orders, or both the capacity linkage and the uncertainty of demand arrival times, then the problem would become much simpler. However, as we show from numerical studies, the costs of doing those can be high most of the time. The cost of ignoring the capacity linkage is high when the total number of orders is large and the orders are very different from one another in their lead times. The cost of ignoring demand uncertainty is consistently high and it gets higher also when the orders are very different from one another in their lead times.

The rest of this paper is organized as follows. In Sect. 2, we introduce assumptions, notation, and the dynamic programming formulation of the model. In Sect. 3, we investigate the special case when there is only one unfilled order, which lays the groundwork for the general case. In Sect. 4, we consider the case when there are two unfilled orders, one of which is confirmed and the other is not and its confirmation time is random. In Sect. 5, we show by counter examples that there are no simple structure when there are multiple unconfirmed products. We then quantify the value of jointly managing the order fulfillment processes of multiple orders and the value of taking into account demand arrival time uncertainty. In Sect. 6, we discuss several extensions of our model and conclude the paper. All proofs are presented in Appendix.

2 Model description and formulation

Suppose that the supplier needs to fill n different orders, which share the same production capacity. Due to constraints on production capacity, the supplier can process at most one order at a time. Production of any product, once started, cannot be interrupted until it is finished. Production of the product i takes a lead time of l_i, for $1 \le i \le n$. We assume that the lead times are deterministic. At time 1, the order of product i will be confirmed at time X_i, which has a discrete distribution $P\{X_i = j\} = b^i_j$, $j \ge 1$, for $1 \le i \le n$. The random variable X_i is assumed to have a finite mean for all $1 \le i \le n$.

Two types of costs are considered: holding cost and delay penalty cost. If production is completed before the order is confirmed, then the supplier incurs a holding cost h_i per unit time; if the order is confirmed before the product is available on hand, then the supplier incurs a delay penalty cost π_i per unit time. Here we assume that the supplier starts to incur a delay penalty as soon as the order is confirmed if the product is not ready. This assumption is not critical and the issues of due date or customer order lead time will be discussed in Sect. 6. The penalty cost can take various forms. First, because the bulk of the purchase price is paid at the time of delivery, late delivery leads to late payment and the penalty cost is the time value of money. Second, the penalty cost can be compensation specified in contracts for late delivery. The compensation can either be paid via cash arrangements or noncash arrangements, such as discounts on spare parts purchases and crew training.[3] Finally, late delivery has a negative impact on firms' image, which will affect their future sales.

Let vector $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ represent the status of all orders. The status of an order i can be "production started" ($x_i = 2$), "order confirmed but production not started" ($x_i = 1$), or "order unconfirmed" ($x_i = 0$). Let F represent the set of orders whose production has been started and $\bar{F}$ the rest. Among all the orders in $\bar{F}$, let C and $\bar{C}$ be the set of orders that have been confirmed, and the set of orders that have not been confirmed, respectively. Then we have $\bar{F} = C \cup \bar{C}$. We use $|\cdot|$ for the cardinality of a set.

The timing of events in a period is as follows. First, the state is observed and a decision is made. Second, if the production is due to be completed in that period, then the product is available to fill demand. Third, orders, if any, are confirmed. Fourth, costs are incurred. Throughout the paper, we assume that if the production of product i is started in period s, the product will be available at the beginning of period $s + l_i$ before the order is confirmed in that period.

We assume that the planning horizon of the supplier is infinite. The decision that the supplier has to make is when to start the production of each product. At the beginning of each period, if the production capacity is unavailable (i.e., a production is on-going), there is no decision to be made. If the production capacity is available, the supplier decides whether or not to start the production of a product: "wait" or "produce". If the action is to produce, the supplier has to decide which product to produce.

Let $u(\mathbf{x}, s)$ be the minimal expected cost if the status of orders is $\mathbf{x}$ at the beginning of period s. The action space, for a given state $(\mathbf{x}, s)$, is to either wait, or start the production of one of the orders in $\bar{F}$. Let $w(\mathbf{x}, s, i)$ be the expected cost if action i is taken, where $i = 0$ means "wait" and $i \in \bar{F}$ means "start the production of product i". Since a decision is necessary only if there is no on-going production, the functions $u(\mathbf{x}, s)$ and $w(\mathbf{x}, s, i)$ are defined only for s such that the production capacity is available. In addition, as soon as the

[3] It is a common practice in the aerospace industry that late delivery penalties are paid via noncash arrangements (Matlack and Holmes 2005).

production of a product has been started, the expected costs associated with that order can be calculated. Thus, $u(\mathbf{x}, s)$ does not include the costs associated with any order $i \in F$. For ease of exposition in the following formulation, we arrange the sequence of orders such that the vector $\mathbf{x}$ can be written as $\mathbf{x} = (0, \dots, 0, 1, \dots, 1, 2, \dots, 2)$.

First, we have $u(\mathbf{x}, s) = 0$ if $\bar{F} = \emptyset$, i.e., all elements in $\mathbf{x}$ are equal to 2, or the production of all ordered has been started before period s.

For $\bar{F} \neq \emptyset$, we consider two cases.

(a) If $\bar{C} \neq \emptyset$, then we have

$$u(\mathbf{x}, s) = \min_{i \in \{0\} \cup \bar{F}} w(\mathbf{x}, s, i). \tag{1}$$

The expression for the cost function $w(\mathbf{x}, s, i)$ depends on the time s and the action taken. At the beginning of period s, if order i has not been confirmed yet, then the distribution of X_i will have to be updated and the probability that it will be confirmed j periods later is given by the following conditional distribution:

$$\frac{b^i_{s+j}}{\sum_{k \geq s} b^i_k}$$

for all $s \geq 1$ and $j \geq 0$. Let $\Omega(i)$ represent a set of all vectors of cardinality i and with elements either 0 or 1. For example, $\Omega(2) = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$. Given state $(\mathbf{x}, s)$, the first $|\bar{C}|$ state variables (i.e., the status of the unconfirmed orders) will become $\mathbf{y}$, where $\mathbf{y} \in \Omega(|\bar{C}|)$, with probability

$$p(\mathbf{y}, s) = \prod_{j=1}^{|\bar{C}|} \left(\left(\frac{b^j_s}{\sum_{i \geq s} b^j_i} \right)^{y_j} \left(1 - \frac{b^j_s}{\sum_{i \geq s} b^j_i} \right)^{1-y_j} \right),$$

where y_i is the ith element of $\mathbf{y}$.

If the action is "wait", the supplier needs to pay a unit of penalty cost for each confirmed order and for each order that will be confirmed in the current period; that is,

$$w(\mathbf{x}, s, 0) = \sum_{i \in C} \pi_i + \sum_{\mathbf{y} \in \Omega(|\bar{C}|)} p(\mathbf{y}, s) \left[\sum_{i=1}^{|\bar{C}|} y_i \pi_i + u(\mathbf{y}, 1, \dots, 1, 2, \dots 2, s+1) \right].$$

If the decision is to produce product $k \in \bar{F}$, then the capacity will be used until time $s + l_k$, which is the next decision epoch. The supplier needs to pay l_k units of penalty cost for each confirmed order. For the unconfirmed orders that will be confirmed before time $s + l_k$, the supplier also incurs penalty cost and its amount depends on the confirmation times. We consider two cases. Suppose that $k \in C \neq \emptyset$. Let $\Psi(k)$ represent the set of all vectors of cardinality k and with nonnegative elements. For example, $\Psi(2) = \{(0, 0), (0, 1), (1, 0), (1, 1), (2, 1), \dots\}$. Then,

$$w(\mathbf{x}, s, k) = l_k \sum_{j \in C} \pi_j + \sum_{\mathbf{y} \in \Psi(|\bar{C}|)} \left(\prod_{j=1}^{|\bar{C}|} \frac{b^j_{s+y_j}}{\sum_{i \geq s} b^j_i} \right) \left(\sum_{j=1}^{|\bar{C}|} (l_k - y_j)^+ \pi_j + u(\mathbf{z}, 2, \dots, 2, s + l_k) \right),$$

where $x^+ = \max\{0, x\}$ and $\mathbf{z}$ is an $|\bar{F}|$-dimensional vector and its jth element is given by

$$z_j = \begin{cases} 0, & \text{if } y_j \geq l_k,\ j \in \bar{C}; \\ 1, & \text{if } y_j < l_k,\ j \in \bar{C}; \\ 2, & \text{if } j = k; \\ 1, & \text{if } j \neq k,\ j \in C. \end{cases}$$

Suppose that $k \in \bar{C}$. Then,

$$w(\mathbf{x}, s, k) = l_k \sum_{j \in C} \pi_j + \pi_k \sum_{j=0}^{l_k - 1} (l_k - j) \frac{b_{s+j}^k}{\sum_{i \geq s} b_i^k} + h_k \sum_{j=l_k}^{\infty} (j - l_k) \frac{b_{s+j}^k}{\sum_{i \geq s} b_i^k}$$
$$+ \sum_{\mathbf{y} \in \Psi(|\bar{C}|-1)} \left(\prod_{j \in \bar{C},\, j \neq k} \frac{b_{s+y_j}^j}{\sum_{i \geq s} b_i^j} \right)$$
$$\times \left(\sum_{j \in \bar{C},\, j \neq k} (l_k - y_j)^+ \pi_j + u(\mathbf{z}, 2, \ldots, 2, s + l_k) \right),$$

where $\mathbf{z}$ is an $|\bar{F}|$-dimensional vector and its jth element is given by

$$z_j = \begin{cases} 0 & \text{if } y_j \geq l_k,\ j \in \bar{C},\ j \neq k; \\ 1 & \text{if } y_j < l_k,\ j \in \bar{C},\ j \neq k; \\ 2 & \text{if } j = k; \\ 1 & \text{if } j \in C. \end{cases}$$

(b) If $\bar{C} = \emptyset$, i.e., all orders have been confirmed, then waiting is suboptimal and the supplier should start the production of one of the products in C; then we have,

$$u(\mathbf{x}, s) = \min_{i \in \bar{F}} w(\mathbf{x}, s, i),$$

where

$$w(\mathbf{x}, s, i) = l_i \sum_{j \in C} \pi_j + u(\mathbf{z}, 2, \ldots, 2, s + l_i),$$

and $\mathbf{z}$ is an $|\bar{F}|$-dimensional vector and its ith element is 2 and the rest are all 1.

