Production, Manufacturing and Logistics

A real-time decision rule for an inventory system with committed service time and emergency orders

Shuo Huang\textsuperscript{a,}*\textsuperscript{a}, Sven Axåter\textsuperscript{b}, Yifan Dou\textsuperscript{a}, Jian Chen\textsuperscript{a}

\textsuperscript{a}School of Economics and Management, Tsinghua University, Beijing 100084, China
\textsuperscript{b}Department of Industrial Management and Logistics, Lund University, P.O. Box 118, 22100 Lund, Sweden

1. Introduction

The progress in information technology in recent decades has created new possibilities for efficient inventory management. Inventory managers can now make their decisions based on more complete information. More specifically, it is now possible to track inventory, outstanding orders, and backordered demands continuously with the help of real-time information systems. However, most inventory control systems use only a subset of this type of information. Our model concerns inventory control for an online retailer and the use of real-time information.

A response time for an online retailer is quite common in connection with e-business. That is, customers can find more or less whatever they want on an online retailer’s website, while the retailer may not actually have every listed item in stock. Usually, there is a buffer period in which the retailer may satisfy customer demands. In most cases, customers accept such a waiting time because of the convenience of online purchasing. We call this period of time “committed service time”. Due to this committed service time, the online retailer can reduce its inventory substantially, because there are no backorder costs during this period. However, setting a longer response time may affect demand, i.e., the demand rate faced by an online retailer is assumed to be a decreasing function of the committed service time.

We consider a continuous review \((nQ, R)\) inventory policy for such an online retailer. It is obvious that the costs of an inventory system depend a great deal on what happens to demands when the system is in shortage. In our model, the problem is to handle situations when a customer demand has occurred some time ago and the retailer is still out of stock. Should the demand be backordered, or should it be filled by an emergency order? How is this decision affected by a retailer’s committed service time? In this paper, we present a decision rule that is based on real-time information and minimizes the expected costs under the assumption that future excess demands will be backordered. This rule is then used repeatedly as a heuristic. A simulation study illustrates the performance of the suggested rule.

The remainder of this paper is organized as follows. In Section 2 we discuss the relevant literature. Section 3 provides a detailed problem formulation. The decision rule is derived in Section 4, and numerical examples are presented in Section 5, where we also illustrate the performance of the suggested decision rule. Section 6 concludes the paper.

\* Corresponding author. Tel.: +86 10 62789795; fax: +86 10 62784555. E-mail address: huangsh@sem.tsinghua.edu.cn (S. Huang).
2. Literature review

Our decision rule combines backorders and emergency orders. In our case emergency orders are essentially equivalent to lost sales. Our inventory model is therefore closely related to the literature on partial backordering. Inventory models that consider partial backordering have been widely studied in the past. An early study by Montgomery et al. (1973) considered a mixture of backorders and lost sales. They assumed that a fixed fraction of the unfilled demand is backordered, and the rest is lost. Both deterministic and stochastic demands were considered. Their model for deterministic demand was extended by many researchers. Rosenberg (1979) introduced a “fictitious demand rate” and provided an easier procedure for analyzing the model. Park (1982) studied a similar model by defining a time proportional shortage cost and a fixed penalty cost per unit for a lost sale. He obtained the optimal solution and analyzed the sensitivity numerically. The sensitivity was also considered by Chu and Chung (2004) and Yang (2007).

The stochastic demand model developed by Montgomery et al. (1973) was extended by Kim and Park (1985). In their model, the backorder cost is assumed to be proportional to the length of time for which a backorder exists. They examined partial backordering in the context of a continuous review (Q, R) system. Assuming at most one outstanding order at any point of time, they provided an iterative procedure to determine the optimal order quantity and reorder point. Zeng (2001) studied the effects of using a partial backordering approach for inventory control under deterministic and random demands. She focused on identifying the conditions under which a partial backordering policy outperforms the traditional backordering policy.