Let $a^*(\mathbf{x}, s)$ be the optimal action when the state is $(\mathbf{x}, s)$. That is,

$$a^*(\mathbf{x}, s) = \operatorname*{argmin}_{i \in \{0\} \cup \bar{F}} \{ w(\mathbf{x}, s, i) \}.$$

When different actions lead to the same cost, $a^*(\mathbf{x}, s)$ is defined as the largest number. For example, $a^*(\mathbf{x}, s) = 2$ when starting the production of product 2, product 1, or waiting all lead to the minimal cost.

Using the above formulation, a computational method can be developed for computing the minimum expected total costs and the optimal policy. To find $u(\mathbf{x}, s)$ and the optimal policy, we use induction to break down the problem into smaller and more manageable pieces. To illustrate the key idea, suppose $\mathbf{x} = (0, 0)$. Based on the above dynamic programming, to compute $u(0, 0, s)$, we need to know $u(0, 1, s)$, $u(1, 1, s)$, and $u(0, 2, s)$. To compute $u(0, 1, s)$, we need to know $u(1, 1, s)$ and $u(0, 2, s)$. To compute $u(1, 1, s)$ and $u(0, 2, s)$, we need to know $u(1, 2, s)$, which equals $l_1 \pi_1$.

3 Models with one unfilled order

To understand the form of the optimal policy, we start with the case with only one order to fulfill. If the order is already confirmed, then the optimal policy is to start production immediately and the minimal cost is $l_1 \pi_1$. The case when the order has not been confirmed was previously studied by Katircioglu and Atkins (1996) and Li (2007). Here we remodel it by using our general dynamic programming framework and lay the groundwork for the more complex problem with multiple products. In this case, we need to compute $u(\mathbf{x}, s)$, where $\mathbf{x} = (0, 2, 2, \dots, 2)$. To simplify notation, within this section we leave out the status of orders as state variables and write $u(\mathbf{x}, s)$ as $u(s)$ and $w(\mathbf{x}, s, a)$ as $w(s, a)$ for $a = 0, 1$. We also omit the superscripts or subscripts that are used to differentiate different orders. Then, (1) is simplified to

$$u(s) = \min\{w(s, 0), w(s, 1)\}, \tag{2}$$

where

$$w(s, i) = \begin{cases} \pi \sum_{j=0}^{l-1} (l - j) \frac{b_{s+j}}{\sum_{k \geq s} b_k} + h \sum_{j=l+1}^{\infty} (j - l) \frac{b_{s+j}}{\sum_{k \geq s} b_k} & \text{if } i = 1; \\ \frac{b_s}{\sum_{k \geq s} b_k} (l + 1)\pi + (1 - \frac{b_s}{\sum_{j \geq s} b_j}) u(s + 1) & \text{if } i = 0. \end{cases}$$

As we assume that the expectation of X is finite, $w(s, 1)$ is finite for all s. Consequently, $u(s)$, which is less than $w(s, 1)$, is finite. The problem is known as a stopping problem (see, for example, Ross 1983). The following definition is needed for the characterization of the optimal policy.

Definition 1 (Barlow and Proschan 1996) A discrete distribution $\{b_k\}_{k=1}^{\infty}$ has an increasing failure rate (IFR) if and only if $b_k / \sum_{i=k}^{\infty} b_i$ is increasing in k.

An equivalent condition for IFR is that $\sum_{i \geq k+j} b_i / \sum_{i \geq k} b_i$ is decreasing in k for any $j \geq 0$. Most common discrete distributions have IFR, including, for example, Binomial, Poisson, Geometric, and Negative binomial (r, p) with $r \geq 1$. Let

$$f(s) = \frac{b_s}{\sum_{i \geq s} b_i} (l + 1)\pi + \left(1 - \frac{b_s}{\sum_{j \geq s} b_j}\right) w(s + 1, 1) - w(s, 1).$$

Then $f(s)$ represents the difference in cost between starting production immediately and waiting for exactly one more period and then starting production. The following lemma shows a useful property of $f(s)$.

Lemma 1 *If $\{b_k\}_{k=1}^{\infty}$ has an IFR, then $f(s) \geq 0$ if and only if $s \geq s^*$, where*

$$s^* = \inf\{s : f(s) \geq 0\}. \tag{3}$$

The set $\{s : f(s) \geq 0\}$ represents the set of state for which starting production immediately is at least as good as waiting for exactly one more period and then starting production. In addition, s^* is the first time that the process enters the set $\{s : f(s) \geq 0\}$. By substituting the costs into $f(s)$, we can see that s^* is the smallest s such that

$$\frac{\sum_{j \geq s+l+1} b_j}{\sum_{j \geq s} b_j} - \frac{\pi}{\pi + h} \leq 0.$$

A direct consequence of the above lemma is that the set $\{s : f(s) \geq 0\}$ is a closed set. That is, if $i \in \{s : f(s) \geq 0\}$, then any $j \geq i$ satisfies $j \in \{s : f(s) \geq 0\}$. Proposition 1 gives the structure of the optimal policy.

Proposition 1 *Assume that $\{b_k\}_{k=1}^{\infty}$ has an IFR. Then, the optimal policy for (2) is to start production if and only if $s \geq s^*$.*

According to the above proposition, production should be started the first time the process enters a state in $\{s : f(s) \geq 0\}$ and s^* is the threshold for ordering. As the computation of s^* (or that of $f(s)$) involves only the one-step cost, the policy is a *one-step-look-ahead* policy (Ross 1983). As a special case, when X follows a Geometric distribution, which has the memoryless property, s^* is either zero or infinity. That is, the optimal policy is either to start production or wait until the order is confirmed and exactly which action is optimal only depends on the cost parameters and the lead time, but not on the state. In addition, the optimal expected cost $u(s)$ is also independent of the state.

4 Models with one confirmed order and one unconfirmed order

We now consider two unfilled orders in this section. Suppose at the beginning of time s, the order for product 2 has been confirmed, but the order for product 1 has not. In this case, we need to compute $u(\mathbf{x}, s)$, where $\mathbf{x} = (0, 1, 2, \dots, 2)$. To simplify notation, within this section, we write X_1 and b_j^1 as X and b_j, respectively. We do, however, use $\{\pi_1, \pi_2, l_1, l_2, h_1\}$ for the penalty costs, production times, and holding cost for orders 1 and 2. We also remove the last $N - 2$ elements in the state vector $\mathbf{x}$; for example, we write $u(0, 1, 2, \dots, 2, s)$ as $u(0, 1, s)$.

The dynamic programming recursion (1) becomes

$$u(0, 1, s) = \min\big\{w(0, 1, s, 0), w(0, 1, s, 1), w(0, 1, s, 2)\big\}, \tag{4}$$

where

$$w(0, 1, s, 0) = \left(1 - \frac{b_s}{\sum_{i \geq s} b_i}\right)\pi_2 + \frac{b_s}{\sum_{i \geq s} b_i}\big[(l_1 + 1)\pi_1 + (l_2 + 1)\pi_2 + \min\{l_1\pi_2, l_2\pi_1\}\big] + \left(1 - \frac{b_s}{\sum_{j \geq s} b_j}\right)u(0, 1, s + 1),$$

$$w(0, 1, s, 1) = \pi_1 \sum_{j=0}^{l_1 - 1}(l_1 - j)\frac{b_{s+j}}{\sum_{i \geq s} b_i} + h_1 \sum_{j=l_1+1}^{\infty}(j - l_1)\frac{b_{s+j}}{\sum_{i \geq s} b_i} + (l_1 + l_2)\pi_2,$$

and

$$w(0, 1, s, 2) = l_2\pi_2 + \frac{\sum_{j=0}^{l_2 - 1}(l_1 + l_2 - j)b_{s+j}}{\sum_{i \geq s} b_i}\pi_1 + \frac{\sum_{j=l_2}^{\infty} b_{s+j}}{\sum_{i \geq s} b_i}u(0, 2, s + l_2).$$

Here $u(0, 2, s + l_2)$ is for the expected costs of order 1 starting from period $s + l_2$.

When the order for product 1 is confirmed, the supplier should start production immediately and the supplier must decide which product to start first. The choice depends on the ratio of the delay cost to the lead time, i.e., π_i/l_i (see the expression for $w(0, 1, s, 0)$). In other words, the ratio of the delay cost to the lead time determines the priority of the orders. Therefore, for ease of exposition, we say that the order for product 1 has a higher priority over the order for product 2 if $\pi_1 l_2 - \pi_2 l_1 \geq 0$ and the order for product 2 has a higher priority over product 1 otherwise. This condition is similar to the celebrated $c\mu$ rule for scheduling customers of different types in a potentially congested service system, where c is the waiting cost rate and μ is the service rate. A similar condition has also been used in the literature of make-to-stock queues (see, for example, Porteus 2002).

4.1 The order for product 2 has a higher priority

In this subsection, we consider the case when product 2 has a higher priority (i.e., $\pi_2 l_1 \geq \pi_1 l_2$). Let

$$f_2(s) = \left(1 - \frac{b_s}{\sum_{i\geq s} b_i}\right)\pi_2 + \frac{b_s}{\sum_{i\geq s} b_i}\big[(l_1+1)\pi_1 + (l_2+1)\pi_2 + l_2\pi_1\big] + \left(1 - \frac{b_s}{\sum_{i\geq s} b_i}\right) w(0,1,s+1,2) - w(0,1,s,2).$$

Then $f_2(s)$ represents the difference in cost between starting the production of product 2 immediately and waiting for exactly one period and then start the production of product 2. The following lemma discusses the property of $f_2(s)$ and determines the optimal sequence if the production of one of the two products must be started.

Lemma 2 *If* $\pi_1 l_2 \leq \pi_2 l_1$, *then* (a) $f_2(s) \geq 0$ *for any* s; (b) $w(0,1,s,1) \geq w(0,1,s,2)$ *for any* s.

According to part (a) of the lemma, under the condition $\pi_1 l_2 \leq \pi_2 l_1$, if the supplier is to start the production of product 2 before the order for product 1 is confirmed, then it should do that immediately at state s; the cost is higher to wait for a while and then start producing product 2. Part (b) means that at any time state s, if the supplier is to start production, then the production of product 2 should be started first. The optimal policy follows directly from Lemma 2.