Partial backordering can also be modeled by taking customer behavior into account. For example, Posner and Yansouni (1972) analyzed a model with Poisson demand and exponential lead times. Their inventory policy was similar to that in Kim and Park (1985). However, they assumed that customers are impatient, and that customers are willing to wait for a random period of time. Das (1977) used an (S – 1, S) policy and assumed that customers are willing to wait for a fixed period of time before canceling their orders. Moinzadeh (1989) also studied an (S – 1, S) policy with partial backordering, but assumed a constant lead time.

In the last few years, different customer impatience functions have been used to describe customer behavior when shortages occur. Abad (1996) was the first to propose customer impatience functions in connection with partial backordering. He developed an inventory model with dynamic pricing and lot-sizing for a retailer who sells a perishable item. He assumed that demand can be partially backordered and uses a customer impatience function to model it. There are different customer impatience functions. In Abad (2001), when a shortage occurs, a fraction of the demand at a given time is backordered. Different from Montgomery et al. (1973), this fraction is not fixed, but is a decreasing function of the waiting time. San José et al. (2005) studied a continuous review inventory model with deterministic demand. The fraction of backordered demand is characterized by an arbitrary customer impatience function that is a nonincreasing function of the waiting time. They also studied a special case where unsatisfied demand is partially backlogged according to an exponential function and they developed a general method to find the optimal policy (San José et al., 2006).

Although the decisions about waiting or leaving the system are typically made by customers, a retailer can influence the choice of customers. A customer may, for example, be given a rain check if she is willing to wait. DeCroix and Arreola-Risa (1998) and Cheung (1998) considered inventory systems that offer economic incentives (time-based price discounts) to customers who are willing to wait for longer than the normal delivery time.

The ability to control customer behavior gives retailers a chance to use partial backordering to improve the performance of their inventory systems. For such cases, a number of new inventory control policies have been presented. These policies usually incorporate new parameters to control the excess demands. Rabinowitz et al. (1995) introduced a third control parameter into the traditional (Q, R) policy: the maximum number of backorders allowed. During a shortage period, the first b units of incoming demand are backordered, and the rest are lost. They employed a search procedure to find the optimal values of Q, R, and b. However, Zhang et al. (2003) pointed out that this policy may not be the most cost-effective one because of the time-dependent cost of backorders. They presented another policy to limit backordered demands. Still in a (Q, R) system, during the lead time there is a cutoff point r. Before this point, if a shortage occurs, incoming demands are filled by emergency orders; after this point all unfilled demands are backordered. They also provided a solution procedure to find the three decision variables.

Another related, but somewhat different stream of the literature focuses on multiple supply modes and emergency supplies. It is quite common in practice for a retailer to have more than one supply mode and to choose from among these different modes depending on the situation. When different modes provide different lead times, a complex decision faced by the retailer is when to use a particular mode to replenish inventory. A large number of models in the literature address this problem. The simplest considers only two supply modes: a normal mode with a longer lead time and a lower cost, and an expedited mode with a shorter lead time but a higher cost. Barankin (1961) was the first to study this problem. He developed a one-product, single-period inventory model with two supply modes; and the two options are a one-period and a zero-period lag only. Daniel (1962) and Neuts (1964) extended this model to multiple periods and derived similar forms of the optimal policy. Fukuda (1964) also considered a fixed set-up cost when ordering and allowed both expedited and regular orders to be issued simultaneously. Wright (1968) investigated an inventory system with multiple products. However, in his model it was assumed that the normal lead time and the expedited lead time differed by one period only. Whitmire and Saunders (1977) analyzed a more general case by considering a multi-period dynamic model and allowing both the normal and regular lead times to be of arbitrary lengths. Unfortunately, the form of the optimal policy they derived is extremely complex. Furthermore, they did not include a fixed ordering cost. They were able to obtain explicit results only for the case where the two lead times differed by one period. Chiang and Gutierrez (1996) also studied a model with two supply modes, but they considered a periodic review system where the supply lead times may be shorter than the review periods.