Theorem 2 *If* $\pi_1 l_2 \leq \pi_2 l_1$, *then for any* s, *the optimal decision is to start the production of product* 2; *that is*, $a^*(0,1,s) = 2$ *for all* s.

The optimal policy in this case is relatively simple because order 2 has a higher priority and it has been confirmed. Once the optimal policy for one of the two products has been decided, the remaining problem is a single-product problem. Specifically, if the order for product 1 is confirmed before $s + l_2$, the time when the capacity is released from producing product 2, then the production of product 1 needs to be started at $s + l_2$. If the order for product 1 is confirmed after $s + l_2$, then the problem facing the supplier is essentially the same as what is analyzed in Sect. 3 and the order confirmation time is X conditional on $X \geq l_2 + s$.

4.2 The order for product 1 has a higher priority

In this subsection, we consider the case when $\pi_1 l_2 > \pi_2 l_1$. We first introduce the following concept.

Definition 2 A discrete distribution $\{b_k\}_{k=1}^{\infty}$ is said to have an *increasing likelihood ratio* (ILR) if b_{k+i}/b_k is decreasing in k for all $i \geq 0$.

ILR is (slightly) stronger than IFR, but it still includes most of the common distributions (the IFR examples mentioned in Sect. 3 are all ILR). The following lemma further characterizes $f_2(s)$.

Lemma 3 *If* $\pi_1 l_2 > \pi_2 l_1$, $\pi_2 \geq h_1$, *and* X *has ILR*, *then* $f_2(s) \leq 0$ *if and only if* $s \geq \bar{s}_2$, *where*

$$\bar{s}_2 = \inf\{s : f_2(s) \leq 0\}.$$

The sign of $f_2(s)$ determines whether the supplier should immediately start the production of product 2 or delay for exactly one period and then start the production. As time progresses, since the order for product 1 is more and more likely to be confirmed due to IFR assumption, waiting is less likely to be the better choice.

Let

$$f_1(s) = \left(1 - \frac{b_s}{\sum_{i \geq s} b_i}\right)\pi_2 + \frac{b_s}{\sum_{i \geq s} b_i}\left[(l_1 + 1)\pi_1 + (l_2 + 1)\pi_2 + l_1\pi_2\right] + \left(1 - \frac{b_s}{\sum_{i \geq s} b_i}\right)w(0, 1, s + 1, 1) - w(0, 1, s, 1).$$

Similar to $f_2(s)$, $f_1(s)$ is the difference in cost between starting the production of product 1 immediately and waiting for exactly one period and then start the production of product 1. Lemma 4 provides useful properties of $f_1(s)$.

Lemma 4 *Assume that* $\pi_1 l_2 > \pi_2 l_1$ *and* X *has an IFR.*

(a) $f_1(s) \geq 0$ *if and only if* $s \geq \bar{s}_1$, *where*

$$\bar{s}_1 = \inf\{s : f_1(s) \geq 0\}.$$

(b) $\bar{s}_1 \leq s_1^*$, *where* s_1^* *is the smallest* s *such that*

$$\frac{\sum_{j \geq s + l_1 + 1} b_j}{\sum_{j \geq s} b_j} - \frac{\pi_1}{\pi_1 + h_1} \leq 0.$$

(c) *If, in addition,* $\pi_2 \geq h_1$, *then,* $f_1(s) \geq 0$ *for all* s.

The interpretation of part (a) is similar to that of Lemma 1. As time progresses, the order for product 1 is more likely to be confirmed and hence there is a point in time, $\bar{s}_1$, at which the supplier should stop waiting and start the production of product 1. What is different from Lemma 1 is that here the order for product 2 has already been confirmed, and hence delaying the production of product 1 will not only increase the expected penalty cost from product 1 but also delay the fulfillment of the order for product 2. As a result, the supplier should start the production of product 1 earlier than s_1^*, which is the stopping time when product 1 is the only product whose order needs to be filled. This is what we show in part (b). Finally, if $\pi_2 \geq h_1$, then starting the production of product 1 immediately is at least as good as delaying for exactly one period and then starting the production of product 1. This is because delaying for one period increases the penalty cost at least by π_2, but the saving on holding cost is no more than h_1. So $f_1(s) \geq 0$ for all s. By the above two lemmas, we can arrive at the optimal policy.

Theorem 3 *Assume that* $\pi_1 l_2 > \pi_2 l_1$, $\pi_2 \geq h_1$, *and* X *has ILR.*

(a) $u(0, 1, s) = \min\{w(0, 1, s, 1), w(0, 1, s, 2)\}$ *for all* s; *that is, for all* s, $a^*(0, 1, s) = 1$ *or* 2;
(b) *Let* $z_1^* = \inf\{s : w(0, 1, s, 1) - w(0, 1, s, 2) \geq 0\}$. *If* $s < z_1^*$, *then* $a^*(0, 1, s) = 2$; *if* $s \geq z_1^*$, *then* $a^*(0, 1, s) = 1$.

Theorem 3 provides a full characterization of the optimal policy for the case when product 1 has a higher priority. Again, once the production of one of the products has been started, the remaining problem is a one-product problem, for which the optimal policy is provided in Proposition 1. From a practical standpoint, the condition $\pi_2 \geq h_1$ is not at all restrictive. While the penalty costs may vary greatly among different products, the costs of production and logistics of different products that share the same capacity (different versions of 787 Dreamliners created by different customization, for examples) tend to be similar. This implies that the holding costs of products that share the same capacity are also similar. Furthermore, it is widely accepted in inventory research that, for the same product, delay penalty cost is much greater than holding cost. Take the newsvendor problem as an example. A service level of at least 90 % is common, which means that the penalty cost is at least 9 times greater than the holding cost. The following theorem further characterizes z_1^*.

Theorem 4

(a) *The threshold z_1^* is increasing in π_2 but decreasing in l_2;*
(b) *The threshold z_1^* is increasing in h_1 and l_1 but decreasing in π_1.*

As we can see from Theorem 4, while the ratio of the penalty cost to the lead time along cannot determines which product should be started first, it continues to play a crucial role in determining the optimal sequence. In particular, if π_i/l_i increases, then the set of state for which starting the production of product i is at least as good as starting that of the other product increases. The holding cost of product 1 also matters. As it increases, the risk of starting the production of product 1 too soon increases, which reduces the set of state for which the production of product 1 should be started first.

5 Models with multiple unconfirmed orders

For the general problem with at least two unconfirmed orders, we can compute the optimal cost and find the optimal policy by using induction. However, the optimal policy of the problem in general no longer possesses any structure even under ILR. To see this, let's consider the following example with three orders.

Example 1 The distributions of X_i, $i = 1, 2, 3$, are given by the following common formulas: $b_j = c\rho_1^j$ for $j = 1, 2, 3$, $b_j = c\rho_1^3\rho_2^{j-3}$ for $j = 4, 5, 6$, $b_j = c\rho_1^3\rho_2^3\rho_3^{j-6}$ for $j = 7, 8, 9$, $b_j = c\rho_1^3\rho_2^3\rho_3^3\rho_4^{j-9}$ for $j = 10, 11, 12$, and $b_j = 0$ for $j > 12$. The value of c is chosen such that the sum of b_i is one in each case. For product 1, $\pi_1 = 4$, $h_1 = 2$, $l_1 = 2$, $\rho_1 = \rho_2 = \rho_3 = \rho_4 = 1$. For product 2, $\pi_2 = 3.1$, $h_2 = 3$, $l_2 = 2$, $\rho_1 = \rho_2 = 2$, $\rho_3 = 0.5$, and $\rho_4 = 0.1$. For product 3, $\pi_3 = 4$, $h_3 = 3$, $l_3 = 1$, $\rho_1 = \rho_2 = \rho_3 = 3$, and $\rho_4 = 2$.

It is easy to verify that all the confirmation times in Example 1 have ILR. The optimal policies of different cases are reported in Table 1. As we can see, when both the orders for product 1 and product 2 are unconfirmed (the row for $a^*(0, 0, 2)$), neither the optimal time to start the production of the first product nor the optimal sequence has a threshold structure. The thresholds for product 1 and 2 are both at $s = 4$, which causes congestion. Therefore, it is optimal to start product 1, which has a lower holding cost, at $s = 2$, which is sooner than its threshold. For $s \leq 1$, the optimal action is "wait" because neither of the orders is very likely to be confirmed soon. For $s = 3$, the optimal action is also "wait" because at this

Table 1 The optimal policies in Example 1

s	1	2	3	4	5	6	7	8	9	10	11	12
$a^*(0, 2, 2)$	0	0	0	1	1	1	1	1	1	1	1	1
$a^*(2, 0, 2)$	0	0	0	2	2	2	2	2	2	2	2	2
$a^*(2, 2, 0)$	0	0	0	0	0	0	0	0	0	3	3	3
$a^*(0, 0, 2)$	0	1	0	2	2	2	2	2	1	1	1	1
$a^*(0, 2, 0)$	0	0	0	1	1	1	1	1	1	1	3	3
$a^*(2, 0, 0)$	0	0	0	2	2	2	2	2	2	3	3	3
$a^*(0, 0, 0)$	0	1	0	2	2	2	2	2	1	3	3	3

point, starting product 1 will delay the production of product 2, and stating product 2 will cause a high holding cost. The "wait-produce-wait" phenomenon also appears in the case when all products are unconfirmed (the row for $a^*(0, 0, 0)$).

The optimal policy in general is complex because of, first, the linkage across orders through capacity constraint, and second, the uncertain arrival times. What happens if we ignore the linkage and/or the uncertainty when setting the thresholds? In what follows, we consider two heuristics. In the threshold-based heuristic, we ignore the linkage across orders and compute their thresholds independent of others. If it is not feasible due to capacity constraint, priority will be given to the order with the highest π_i/l_i ratio. The difference in cost between the optimal policy and the heuristic measures the value of explicitly modeling the capacity constraint. In the mean-based heuristic, not only do we ignore the linkage, we also ignore the uncertainty of the order confirmation times. Every period, the supplier assigns a time slot to each order based on their expected confirmation times. The assignments are then updated over time when orders are confirmed. The difference in cost between the two heuristics measures the value of capturing uncertainty.

5.1 Threshold-based heuristic

To measure the cost of not considering the capacity linkage across products when setting thresholds, we construct the following threshold-based heuristic. For all unconfirmed orders, we first compute the thresholds as if there were infinite capacity and hence orders could be managed independently. That is, s_i^* is the smallest s such that

$$\frac{\sum_{j \geq s+l_i+1} b_j^i}{\sum_{j \geq s} b_j^i} - \frac{\pi_i}{\pi_i + h_i} \leq 0. \tag{5}$$

Let $k \in C$ such that

$$\frac{\pi_k}{l_k} \geq \frac{\pi_i}{l_i} \quad \text{for all } i \in C.$$

Let $\bar{C}_1$ be a subset of $\bar{C}$ such that

$$s \geq s_i^* \quad \text{for all } i \in \bar{C}_1.$$

If $C \cup \bar{C}_1$ is empty, then we wait; otherwise, we start the production of the product with the greatest π_i/l_i ratio in $C \cup \bar{C}_1$. As a special case, if $C \neq \emptyset$ and $\bar{C} = \emptyset$, according to the heuristic, we will start the production of product k, the product with the highest π_i/l_i ratio for $i \in C$. In this case, the heuristic is optimal. In Appendix we provide a theoretical bound for the cost of the heuristic (see Theorem 5).

Table 2 The costs as a function of a holding cost

h_1	0.2	0.6	1	1.4	1.8	2.2	2.6	3	3.4	3.8
Δ_c	2.8	3.5	4.4	5.2	9.4	5.4	4.1	4.5	2.6	2.8
Δ_{cu}	15.9	13.2	10.7	8.3	6.1	4.5	3.7	2.9	2.3	1.8

Table 3 The costs as a function of a delay penalty cost

π_1	1	2	3	4	5	6	7	8	9	10
Δ_c	0	0.1	0.3	9.6	5.2	2.9	1.9	0	0	0
Δ_{cu}	3.0	3.0	4.1	5.3	8.4	11.7	15.0	18.6	22.8	26.7

5.2 Mean-based heuristic

To further measure the cost of ignoring the demand arrival time uncertainty, we construct another heuristic, mean-based heuristic. In this heuristic, besides the capacity linkage, we also ignore the uncertainty of the order confirmation times. Every period, the supplier assigns a time slot to each order based on their expected confirmation times. The assignments are then updated over time when orders are confirmed.

At time s, let μ_i represent the conditional mean of X_i for all $i \in \bar{C}$ and let $\mu_i = 0$ for all $i \in C$. In the mean-based heuristic, the supplier ignores the capacity constraint as well as the uncertainty and plans as if order i would be confirmed in time μ_i in the future. As such, if there were infinite capacity, then the supplier would start the production of product i at time $\max\{\mu_i - l_i, s\}$, or when it is confirmed, whichever is sooner. Similar to the threshold-based heuristic, at a certain time s, if $s < \mu_i - l_i$ for multiple products, the priority will be given to the one with the highest π_i / l_i ratio.

5.3 Computational results

We let u_{mb} be the cost of implementing the mean-based heuristic and recall that u_{tb} is the cost of implementing the threshold-based heuristic. We define

$$\Delta_c = \frac{u_{tb} - u}{u} \times 100,$$

and

$$\Delta_{cu} = \frac{u_{mb} - u}{u} \times 100.$$

Here Δ_c measure the cost of ignoring the capacity linkage across products and Δ_{cu} measures the cost of ignoring both the capacity linkage and uncertainty.

For Tables 2 to 6, we use the parameters that we have used in Example 1 and assume that there are three unconfirmed orders. In each of the tables, we change one parameter related to order 1 while keeping all others unchanged. For the distributions that we use, all orders will be confirmed at or before $s = 12$. We then report Δ_c and Δ_{cu} in each case. In each case, we also calculate the analytical bound in Theorem 5 as a percentage over the minimal cost.

We can see from the tables that overall, even for three-product cases, the cost of ignoring the capacity linkage when setting the thresholds can be quite high, and the cost of ignoring both capacity linkage and demand uncertainty, in most cases, can be ever higher.

Table 4 The costs as a function of a lead time ($l_2 = 2, l_3 = 1$)

l_1	1	2	3	4	5	6	7	8	9	10
Δ_c	1.8	9.6	0.3	1.7	3.2	5.6	7.5	9.1	10.9	13.0
Δ_{cu}	6.5	5.3	7.0	6.2	6.6	8.1	9.5	11.3	13.2	15.0

Table 5 The costs as a function of a lead time ($l_2 = 8, l_3 = 1$)

l_1	1	2	3	4	5	6	7	8	9	10
Δ_c	21.2	10.8	7.0	6.7	6.8	7.1	6.8	6.0	5.0	8.0
Δ_{cu}	41.9	29.4	18.3	13.4	12.6	12.1	11.8	11.0	10	10.9

Table 6 The costs as a function of a lead time ($l_2 = 8, l_3 = 8$)

l_1	1	2	3	4	5	6	7	8	9	10
Δ_c	6.8	3.4	0	0	0	0	0	0	0	0
Δ_{cu}	1.5	0.8	0	0	0	0	0	0	0	0

In most, although not all, cases, ignoring the demand uncertainty is costly. The threshold-based heuristic is not always better than the mean-based heuristic (e.g., Table 4), which means that in our context, ignoring the demand uncertainty can sometimes be good. One possible outcome of ignoring capacity linkage when setting thresholds is congestion. To reduce congestion, one should shift the thresholds sooner so that they are spaced sufficiently apart. The thresholds set under the mean-based heuristics are sometimes smaller than those set under the threshold-based heuristic, which may reduce congestion.

The effects of lead time are summarized in Tables 4, 5, and 6. We first note from Table 4 that the costs of ignoring capacity linkage and demand uncertainty increase as l_1 increases. Is the value of l_1 or the differences between l_1 and the other two lead times that is driving the results? To see this, we increase l_2 to 8 in Table 5 and increase both l_2 and l_3 to 8 in Table 6. Interestingly, in Tables 5 and 6, the Δ_c and Δ_{cu} actually become smaller as l_1 increases, a stark contrast to what happens in Table 4. In fact, as we can see in Table 6, when $l_2 = l_3 = 8$ and l_1 is not too far from 8, both Δ_c and Δ_{cu} are zero. This suggests that it is the differences among the lead times that matter (more).

To understand the effects of lead time, we need to first understand the effects of lead time on the optimal ranking of orders. If we look at the π_i/l_i ratio, then an order with a shorter lead time has a higher π_i/l_i ratio and hence should be processed sooner. This is because processing an order with a short lead time will not cause much delay for other orders. If we use the thresholds defined in (5) (i.e., s_i^*) to rank orders, then the opposite is true; the threshold is lower when the lead time is longer, therefore the order with a longer lead time should be processed sooner. This is because the longer the lead time, the earlier one should start production preemptively to avoid delay. The optimal response to an increase in lead time is nontrivial.

In the threshold-based heuristic, we first compute s_i^*. For the orders whose s_i^* is smaller than the current time s, we rank them together with the already confirmed orders by their π_i/l_i ratio. For the orders whose s_i^* is greater than the current time s, however, we rank them by their s_i^* values. We do the same in the mean heuristic except that we compute the thresholds differently. This treatment, while natural and intuitive, is suboptimal in general.

Table 7 The costs as a function of the number of products

Number of products	1	2	3	4	5	6
Δ_c	0	7	6.3	33	23	23.6
Δ_{cu}	106.8	83.6	100.2	157.7	116	115.8

Table 8 The costs change when the products become more different

Parameter range	Original	$\pi[7, 11]$	$l[1, 5]$	$h[0, 4]$
Δ_c	20.9	22.8	43.8	24.3
Δ_{cu}	83.9	88.2	130.5	81.1

When the lead times of different orders are significantly different, it leads to significantly different ranking of orders from the optimal solution.

Finally, in order to see the impact on the costs when the number of products increases, we change the number of products from 1 to 6. We add product to the system one at a time. To get a robust result, we randomly generate 10,000 examples. We generate π_i and h_i by using continuous uniform distributions over [8, 10] and [1, 3], respectively. The lead times are generated from a discrete uniform distribution over [2, 4]. As we can see from Table 7, as the number of products increases, both Δ_c and Δ_{cu} increase, although Δ_c is more sensitive to the number of products. In Table 8, we keep the number of unconfirmed product to be six. We increase the range of the uniform distributions used for generating the penalty cost, holding cost, and lead time one at a time. For example, we increasing the range from [8, 10] to [7, 11] for the penalty cost whiling keeping other parameters unchanged. By doing that, we make the six products more different from each other. We again randomly generate 8,000 examples. As we can see from the table, Δ_c and Δ_{cu} are particularly sensitive to the change in lead time and they both become greater when the lead times are more different. From the our numerical studies, we can conclude that the cost of ignoring the capacity linkage across product is high when there are many products and their lead times are very different from one another. It is always a bad idea to ignore demand uncertainty.

6 Discussion and concluding remarks

Motivated by applications in capital goods industries, we have constructed a model by discrete time Markov decision process to capture the essence of the order fulfillment processes when there are multiple different orders. We have provided a procedure to compute the optimal policy recursively. However, the computational complexity increases exponentially when the number of products or lead times increase. For problems with very long lead times and many products, one may consider approximate dynamic programming or simpler heuristic policies. In addition, there are a few practical issues that we have not in the model directly considered, but one may (sometimes) encounter in capital goods industries. Some of these issues can be addressed by slightly calibrating the basic model, while others require new models and are perhaps well beyond the scope of the current paper.

In many cases, customers may not expect to have the product as soon as they place their orders. Instead, they may allow the supplier some time, which is sometimes called order lead time, to produce and deliver the order. The supplier won't incur a delay penalty cost until the customers expect to have the product. In this case, instead of the order confirmation time, we need to reinterpret X_i as the time when the customers expect to have the product.

Note that in practice the order lead time is often shorter than the time required to prepare and deliver the order and the difference between the two is known as lead time gap (Christopher 2000).