Inventory systems with more than two supply modes have also been studied by a number of researchers. For example, Fukuda (1964) first addressed the three-supply-modes problem. However, he assumed that orders could only be placed every second period. Under this assumption, the problem is reduced to a dual-supply-mode problem that can be solved with previously known techniques. Zhang (1996) extended Fukuda’s work to a general three-supply-modes inventory system. Due to the complex form of the optimal ordering policy, she presented a heuristic policy to estimate the optimal base-stock levels for the three supply modes. Feng et al. (2005) incorporated forecast updates into a three-supply-modes model; they proved the existence of an optimal policy for this model, and obtained the structure of the optimal policy.
The optimal policy in a continuous review system with emergency orders appears to be extremely difficult to obtain; thus, most of the related research focuses on certain extensions of the well-known \((Q,R)\) policy. Kalpakam and Sapna (1995) considered a policy that requires an emergency order for \(R\) units to be issued whenever the on-hand inventory reaches zero. Mooinzadeh and Nahmias (1988) formulated a more general continuous review emergency order model. They also considered an extension of the standard \((Q,R)\) policy to allow for different lot sizes, \(Q_1\) and \(Q_2\), and two different reorder points, \(R_1\) and \(R_2\), where \(Q_1\) and \(R_1\) are the normal order quantity and reorder point, and \(Q_2\) and \(R_2\) are the corresponding parameters for emergency orders. Based on the assumption that there is at most one normal outstanding order, they obtained expressions for the expected on-hand inventory and the expected backorders, and a simple search procedure for determining the policy parameters is also given. However, their model does not include a time-dependent backorder cost. Johansen and Thorstenson (1998) considered a more realistic cost structure that includes fixed and variable ordering costs, backordering costs, and holding costs. They also assumed that there is at most one normal outstanding order. When a normal order is outstanding, emergency orders are issued depending on the remaining delivery time of this normal order. In a recent study, Axsäter (2007) relaxed some assumptions made in Johansen and Thorstenson (1998). Given the reorder point and lot size for normal orders, and based on real-time information regarding the remaining delivery time, he developed a heuristic decision rule to determine the timing and size of the emergency orders. Simulation studies indicate that this technique performs quite well in some situations of practical interest.

In this paper, we consider a system in which a retailer can determine whether unfilled demand should be filled by an emergency order or backordered. Our study differs from the existing literature in the following aspects. First, in most of the related studies, retailers only control unfilled demand according to a predetermined policy, without considering real-time information when a shortage occurs. However, due to the rapid development of information technology, in most real systems, this type of information is readily available. Axsäter (2003) studied lateral shipments in inventory systems and derived a decision rule based on real-time information to make transshipment decisions. Axsäter (2007) considered emergency orders based on real-time information. This seems to be the work that is most related to our model. However, our work differs from Axsäter (2007) in that we focus on an online retailer facing demand that depends on the committed service time. That is, we consider an online retailer with a committed service time in conjunction with partial backordering. A committed service time is quite common in supply networks (Simchi-Levi et al., 2005), especially in the e-business environment. Although the committed service time may be very short because of competition among online retailers, it does exist and will affect the analysis. Second, we assume that the demand rate faced by the online retailer is a decreasing function of the committed service time. This is a reasonable assumption because different customers may have different sensitivities with respect to the delivery lead time. If a retailer can provide only a relatively long committed service time, it will lose customers who require quicker deliveries.

A committed service time is somewhat similar to the demand lead time defined by Harirhan and Zipkin (1995). They assumed that each customer order has a due date, and the time from a customer’s order until that due date is defined as the demand lead time. After modeling supply and demand lead times in different ways, they came to a quite simple and intuitively appealing conclusion: demand lead times directly offset supply lead times. Unfortunately, their simple conclusion is not valid in our model, which is because their model does not allow orders to be filled early, but ours does. In a just-in-time (JIT) purchasing environment, their assumption is reasonable. However, as online customers, we have all probably experienced the following situation. After we have placed our order, we are notified that the delivery will take place before a certain date, but not necessarily on the due date (Wang and Toktay, 2008). For example, Amazon China uses this method in its order fulfillment process. Furthermore, there is a possibility of rejecting demands in our model, while Harirhan and Zipkin (1995) assume complete backordering. Demand lead times have also been considered in other advance demand information models, such as Graves and Willems (2000), Marklund (2006), etc. For further discussion of advance demand/order information, we refer readers to Gallego and Özer (2002) and the references therein.

3. Problem formulation

We consider an online retailer that faces compound Poisson demand. This distribution is widely used in literature, including that on online business. Specifically, based on real data, a study (Tatsiopoulos et al., 2002) on online clothing industry shows that the time between two successive order arrivals is exponentially distributed, which justifies our assumption on demand. The retailer has a committed service time, which affects its demand rate. Customers view this period of time as a necessary time for handling their orders by the retailer, and are willing to wait for this period of time after they have placed their orders. They may accept this waiting time because of the advantages of online business, such as the variety of products, the purchasing convenience, etc. When a demand occurs, if the retailer has stock on-hand, it will be satisfied immediately. But in case of shortage, the retailer can use this period of time as a buffer and there are no backorder costs as long as this demand is satisfied within the committed service time. However, after this time has passed, if the retailer is still out of stock, he must decide whether to fill the demand in question by an emergency order or backorder it with a time-proportional backorder cost.

An emergency order incurs an additional cost per unit. (Recall that we can also use the interpretation that the demand is lost and that this incurs a lost profit per unit.) If the retailer decides to backorder the considered demand, time-proportional backorder costs are incurred. This can be interpreted as economic compensation for customers who are willing to wait for longer than the committed service time. Hence, a problem faced by the retailer is whether to choose emergency orders or backorders, considering the available real-time information.

Normally, the retailer replenishes his stock from an outside supplier. We assume that this supplier has an infinite supply; hence, the normal lead time of the retailer is constant. We assume also that an emergency order takes no time. A meaningful committed service time is usually very short. Specifically, it should be shorter than the lead time. Otherwise, the retailer would be able to satisfy all demands without holding any inventory. Demands that are backordered after the committed service time are satisfied later on a first-come, first-served basis.

We assume that the retailer uses a continuous review \((nQ,R)\) policy for normal replenishments, i.e., when the inventory position \(nQ\) on-hand plus outstanding orders minus backorders\) de- clines to or below the reorder point \(R\), he orders a number of batches so that the resulting inventory position is in the interval \([R+1, R+Q]\). Note that the resulting number of batches is unique. We consider ordering, holding, backordering, and emergency ordering costs.

Before proceeding, we introduce the following notation, which is used throughout the paper:
In Section 4.1, we provide a method to describe the state of the system. An emergency order affects the expected costs indirectly. Then in Section 4.2, we derive the long-term expected costs with complete backordering. In Section 4.3, we compare the costs for emergency orders can be placed in the future. This idea has previously been used by Axsäter (2003, 2007).

An emergency order affects a retailer's cost in two ways. First, an emergency order imposes a direct cost per unit, but it also eliminates the corresponding profit, i.e., \( ac \) per unit, in the same way as for emergency orders. We first derive a decision rule for emergency orders (or lost sales), given the reorder point and batch quantity for normal replenishments. Specifically, after the committed service time, should a waiting demand be covered by an emergency order or backordered? When making such decisions about emergency orders or backorders, we assume that the retailer has complete information about the current state of the system, such as the remaining delivery times of the outstanding orders and the incoming times of the existing waiting demands.

### 4. Decision rule

In this section, we present a decision rule for using emergency orders. The principal idea of this rule is to make the best choice under real-time information and the assumption that no further emergency orders can be placed in the future. This idea has previously been used by Axsäter (2003, 2007).