We have assumed that the order lead time is determined by the customers. When it is determined by the supplier, the market, or regulators, it is better known as due date. If the due date is exogenously given by the market as an industry standard, or constrained by technological limit, or by the regulators, then X_i should be interpreted as the time when the order is due (Duenyas and Hopp 1995) and our previous analysis continues to apply. However, if the due date is determined by the supplier, a new model is needed since the action space is now different. It is a promising extension to combine due date setting and timing of order fulfillment.

Other than order lead time and due date, additional issues include the possibility of order cancellation, lead time uncertainty, common modules among orders, etc. Incorporating some or all of these into the model will make the model formulation quite tedious. However, we believe that these issues will not alter the qualitative results in a significant way. Neither do these issues present new computational challenges. Or the demand arrival time depends on additional factors, such as stock market, weather, etc., that are observable to the supplier. The supplier will update the distribution based on the latest information. In this case, the addition of state variable will make the computation of the optimal policy more challenging. Despite all these issues in practice, we believe that the results in this study provide insights into the problem and are useful for firms in such industries to understand the trade-offs they face. Our formulation is amenable for computation of optimal policies. We highlight that it is important to incorporate uncertainty in planning and it is also important to coordinate the fulfillment processes across products, especially when there are many products.

Acknowledgements The authors are grateful to Xiangtong Qi and Albert Ha for helpful discussions on this topic. The authors also wish to thank the Guest Editor and two anonymous referees for their many critical but constructive comments and suggestions, which have improved this paper. The research was in part supported by HKUST Business School (SBI09/10.BM16), National Natural Science Foundation of China under Grant No.71131003,71201061, the Fundamental Research Funds for the Central Universities (2012ZM0032).

Appendix

Proof of Lemma 1 Note first that

$$\begin{aligned} f(s) &= \frac{b_s}{\sum_{i\geq s} b_i}(l+1)\pi + \frac{\sum_{i\geq s+1} b_i}{\sum_{i\geq s} b_i}\left[\pi\sum_{j=0}^{l-1}(l-j)\frac{b_{s+1+j}}{\sum_{i\geq s+1} b_i} + h\sum_{j=l+1}^{\infty}(j-l)\frac{b_{s+1+j}}{\sum_{i\geq s+1} b_i}\right] \\ &\quad -\pi\sum_{j=0}^{l-1}(l-j)\frac{b_{s+j}}{\sum_{i\geq s} b_i} - h\sum_{j=l+1}^{\infty}(j-l)\frac{b_{s+j}}{\sum_{i\geq s} b_i} \\ &= \frac{b_s}{\sum_{i\geq s} b_i}(l+1)\pi + \frac{1}{\sum_{i\geq s} b_i}\left[\pi\sum_{j=0}^{l-1}(l-j)(b_{s+1+j}-b_{s+j})\right. \\ &\quad \left. + h\sum_{j=l+1}^{\infty}(j-l)(b_{s+1+j}-b_{s+j})\right] \end{aligned}$$

$$= \frac{1}{\sum_{i\geq s} b_i}\left[b_s(l+1)\pi + \pi\sum_{j=0}^{l-1}(l-j)(b_{s+1+j}-b_{s+j}) \right.$$
$$\left. + h\sum_{j=l+1}^{\infty}(j-l)(b_{s+1+j}-b_{s+j})\right].$$

By using the following identities:

$$\sum_{j=0}^{l-1}(l-j)b_{s+1+j} = \sum_{j=0}^{l-1}\sum_{i=s+1}^{j+s+1} b_i; \qquad \sum_{j=0}^{l-1}(l-j)b_{s+j} = \sum_{j=0}^{l-1}\sum_{i=s}^{s+j} b_i; \tag{6}$$

$$\sum_{j=l+1}^{\infty}(j-l)b_{s+1+j} = \sum_{j=l+1}^{\infty}\sum_{i=j+1}^{\infty} b_{s+i}; \qquad \sum_{j=l+1}^{\infty}(j-l)b_{s+j} = \sum_{j=l+1}^{\infty}\sum_{i=j}^{\infty} b_{s+i}, \tag{7}$$

we have

$$\begin{aligned} f(s) &= \frac{1}{\sum_{i\geq s} b_i}\left[\pi\sum_{j=s}^{l+s} b_j - h\sum_{j=l+1+s}^{\infty} b_j\right] \\ &= \frac{1}{\sum_{i\geq s} b_i}\left[\pi\sum_{j=s}^{\infty} b_j - (\pi+h)\sum_{j=s+l+1}^{\infty} b_j\right] \\ &= \pi - (\pi+h)\frac{\sum_{j\geq s+l+1} b_j}{\sum_{j\geq s} b_j}. \end{aligned}$$

Due to IFR assumption, the expression inside the square bracket is increasing in s. Therefore, $f(s) \geq 0$ if and only if $s \geq s^*$. □

Proof of Proposition 1 For any $s < s^*$, starting production can't be optimal because waiting for one period and then starting production costs strictly less than immediately starting production (Lemma 1). We show the optimal policy for $s \geq s^*$ by contradiction. Suppose there is an $s' \geq s^*$ such that $a^*(s') = 0$.

Suppose $a^*(s) = 0$ for all $s \geq s'$. Let

$$c_s = \frac{b_s}{\sum_{i\geq s} b_i}(l+1)\pi.$$

By Lemma 1, for any $s \geq s'$ we have

$$\begin{aligned} w(s,1) &\leq c_s + \left(1 - \frac{b_s}{\sum_{j\geq s} b_j}\right)w(s+1,1) \\ &\leq c_s + \left(1 - \frac{b_s}{\sum_{j\geq s} b_j}\right)\left[c_{s+1} + \left(1 - \frac{b_{s+1}}{\sum_{j\geq s+1} b_j}\right)w(s+2,1)\right] \\ &\leq \cdots \\ &\leq c_s + \left(1 - \frac{b_s}{\sum_{j\geq s} b_j}\right)c_{s+1} + \left(1 - \frac{b_s}{\sum_{j\geq s} b_j}\right)\left(1 - \frac{b_{s+1}}{\sum_{j\geq s+1} b_j}\right)c_{s+2} + \cdots. \end{aligned}$$

Therefore immediately starting production is better than waiting until the order is confirmed and hence there must be a finite $s > s'$ such that $a^*(s) = 1$.

We let

$$s'' = \min\{s : a^*(s) = 1 \text{ and } s > s'\}.$$

By the definitions of s' and s'', we have $a^*(s''-1) = 0$ and $a^*(s'') = 1$. This is a contradiction because we know from Lemma 1 that for any $s \geq s^*$, immediately starting production costs less than waiting for one period and then starting production. In conclusion, such s' does not exist and hence $a^*(s) = 1$ for all $s \geq s^*$. □

Proof of Lemma 2 Let

$$c_s = \left(1 - \frac{b_s}{\sum_{i\geq s} b_i}\right)\pi_2 + \frac{b_s}{\sum_{i\geq s} b_i}\big[(l_1+1)\pi_1 + (l_2+1)\pi_2 + l_2\pi_1\big].$$

(a) Note first that

$$\begin{aligned}
&c_s + \left(1 - \frac{b_s}{\sum_{j\geq s} b_j}\right) w(0,1,s+1,2)\\
&= \left(1 - \frac{b_s}{\sum_{j\geq s} b_j}\right)\pi_2 + \frac{b_s}{\sum_{j\geq s} b_j}(l_1\pi_1 + l_2\pi_2 + l_2\pi_1 + \pi_1 + \pi_2)\\
&\quad + \left(1 - \frac{b_s}{\sum_{j\geq s} b_j}\right)\Bigg[l_2\pi_2 + \frac{\sum_{j=0}^{l_2-1}(l_1+l_2-j)b_{s+j+1}}{\sum_{i\geq s+1} b_i}\pi_1\\
&\quad + \frac{\sum_{j=l_2}^{\infty} b_{s+j+1}}{\sum_{i\geq s+1} b_i} u(0,2,s+1+l_2)\Bigg]\\
&= \left(1 - \frac{b_s}{\sum_{j\geq s} b_j}\right)\pi_2 + l_2\pi_2 + \frac{b_s}{\sum_{j\geq s} b_j}(l_1\pi_1 + l_2\pi_1 + \pi_1 + \pi_2)\\
&\quad + \frac{1}{\sum_{j\geq s} b_j}\left[\pi_1 \sum_{j=0}^{l_2-1}(l_1+l_2-j)b_{s+j+1} + \sum_{j=l_1}^{\infty} b_{s+j+1} u(0,2,s+1+l_2)\right].
\end{aligned}$$

Also, by the definition of $u(0,2,s)$,

$$\begin{aligned}
w(0,1,s,2) &\leq l_2\pi_2 + \frac{\sum_{j=0}^{l_2-1}(l_1+l_2-j)b_{s+j}}{\sum_{i\geq s} b_i}\pi_1\\
&\quad + \frac{\sum_{j=l_2}^{\infty} b_{s+j}}{\sum_{i\geq s} b_i}\left[\frac{b_{s+l_2}}{\sum_{i\geq s+l_2} b_i} l_1\pi_1 + \left(1 - \frac{b_{s+l_2}}{\sum_{j\geq s+l_2} b_j}\right) u(0,2,s+1+l_2)\right].
\end{aligned}$$

Therefore,

$$\begin{aligned}
f_2(s) &\geq \left(1 - \frac{b_s}{\sum_{j\geq s} b_j}\right)\pi_2 + \frac{b_s}{\sum_{j\geq s} b_j}(l_1\pi_1 + l_2\pi_1 + \pi_1 + \pi_2)\\
&\quad + \pi_1 \frac{\sum_{j=0}^{l_2-1}(l_1+l_2-j)b_{s+j+1}}{\sum_{j\geq s} b_j}\\
&\quad - \frac{\sum_{j=0}^{l_2-1}(l_1+l_2-j)b_{s+j}}{\sum_{i\geq s} b_i}\pi_1 - l_1\pi_1 \frac{b_{s+l_2}}{\sum_{j\geq s} b_j}\\
&= \pi_1 \frac{\sum_{i=s}^{s+l_2} b_i}{\sum_{i\geq s} b_i} + \pi_2\\
&\geq 0.
\end{aligned}$$

The result hence follows.