An emergency order affects a retailer's cost in two ways. First, an emergency order imposes a direct cost per unit, but it also eliminates the corresponding share of the ordering cost for a normal replenishment. An emergency order fills demands but is not ordered from the normal supplier. Therefore we avoid the ordering cost for a normal replenishment. Second, since an emergency order means that a demand disappears from the backordered queue without using a normal replenishment, it changes the state of the system. Suppose that an emergency order takes place at time \( t \). It is obvious that the expected cost for the retailer after \( t \) depends on the state of the system at time \( t \); hence, the emergency order affects the expected costs indirectly.

We derive our decision rule in the following way. First in Section 4.1, we provide a method to describe the state of the system. Then in Section 4.2, we derive the long-term expected costs with complete backordering. In Section 4.3, we compare the costs for a given initial state with the long-term expected costs derived in Section 4.2, using the system state defined in Section 4.1. We make this comparison to see whether we should satisfy a certain demand by an emergency order. Based on these cost differences, the decision rule is given in Section 4.4. Recall that in this section, it is assumed that the batch quantity and reorder point for normal replenishments are given.

#### 4.1. System state

To simplify the derivation of the decision rule, we need to set up our description of the system state. At any given time, the retailer may have stock on-hand or backorders. That is, the retailer's inventory level may be positive or negative. Corresponding to each inventory level (stock on-hand minus backorders), there is a unique inventory position, and the difference between the inventory position and the inventory level is the quantity of outstanding orders. Suppose the retailer's inventory level is \( IL \). We define

\[
(II) = IL + iQ
\]

for \( IL \leq R + Q \), where \( i \) is the smallest integer that satisfies \( (II) \geq R + 1 \). With this definition, \( (II) \) is the inventory position corresponding to a certain inventory level \( IL \), and \( i \) is the number of outstanding orders. For \( IL > R + Q \), we let \( (II) = IL \).

The inventory state of the retailer can be defined as

\[
X = (IL, IP_1, IP_{-1}, \ldots, IP_{1-(II)}, WP_{-1}, WP_{-2}, \ldots, W_{1-(II)}),
\]

where \( IL \) is the inventory level, and \( IP \) is the inventory position corresponding to \( IL \). We assume that we have complete information about outstanding orders; thus, \( t \) is the remaining delivery time for the item which will be used to satisfy the ith future demand from now. Similarly, \( w \) is the waiting time already incurred by an existing backordered demand. Of course, if the corresponding demand has not occurred, \( w = 0 \).

Note that (2) indicates that all items on-hand or outstanding correspond to certain demands. If \( IL \geq 0 \), all units, no matter on-hand or outstanding, will be used for future demands, and the units on-hand have zero remaining delivery time. If \( IL < 0 \), some demands have already been backordered and no units have zero remaining delivery time. Some of the units in outstanding orders will be used to satisfy demands that have already come. Furthermore, since the retailer issues orders in batches of \( Q \) units, all of the units in the same order must have the same remaining delivery time.

Besides the information about the on-hand inventory and outstanding orders, the retailer also keeps track of the customers' waiting times. Thus, if there is a shortage, the retailer knows exactly how long the customers have been waiting. The waiting times are denoted as \( WP_{-1}, WP_{-2}, \ldots, W_{1-(II)} \). Clearly, \( W_s = 0 \) for \( s > 0 \) because the corresponding demand has not yet occurred. If \( IL \geq 0 \), no customers are waiting. However, if \( IL = -2 \), for example, two customers are waiting with waiting times \( W_{-1} \geq WP_{-2} \).

#### 4.2. Costs in the traditional backorder model

In this section, we derive the long-term expected costs per unit of time for a traditional inventory model with backorders that incur time-proportional backorder costs after a committed service time. The costs derived in this section are used to analyze the cost for a given initial state in the next section.

The approach to evaluating retailer costs is standard, except for the calculation of the backorder cost, which is different because of the committed service time. Consider the inventory position at any given time \( t \) and the inventory level at \( t + L \). Everything that is included in the inventory position at time \( t \) will have arrived by time \( t + L \). Furthermore, possible orders triggered during \( (t, t + L) \)
This means that we have to replace the inventory position at time $t$ with the expected rate of the total holding and backorder costs at time $t + L$. If the demands or emergency orders are different from the average demand, the state of the retailer will change. However, if this occurs in a short time, the waiting time at this stage is not incurred. Thus, we only need to consider the backorder cost if the item arrives during $(T - w, L]$ (the grey area in the left-hand figure in Fig. 1). However, if $w > T$, then no matter when the corresponding item arrives, the retailer will be charged a backorder cost for this item. The total cost of this item in this case is

$$
\begin{align*}
C(k) &= h\{k - D(L)\}^+ + b(k - D(L - T))^- \\
&= h\{k - D(L)\}^+ + b(k - D(L - T))^- - b(k - \hat{i}(T)(L - T)\mu_i) \\
&= he^{-\lambda(T)} \sum_{i=0}^{k-1} \frac{(\lambda(T)L)^i}{i!} + b e^{-\lambda(T)(L-T)} \\
&\times \sum_{i=0}^{k-1} \frac{(\lambda(T)(L-T))^i}{i!} - b(k - \hat{i}(T)(L - T)\mu_i).
\end{align*}
$$

(3)

When applying a $(nQ, R)$ policy, it is well known that in steady state, the inventory position is uniformly distributed on the integers $\{R + 1, R + 2, \ldots, Q \}$ (Hadley and Whitin, 1963). Thus, the average long-term holding and backorder costs per unit time can be obtained as

$$
C = \frac{1}{Q} \sum_{k=R+1}^{Q} C(k).
$$

(4)

4.3. Retailer costs for a given initial state

It is obvious that, given a certain initial state, the total expected future holding and backorder costs for a retailer (without rejected demands or emergency orders) will be different from the average costs derived in (4). In this section, we compare the retailer holding and backorder costs for a given initial state with the long-term expected costs and obtain their difference. If a demand is backordered, the state of the retailer will change. However, if this demand is satisfied by an emergency order or rejected, this state remains unchanged, but an additional cost per unit is incurred. Hence, by determining the cost difference for a given initial state, we can make the optimal decision concerning whether to satisfy a demand by an emergency order under the assumption that no further emergency orders are allowed.

For a given initial state $X$, we define $\alpha(X)$ as the expected cost difference between the lead time costs under this initial state $X$ and the long-term expected holding and backorder costs, $\beta(IP)$ as the expected cost difference from the long-term average holding and backorder costs after the lead time, and $\gamma(X) = \alpha(X) + \beta(IP)$ as the total expected cost difference from the long-term average holding and backorder costs given initial state $X$. We note that $\beta(IP)$ is a function of the inventory position alone, because the costs after the lead time depend only on the inventory position and the lead time demand.

Consider now an item that has a remaining delivery time $t$ until it arrives at the retailer. If $t = 0$, it means that the item is already on-hand. Assume that this item will meet the nth customer demand from now. If $s \leq 0$, it means that this demand has already occurred. When deriving $\gamma(X)$, we start by evaluating the associated holding and backorder costs for this item. For this purpose, we define $X(s, t, w)$ as the expected holding and shortage costs for an item during the lead time, given that the remaining delivery time for this item is $t$, the waiting time at this stage is $w$, and the item will be used to satisfy the sth customer demand. Recall that, for $s > 0$, we have $w = 0$.

The calculation of $X(s, t, w)$ can be divided into four cases.

(1) $s \leq 0$, $t \leq L$. This means that the demand has already occurred and that the order for the corresponding item has been triggered. This demand has been waiting at the retailer for time $w$. We need to consider the situations for $w < T$ and $w \geq T$, as shown in Fig. 1. The considered time is represented as time 0 in the figure.