(b) We show that under the condition, $w(0,1,s,1) \geq w(0,1,s,2)$ for all s. Let X_s be the time when the order for product 1 is confirmed conditioning on that it is greater than s; that is, $Prob(X_s = i) = Prob(X = i | X \geq s)$ for all $i \geq s$. Then,

$$\begin{aligned} w(0,1,s,2) &\leq l_2\pi_2 + \mathsf{E}\pi_1(l_1+l_2-X_s)^+ + \mathsf{E}h_1(X_s-l_1-l_2)^+ \\ &\leq l_2\pi_2 + \mathsf{E}\pi_1(l_1+l_2-X_s)^+ + \mathsf{E}h_1(X_s-l_1)^+ \\ &\leq l_2\pi_2 + l_2\pi_1 + \mathsf{E}\pi_1(l_1-X_s)^+ + \mathsf{E}h_1(X_s-l_1)^+ \\ &\leq l_2\pi_2 + l_1\pi_2 + \mathsf{E}\pi_1(l_1-X_s)^+ + \mathsf{E}h_1(X_s-l_1)^+ \\ &= w(0,1,s,1). \end{aligned}$$

The first inequality is because it is feasible but not necessarily optimal to start the production of product 1 at time l_2. The second and the third inequalities are obvious. The fourth follows from the condition $\pi_1 l_2 \leq \pi_2 l_1$. □

Proof of Theorem 2 Because $w(0,1,s,1) \geq w(0,1,s,2)$, we can simplify (4) to

$$u(0,1,s) = \min\{w(0,1,s,0), w(0,1,s,2)\}.$$

The rest of the proof is similar to the proof of Proposition 1 and hence omitted. □

Proof of Lemma 3 Let s_1^* be the threshold for product 1 defined in Sect. 3. For $s \leq s_1^* - l_2 - 2$, following from similar arguments used in the proof of Lemma 2, we have

$$f_2(s) = \frac{b_s}{\sum_{i\geq s} b_i}\left[\pi_1 \sum_{i=0}^{l_2} \frac{b_{s+i}}{b_s} + \pi_2 \sum_{i\geq 0} \frac{b_{s+i}}{b_s} + \pi_2 l_1 - \pi_1 l_2\right].$$

Because of the ILR assumption, the expression inside the square bracket is decreasing in s.

For $s \geq s^* - l_2$, then,

$$u(0,2,s+l_2+1) = \pi_1 \sum_{j=0}^{l_1-1}(l_1-j)\frac{b_{s+1+l_2+j}}{\sum_{i\geq s+1+l_2} b_i} + h_1 \sum_{j=l_1+1}^{\infty}(j-l_1)\frac{b_{s+1+l_2+j}}{\sum_{i\geq s+1+l_2} b_i},$$

$$u(0,2,s+l_2) = \pi_1 \sum_{j=0}^{l_1-1}(l_1-j)\frac{b_{s+l_2+j}}{\sum_{i\geq s+l_2} b_i} + h_1 \sum_{j=l_1+1}^{\infty}(j-l_1)\frac{b_{s+l_2+j}}{\sum_{i\geq s+l_2} b_i}.$$

By using identities (6) and (7), we have

$$\begin{aligned} f_2(s) = \frac{b_s}{\sum_{i\geq s} b_i}\Bigg[&(\pi_1+h_1)\sum_{i=0}^{l_1+l_2-1}\frac{b_{s+1+i}}{b_s} + (\pi_2-h_1)\sum_{i=1}^{\infty}\frac{b_{s+i}}{b_s} \\ &+ (\pi_2 l_1 - \pi_1 l_2) + (\pi_1+\pi_2)\Bigg]. \end{aligned}$$

Under the conditions that X has ILR and $\pi_2 \geq h_1$, $f_2(s)$ crosses zero at most once, and, if it does, it is from above.

Finally, we shall show

$$\lim_{s\to(s_1^*-l_2-1)^-} f_2(s) \geq f_2(s_1^*-l_2-1) \geq \lim_{s\to(s_1^*-l_2-1)^+} f_2(s).$$

The first inequality is true because based on the definition of s_1^*, when $s = s_1^*$, $a^*(0, 2, s) = 1$, not 0. The second inequality is because when $s = s_1^* - 1$, $a^*(0, 2, s) = 0$, not 1. In summary, $f_2(s)$ may cross zero at most once, and, if it does, it is from above. We thus can conclude that $f_2(s) \le 0$ if and only if $s \ge \bar{s}_2$. □

Proof of Lemma 4 We show that under the conditions, $f_1(s)$ crosses zero at most once when s traverses from zero to infinity, and, if it does, it is from below.

$$
\begin{aligned}
f_1(s) = {}& \left(1 - \frac{b_s}{\sum_{i\ge s} b_i}\right)\pi_2 + \frac{b_s}{\sum_{i\ge s} b_i}\big[(l_1+1)\pi_1 + (l_2+1)\pi_2 + l_1\pi_2\big] \\
& + \frac{\sum_{i\ge s+1} b_i}{\sum_{i\ge s} b_i}\left[\pi_1 \sum_{j=0}^{l_1-1}(l_1 - j)\frac{b_{s+1+j}}{\sum_{i\ge s+1} b_i}\right. \\
& \left. + h_1 \sum_{j=l_1+1}^{\infty}(j - l_1)\frac{b_{s+1+j}}{\sum_{i\ge s+1} b_i} + (l_1+l_2)\pi_2\right] \\
& - \pi_1 \sum_{j=0}^{l_1-1}(l_1 - j)\frac{b_{s+j}}{\sum_{i\ge s} b_i} - h_1 \sum_{j=l_1+1}^{\infty}(j - l_1)\frac{b_{s+j}}{\sum_{i\ge s} b_i} - (l_1+l_2)\pi_2 \\
= {}& \left(1 - \frac{b_s}{\sum_{i\ge s} b_i}\right)\pi_2 + (\pi_1+\pi_2)\frac{b_s}{\sum_{i\ge s} b_i} \\
& + \frac{1}{\sum_{i\ge s} b_i}\left[b_s l_1 \pi_1 + \pi_1 \sum_{j=0}^{l_1-1}(l_1 - j)(b_{s+1+j} - b_{s+j})\right. \\
& \left. + h_1 \sum_{j=l_1+1}^{\infty}(j - l_1)(b_{s+1+j} - b_{s+j})\right].
\end{aligned}
$$

By using identities (6) and (7), we have

$$
f_1(s) = \pi_1 + \pi_2 - (\pi_1 + h_1)\frac{\sum_{j\ge s+1+l_1} b_j}{\sum_{j\ge s} b_j}.
$$

Part (c) is immediate. Part (a) follows because $\sum_{j\ge s+1+l_1} b_j / \sum_{j\ge s} b_j$ is decreasing in s if X has an IFR. Part (b) follows because $f_1(s) \ge f(s)$ when π and h in $f(s)$ are replaced by π_1 and h_1, respectively. □

Proof of Theorem 3 (a) The proof is by contradiction. Suppose there is a $\hat{s}$ such that $a^*(0, 1, \hat{s}) = 0$.

We first show that $a^*(0, 1, \hat{s}) = 0$ implies $a^*(0, 1, s) = 0$ for all $s \ge \hat{s}$. This can be showed by contradiction. If there is an $s \ge \hat{s}$ such that $a^*(0, 1, s) = 0$ and $a^*(0, 1, s+1) = 2$. This implies $f_2(s) \le 0$. According to Lemma 3, we have $f_2(s+1) \le 0$, which means that starting the production of product 2 at $s+1$ would be more costly than delaying the production until one period later. So $a^*(0, 1, s+1) = 0$, which is a contradiction. Also, because of part (c) of Lemma 4, if $a^*(0, 1, s) = 0$, then $a^*(0, 1, s+1) \ne 1$. Therefore, if $a^*(0, 1, \hat{s}) = 0$, then $a^*(0, 1, s) = 0$ for all $s \ge \hat{s}$.

We next show that such $\hat{s}$ does not exist. Suppose it did, then

$$
\begin{aligned}
u(0, 1, \hat{s}) &= l_1\pi_1 + (l_1+l_2)\pi_2 + \pi_2 \mathsf{E}X_{\hat{s}} \\
&\ge l_1\pi_1 + (l_1+l_2)\pi_2 + h_1 \mathsf{E}X_{\hat{s}}
\end{aligned}
$$

$$\geq \pi_1 \mathsf{E}(l_1 - X_{\hat{s}})^+ + h_1 \mathsf{E}(X_{\hat{s}} - l_1)^+ + (l_1 + l_2)\pi_2$$
$$= w(0, 1, \hat{s}, 1),$$

which is a contradiction. Therefore, such $\hat{s}$ does not exist and "wait" can never be the optimal action.

(b) Let $\delta = (w(0, 1, s, 2) - w(0, 1, s, 1)) \sum_{i \geq s} b_i$. To show the result, it suffices to show that δ as a function of s crosses zero at most once when s traverses from zero to infinity, and if it does, it does so from below. For $s < s_1^* - l_2$,

$$\begin{aligned}
\delta &= \pi_1 \sum_{j=0}^{l_2-1} (l_1 + l_2 - j) b_{s+j} + (l_1 + 1)\pi_1 \sum_{j=s+l_2}^{s_1^*-1} b_j + \pi_1 \sum_{j=0}^{l_1-1} (l_1 - j)(b_{s_1^*+j} - b_{s+j}) \\
&\quad + h_1 \sum_{j=l_1+1}^{\infty} (j - l_1)(b_{s_1^*+j} - b_{s+j}) - l_1 \pi_2 \sum_{i=s}^{\infty} b_i \\
&= (\pi_1 l_2 - l_1 \pi_2) \sum_{i=s}^{\infty} b_i - \pi_1 \sum_{j=0}^{l_2-1} \sum_{i=s+j+1}^{\infty} b_i + \pi_1 l_1 \sum_{i=s}^{s_1^*-1} b_i + \pi_1 \sum_{j=s+l_2}^{s_1^*-1} b_j \\
&\quad + \pi_1 \sum_{j=0}^{l_1-1} \sum_{i=0}^{j} (b_{s_1^*+i} - b_{s+i}) + h_1 \sum_{j=l_1+1}^{\infty} \sum_{i=j}^{\infty} (b_{s_1^*+i} - b_{s+i}) \\
&= (\pi_1 l_2 - l_1 \pi_2) \sum_{i=s}^{\infty} b_i - (\pi_1 + h_1) \sum_{j=l_1+s+1}^{l_1+s_1^*} \sum_{i=j}^{\infty} b_i + \pi_1 \sum_{j=s+l_2+1}^{s_1^*} \sum_{i=j}^{\infty} b_i + \pi_1 \sum_{j=s+l_2}^{s_1^*-1} b_j.
\end{aligned}$$

For $s \geq s_1^* - l_2$,

$$\begin{aligned}
\delta &= \pi_1 \sum_{j=0}^{l_2-1} (l_1 + l_2 - j) b_{s+j} - l_1 \pi_2 \sum_{i=s}^{\infty} b_i \\
&\quad + \pi_1 \sum_{j=0}^{l_1-1} (l_1 - j)(b_{s+l_2+j} - b_{s+j}) + h_1 \sum_{j=l_1+1}^{\infty} (j - l_1)(b_{s+l_2+j} - b_{s+j}) \\
&= \pi_1 \sum_{j=0}^{l_2-1} \sum_{i=0}^{j} b_{s+i} + \pi_1 l_1 \sum_{i=s}^{l_2+s-1} b_i - l_1 \pi_2 \sum_{i=s}^{\infty} b_i \\
&\quad + \pi_1 \sum_{j=0}^{l_1-1} \sum_{i=0}^{j} (b_{s+l_2+i} - b_{s+i}) + h_1 \sum_{j=l_1+1}^{\infty} \sum_{i=j}^{\infty} (b_{s+l_2+i} - b_{s+i}) \\
&= \sum_{i=s}^{\infty} b_i \left[\pi_1 l_2 - \pi_2 l_1 - (\pi_1 + h_1) \sum_{j=l_1+1}^{l_1+l_2} \frac{\sum_{i \geq j+s} b_i}{\sum_{i \geq s} b_i} \right],
\end{aligned}$$

which, under ILR assumption, can cross zero at most once and if it does, it does so from below. Clearly, if $s_1^* \leq l_2$, then $w(0, 1, s, 2) - w(0, 1, s, 2)$ is increasing in $s \in [0, \infty)$ and the result is immediate. So we just need to show the result when $s_1^* > l_2$. Let

$$g_1(s) = (\pi_1 l_2 - l_1 \pi_2) \sum_{i=s}^{\infty} b_i - (\pi_1 + h_1) \sum_{j=l_1+s+1}^{l_1+s_1^*} \sum_{i=j}^{\infty} b_i + \pi_1 \sum_{j=s+l_2+1}^{s_1^*} \sum_{i=j}^{\infty} b_i + \pi_1 \sum_{j=s+l_2}^{s_1^*-1} b_j,$$

and

$$g_2(s) = (\pi_1 l_2 - \pi_2 l_1) \sum_{i=s}^{\infty} b_i - (\pi_1 + h_1) \sum_{j=l_1+s+1}^{l_1+l_2+s} \sum_{i=j}^{\infty} b_i.$$

Both g_1 and g_2 are defined in $[0, \infty)$ and $\delta = g_1(s)$ for $s < s_1^* - l_2$ and $\delta = g_2(s)$ for $s \geq s_1^* - l_2$. We first show that $\delta(s)$ is increasing initially and then decreasing in $s \in (0, s_1^* - l_2]$ (i.e., quasi-concave if s were continuous).

$$g_1(s) - g_1(s-1) = -(\pi_1 l_2 - \pi_2 l_1) b_{s-1} + (\pi_1 + h_1) \sum_{i=l_1+s}^{\infty} b_i - \pi_1 \sum_{i=s+l_2-1}^{\infty} b_i.$$

Consider two cases. First, if $l_2 - 1 > l_1$, then,

$$g_1(s) - g_1(s-1) = b_{s-1} \left[-(\pi_1 l_2 - \pi_2 l_1) + \pi_1 \sum_{i=l_1+1}^{l_2-1} \frac{b_{i+s-1}}{b_{s-1}} + h_1 \sum_{i=l_1+1}^{\infty} \frac{b_{i+s-1}}{b_{s-1}} \right].$$

Since the expression inside the square bracket is decreasing in s under ILR assumption, g_1 is quasi-concave. Second, if $l_2 - 1 \leq l_1$, then,

$$\begin{aligned} & g_1(s) - g_1(s-1) \\ & = b_{s-1} \left\{ -(\pi_1 l_2 - \pi_2 l_1) + (\pi_1 + h_1) \sum_{i=l_2}^{\infty} \frac{b_{i+s-1}}{b_{s-1}} \left[\frac{\sum_{i \geq s+l_1} b_i}{\sum_{i \geq s+l_2-1} b_i} - \frac{\pi_1}{\pi_1 + h_1} \right] \right\}. \end{aligned} \tag{8}$$

Because of IFR assumption, we know

$$\begin{aligned} \frac{\sum_{i \geq s+l_1} b_i}{\sum_{i \geq s+l_2-1} b_i} & \geq \frac{\sum_{i \geq s_1^*-l_2+l_1} b_i}{\sum_{i \geq s_1^*-1} b_i} \\ & \geq \frac{\sum_{i \geq s_1^*+l_1} b_i}{\sum_{i \geq s_1^*-1} b_i} \\ & \geq \frac{\pi_1}{\pi_1 + h_1}, \end{aligned}$$

where the last inequality is because of the definition of s_1^*. Therefore, the expression within $\{\}$ in (8) is decreasing in s under ILR assumption and hence g_1 is also quasi-concave in this case. Finally,

$$\begin{aligned} \delta\big(s_1^* - l_2\big) - \delta\big(s_1^* - l_2 - 1\big) &= g_2\big(s_1^* - l_2\big) - g_1\big(s_1^* - l_2 - 1\big) \\ &= -(\pi_1 l_2 - \pi_2 l_1) b_{s_1^*-l_2-1} + (\pi_1 + h_1) \sum_{i=l_1+s_1^*-l_2}^{\infty} b_i - \pi_1 \sum_{i=s_1^*-1}^{\infty} b_i. \end{aligned}$$

From the previous analysis, we know that if $g_1(s_1^* - l_2 - 1) - g_1(s_1^* - l_2 - 2) \leq 0$, then $\delta(s_1^* - l_2) - \delta(s_1^* - l_2 - 1) \leq 0$. Therefore, we can conclude that δ is quasi-concave in $(0, s_1^* - l_2]$.

Note then that $g_1(s) \leq g_2(s)$ for $s < s_1^* - l_2$ because of the threshold structure of the optimal policy for product 1. Consider the following two cases. If $g_2(s_1^* - l_2) \leq 0$, then $g_1(s) \leq g_2(s) \leq 0$ for all $s < s_1^* - l_2$. So in this case, if δ crosses zero at all, it must do so at some $s \in [s_1^* - l_2, \infty)$ and from below. Suppose $g_2(s_1^* - l_2) > 0$. Then because $\delta(s)$ is quasi-concave in $(0, s_1^* - l_2]$, if δ crosses zero, then it must do so at some $s \in (0, s_1^* - l_2 - 1)$ from below. Let $z_1^* = \inf\{s : w(0,1,s,2) - w(0,1,s,1) \ge 0\}$. Then, $a^*(0,1,s) = 2$ for $s < z_1^*$ and $a^*(0,1,s) = 1$ otherwise. □

Proof of Theorem 4 (a) Examining the behavior of z_1^* is equivalent to examining the behavior of δ; that is, to show the results, it suffices to show that δ is decreasing in π_2 but increasing in l_2. This is obviously true for π_2. For l_2, denote δ and g_i, where $i = 1, 2$, as $\delta(l_2)$ and $g_i(s, l_2)$, respectively, to make their dependence on l_2 explicit. Suppose $s_1^* > s$. Then, for $l_2 < s_1^* - s$,

$$\delta(l_2) = g_1(s, l_2),$$

which is obviously increasing in l_2. For $l_2 \ge s_1^* - s$, we consider $l_2^2 > l_2^1 \ge s_1^* - s$. Then,

$$\begin{aligned}
\delta\big(l_2^2\big) - \delta\big(l_2^1\big) &= g_2\big(s, l_2^2\big) - g_2\big(s, l_2^1\big) \\
&= \big(l_2^2 - l_2^1\big)\pi_1 \sum_{i=s}^{\infty} b_i - (\pi_1 + h_1) \sum_{j=l_1+l_2^1+1}^{l_1+l_2^2} \sum_{i=j+s}^{\infty} b_i \\
&\ge \big(l_2^2 - l_2^1\big)\pi_1 \sum_{i=s}^{\infty} b_i - (\pi_1 + h_1)\big(l_2^2 - l_2^1\big) \sum_{i=s+l_1+l_2^1+1}^{\infty} b_i \\
&= \big(l_2^2 - l_2^1\big)(\pi_1 + h_1) \sum_{i=s}^{\infty} b_i \left(\frac{\pi_1}{\pi_1 + h_1} - \frac{\sum_{i=s+l_1+l_2^1+1}^{\infty} b_i}{\sum_{i=s}^{\infty} b_i} \right).
\end{aligned}$$

From the definition of s_1^* and the fact that $s_1^* \le s + l_2^1$, we have

$$\frac{\pi_1}{\pi_1 + h_1} \ge \frac{\sum_{i=s_1^*+l_1+1}^{\infty} b_i}{\sum_{i=s_1^*}^{\infty} b_i} \ge \frac{\sum_{i=s+l_2^1+l_1+1}^{\infty} b_i}{\sum_{i=s+l_2^1}^{\infty} b_i} \ge \frac{\sum_{i=s+l_2^1+l_1+1}^{\infty} b_i}{\sum_{i=s}^{\infty} b_i}.$$

Therefore $\delta(l_2^2) - \delta(l_2^1) \ge 0$ and δ is also increasing in this region. Finally, since

$$\begin{aligned}
\delta\big(s_1^* - s\big) - \delta\big(s_1^* - s - 1\big) &= g_2\big(s, s_1^* - s\big) - g_1\big(s, s_1^* - s - 1\big) \\
&= \pi_1 \left(\sum_{i=s}^{\infty} b_i - \sum_{i=s_1^*-1}^{\infty} b_i \right),
\end{aligned}$$

which is positive for $s_1^* > s$. When $s_1^* \le s$, $\delta(l_2) = g_2(s, l_2)$ for all l_2 and hence the result follows easily.