When $w < T$, that is, the demand has arrived just recently, the retailer will not be charged any cost if the corresponding delivery arrives before $T - w$. Thus, we only need to consider the backorder cost if the item arrives during $(T - w, L]$ (the grey area in the left-hand figure in Fig. 1). However, if $w > T$, then no matter when the corresponding item arrives, the retailer will be charged a backorder cost for this item. The total cost of this item in this case is

$$
\begin{align*}
\alpha(s, t, w) = \begin{cases} 
bt & \text{if } w < T, \\
bl & \text{if } w \geq T.
\end{cases}
\end{align*}
$$

(5)

(2) $s < 0$, $t > L$. In this case the demand has arrived, but the order for the corresponding item has not yet been sent out. Although it is not common in real inventory systems, we take this case into consideration for completeness. The total cost is

$$
\alpha(s, t, w) = \begin{cases} 
b(T - w) & \text{if } w < T, \\
bl & \text{if } w \geq T.
\end{cases}
$$

(6)

(3) $s > 0$, $t \leq L$. Because $w = 0$, we write $\alpha(s, t)$ instead of $\alpha(s, t, w)$. The demand has not arrived, but the order for the corresponding item has been triggered. This is common in inventory systems with a positive reorder point. In this case, during the lead time, the retailer will be charged a holding cost if the demand arrives after $t$, but before the end of the lead time, as illustrated in the grey areas in Fig. 2. Otherwise, if the demand arrives before $t$, we must consider the committed service time to determine the backorder cost. If the demand arrives after $t - T$, there is no backorder cost due to the committed service time. However, if the demand arrives earlier than $t - T$, there is a backorder cost for the time before $t - T$ (see the black area in the left-hand part of Fig. 2). Furthermore, obviously there will be an inventory holding cost if the corresponding demand comes after the lead time, which is not shown in the figure. Recall that we consider compound Poisson demand. Define $P_0^n$ as the probability that the nth unit is ordered by the nth customer. We have

$$
\sum_{k=1}^{\infty} \sum_{k=1}^{\infty} P_0^n = \sum_{k=1}^{\infty} P_0^n - \sum_{k=1}^{\infty} P_0^{n-1}.
$$

(7)

From Eq. (7) (note that $P_0^n = 0$), we can conclude that

$$
\sum_{k=1}^{\infty} P_0^n = P_0^{n-1} + \sum_{k=1}^{\infty} P_0^{n-1}.
$$

(8)

Eq. (8) can be used to determine the probabilities, $P_0^n$, recursively.

Also, the time when the nth customer arrives at the retailer has an Erlang ($\lambda(T), n$) distribution with density function

$$
\begin{align*}
\text{Fig. 1. Case 1, } s \leq 0, t \leq L.
\end{align*}
$$
vice time. Thus, no backorder cost is incurred. The last situation is the lead time.

using (14), (13) can be reformulated as

\[ x(s,t) = \begin{cases} b(1-R_t) & \text{if } t > T, \\ h[U_t(L) - U_t(t)] & \text{if } t \leq T. \end{cases} \]

Now we can derive the expected cost difference relative to the long-term average holding and backorder costs during the lead time, \( \alpha(X) \), as follows.

\[ \alpha(X) = \sum_{s=1}^{R_Q} x(s,t,w) + \sum_{s=R_Q+1}^{R_Q+Q} x(s,t,w) - L \times C. \] (16)

In (16), \( t_i \) is the remaining delivery time of the item that corresponds to the \( i \)th demand. The terms in the second sum quickly decline to zero as \( s \) increases. In these terms, \( t_i > L \) is unknown. However, as shown in (6) and (15), we can then replace \( t_i \) with \( L \).