(b) We only provide the proof for the statement about h_1; results about p_1 and l_1 can be shown analogously. To show that z_1^* is increasing in h_1, it suffices to show that δ is decreasing in h_1. Similarly, we denote δ, s_1^*, and g_i, $i = 1, 2$, as $\delta(h_1)$, $s_1^*(h_1)$, and $g_i(s, h_1)$, respectively, for ease of presentation. Let

$$s^{-1} = \sup\left\{ h_1 : \frac{\sum_{j \ge s+l_1+1} b_j}{\sum_{j \ge s} b_j} - \frac{\pi_1}{\pi_1 + h_1} = 0 \right\};$$

that is, s^{-1} is the inverse function of s_1^* and it is increasing. For $h_1 \le s^{-1}(s + l_2)$, $\delta(h_1) = g_2(s, h_1)$, which is obviously decreasing in h_1.

For $h_1 > s^{-1}(s + l_2)$, $\delta(h_1) = g_1(s, h_1)$. Let $h_1^2 > h_1^1 > s^{-1}(s + l_2)$. Then, $s_1^*(h_1^2) \ge s_1^*(h_1^1)$ and

$$
\begin{aligned}
g_1\big(s, h_1^2\big) - g_1\big(s, h_1^1\big) &= -\big(h_1^2 - h_1^1\big) \sum_{j=l_1+s+1}^{l_1+s_1^*(h_1^1)} \sum_{i=j}^{\infty} b_i - \big(\pi_1 + h_1^2\big) \sum_{l_1+s_1^*(h_1^1)+1}^{l_1+s_1^*(h_1^2)} \sum_{i=j}^{\infty} b_i \\
&\quad + \pi_1 \sum_{j=s_1^*(h_1^1)+1}^{s_1^*(h_1^2)} \sum_{i=j}^{\infty} b_i + \pi_1 \sum_{j=s_1^*(h_1^1)}^{s_1^*(h_1^2)-1} b_j \\
&= -\big(h_1^2 - h_1^1\big) \sum_{j=l_1+s+1}^{l_1+s_1^*(h_1^1)} \sum_{i=j}^{\infty} b_i - \big(\pi_1 + h_1^2\big) \sum_{l_1+s_1^*(h_1^1)+1}^{l_1+s_1^*(h_1^2)} \sum_{i=j}^{\infty} b_i \\
&\quad + \pi_1 \sum_{j=s_1^*(h_1^1)}^{s_1^*(h_1^2)-1} \sum_{i=j}^{\infty} b_i \\
&\le -\big(h_1^2 - h_1^1\big) \sum_{j=l_1+s+1}^{l_1+s_1^*(h_1^2)} \sum_{i=j}^{\infty} b_i \\
&\le 0.
\end{aligned}
$$

The first inequality above is because of the definition of s_1^*. So δ is also decreasing in this region. Finally,

$$
\begin{aligned}
&\delta\big(s^{-1}(s+l_2)\big) - \lim_{h_1 \to [s^{-1}(s+l_2)]^+} \delta(h_1) \\
&\quad = g_2\big(s, s^{-1}(s+l_2)\big) - \lim_{h_1 \to [s^{-1}(s+l_2)]^+} g_1(s, h_1) \\
&\quad = -\big(\pi_1 + s^{-1}(s+l_2)\big) \sum_{j=l_1+s+1}^{l_1+l_2+s} \sum_{i=j}^{\infty} b_i - \pi_1 \sum_{j=s+l_2}^{s_1^*-1} b_j \\
&\qquad + \lim_{h_1 \to [s^{-1}(s+l_2)]^+} \left\{ (\pi_1 + h_1) \sum_{j=l_1+s+1}^{l_1+s_1^*} \sum_{i=j}^{\infty} b_i - \pi_1 \sum_{j=s+l_2+1}^{s_1^*} \sum_{i=j}^{\infty} b_i \right\}.
\end{aligned}
$$

Since $s_1^* \to s + l_2 + 1$ as h_1 approaches $s^{-1}(s+l_2)$ from the right, we therefore have

$$
\begin{aligned}
\delta\big(s^{-1}(s+l_2)\big) - \lim_{h_1 \to [s^{-1}(s+l_2)]^+} \delta(h_1) &= \big(\pi_1 + s^{-1}(s+l_2)\big) \sum_{i=s+l_1+l_2+1}^{\infty} b_i - \pi_1 \sum_{i=s+l_2}^{\infty} b_i \\
&= 0,
\end{aligned}
$$

where the last equality is due to the definition of s^{-1}. So δ is decreasing in h_1 in $[0, \infty)$ and the result follows. □

Theorem 5 *Let $u_{tb}(\mathbf{x}, s)$ be the cost of the threshold-based heuristic. Of all N orders, m of them have not been confirmed. Let $\delta(k) = (n_1, n_2, \ldots, n_k)$ represents a permutation of $(1, 2, \ldots, k)$. Let $\Lambda(k)$ be the set of all possible permutations. Then,*

$$
u(\mathbf{x}, s) \le u_{tb}(\mathbf{x}, s) \le \max_{\delta(N) \in \Lambda(N)} \left[l_{n_1}\pi_{n_1} + (l_{n_1} + l_{n_2})\pi_{n_2} + \cdots + \sum_{i=1}^{N} l_{n_i}\pi_{n_N} \right] + \sum_{i=1}^{m} \pi_i .
$$

Proof The first inequality is obvious. For the second, suppose $(n_1, n_2, \ldots, n_N)$ represent the order by which the products are produced; that is, product n_1 is produced first, n_2, the second, etc. By the heuristic, the supplier may start the production of a product either because an order have been confirmed or because the time has passed a threshold and hence production must be started preemptively.

If product $n_i \in C$, then its cost is upper bounded by $(l_{n_1} + \cdots + l_{n_{i-1}} + l_{n_i})\pi_{n_i}$; that is, its cost is highest when its order is confirmed when the production of product n_1 is started.

Suppose that product $n_i \in \bar{C}$. If product n_i is produced after its order is confirmed, then its cost is upper bounded by $(l_{n_1} + \cdots + l_{n_{i-1}} + l_{n_i})\pi_{n_i}$. If product n_i is produced preemptively before its order is confirmed, then its cost is upper bounded by $(l_{n_i} + 1)\pi_{n_i}$ because the strategy of producing it as soon as the order is confirmed but not before is a feasible strategy. But $(l_{n_i} + 1)\pi_{n_i}$ is upper bounded by $(l_{n_1} + \cdots + l_{n_{i-1}} + l_{n_i})\pi_{n_i} + \pi_{n_i}$. The total number of products in $\bar{C}$ is m. Therefore, given a sample path $(n_1, n_2, \ldots, n_N)$, the cost of the heuristic is upper bounded by

$$\left[l_{n_1}\pi_{n_1} + (l_{n_1} + l_{n_2})\pi_{n_2} + \cdots + \sum_{i=1}^{N} l_{n_i}\pi_{n_N} \right] + \sum_{i=1}^{m} \pi_i,$$

and hence the result follows.

The bound is independent of time and distributions of confirmation times. Its computation is straightforward. The cost bound is constructed by computing the maximal cost for each product under the heuristic. Each permutation $(n_1, n_2, \ldots, n_N)$ represents a possible sequence by which the products are produced. For an arbitrary product n_i, its cost can't be higher than $(l_{n_1} + l_{n_2} + \cdots + l_{n_i} + 1)\pi_{n_i}$. From the bound of the heuristic, one can compute the maximal cost of not considering the capacity linkage across products. Based on our numerical studies, unfortunately, the bound is not very tight. □

References

Barlow, R., & Proschan, F. (1996). *Mathematical theory of reliability*. Philadelphia: Society for Industrial and Applied Mathematics.

Benjaafar, S., Cooper, W. L., & Kim, J. S. (2005). On the effects of pooling in production-inventory systems. *Management Science*, *51*, 548–565.

Christopher, M. (2000). *Logistics and supply chain management* (2nd ed.). New York: Prentice Hall.

Cohen, M. A., Ho, T. H., Ren, Z. J., & Terwiesch, C. (2003). Measuring imputed cost in the semiconductor equipment supply chain. *Management Science*, *49*, 1653–1670.

Duenyas, I. (1995). Single facility due date setting with multiple customer classes. *Management Science*, *41*(4), 608–619.

Duenyas, I., & Hopp, W. J. (1995). Quoting customer lead times. *Management Science*, *41*(1), 43–57.

Elhafsi, M. (2002). Optimal leadtimes planning in serial production systems with earliness and tardiness costs. *IIE Transactions*, *34*, 233–243.

Ha, A. (1997). Optimal dynamic scheduling policy for a make-to-stock production system. *Operations Research*, *45*, 42–53.

Ha, A. (2000). Stock rationing in an M/Ek/1 make-to-stock queue. *Management Science*, *46*, 77–87.

Jia, C. (2001). Stochastic single machine scheduling with an exponential distributed due date. *OR Letters*, *28*, 199–203.

Jang, B. (2005). Air Canada slips on Boeing order list. *The Globe and Mail*, July 5, 2005.

Katircioglu, K., & Atkins, D. (1996). *New optimal policies for a unit demand inventory problem* (Working Paper). Faculty of Commerce and Business Administration, The University of British Columbia.

Li, Q. (2007). Risk, risk aversion and the optimal time to produce. *IIE Transactions*, *39*, 145–158.

MacCormack, A., & Mason, R. (2005). *The fate of the Vasa*. Harvard Business School case.

Matlack, C., & Holmes, S. (2005). Why airbus is losing altitude. *BusinessWeek*, June 20, 2005.

Mora, H., Hausman, W., Lee, H., & Amaro, C. (2011). Project-based supply chains with uncertain demand timing and changing technological requirements. In *2011 MSOM annual conference*.

Pinedo, M. (2002). *Scheduling: theory, algorithms, and systems* (2nd ed.). Upper Saddle River: Prentice Hall.
Porteus, E. L. (2002). *Foundations of stochastic inventory theory*. Stanford: Stanford University Press.
Ross, S. (1983). *Introduction to stochastic dynamic programming*. San Diego: Academic Press.
Wein, L. M. (1991). Due-date setting and priority sequencing in a multiclass M/G/1 queue. *Management Science*, *37*(7), 834–850.