Now we can determine the distribution of the time when the corresponding item has not yet been triggered. In this case, the order unit is demanded. The density function of this time is obtained as

Thus, there is no backorder cost, but only a holding cost for the order quantity of the item that corresponds to the \( i \)th demand. The inventory position remains unchanged during \( [t_i, t_{i+1}) \). Considering that the expected length of this time interval is \( 1/\lambda(T) \), the expected cost difference from the long-term average costs during this time interval is

\[ \delta(k) = (C(k) - C)/\lambda(T). \] (17)

Recall that since we apply a (nQ,R) policy, \( k \) is an integer in \([R + 1, R + Q]\). Suppose that the size of the incoming demand is \( d \). The inventory position will then change to \( (k-d) \). The cost difference, \( \beta(k) \), includes the cost difference in the initial step, \( \delta(k) \), and the total cost difference beyond this initial step. For \( k = R + 1, \ldots, R + Q \), we have

\[ \beta(k) = \delta(k) + \sum_{d=1}^{Q} f_d \beta((k-d)). \] (18)

This equation has a unique solution; thus, by solving it, we can obtain \( \beta(k) \) for \( k = R + 1, \ldots, R + Q \).

We do not need to consider \( k < R + 1 \). However, we may need to determine \( \beta(k) \) for \( k > R + Q \). This is because our decision rule is applied after time \( T \) of the customer’s arrival. The demand may have already triggered a normal order, even if it is satisfied by an emergency order at time \( T \) later. As a result, the inventory position may exceed \( R + Q \) in this case.

Let, for example, \( k = R + Q + 1 \). Then after some steps, the inventory position will fall into \([R + 1, R + Q]\) eventually. Thus, for \( k > R + Q \), we can still use Eq. (18) to derive \( \beta(k) \). When determining \( \beta(k) \) for \( k > R + Q \), we can start with \( k = R + Q + 1 \) and then successively go to \( k = R + Q + 2, R + Q + 3, \ldots. \)

Finally, we get the total expected cost difference as

\[ \gamma(X) = \alpha(X) + \beta(IP). \] (19)

4.4. Decision rule for emergency ordering

Given the cost difference derived above, we can evaluate the consequences of an emergency order. An emergency order affects
the retailer costs in two ways. First, it incurs certain direct costs. Second, it changes the future expected costs by changing the state of the inventory system.

Suppose the current system state is $X$. If the retailer decides to satisfy a waiting demand of size $d$ by himself later (backorder the demand), then the state will change to $X'(d)$, and the expected future cost difference compared to the long-term average cost is $\gamma(X'(d))$. However, if the retailer decides to satisfy $q$ demanded units by an emergency order, the state will change to $X(d, q)$. The total expected cost difference is $\gamma(X(d, q)) + q \times ac - q \times K/Q$, where $ac$ is the additional cost per unit, and $K/Q$ is the per-unit fixed ordering cost saved.

Comparing these two cost differences, we get

$$\Delta = \gamma(X'(d)) - \min_{q \leq d} \left[ \gamma(X(d, q)) + q \times ac - q \times K/Q \right].$$

(20)

Let $q^*$ be the minimizing $q$. The decision rule can be expressed as follows. If $\Delta \geq 0$, the retailer should satisfy $q^*$ units of the waiting demands by an emergency order. If $\Delta < 0$, the retailer should let the customers continue to wait. This decision rule is applied at time $T$ after the arrival of a demand that cannot be met from the stock on-hand. In this case, we let the demand be backordered the free time $T$. After this time, we apply the decision rule, provided that the demand has not already been met by a normal order.

We now briefly describe how we obtain $X'(d)$ and $X(d, q)$. Assume that we know the state (see (2)) just before demand $d$, which cannot be satisfied immediately by stock on-hand. This demand, which is then backordered, changes the inventory level and may also trigger a replenishment order. During time $T$, there may be more demands that similarly change the inventory level and the inventory position. Furthermore, the times until delivery and the waiting times also change. The resulting state is $X'(d)$.

Assume then that we omit $q$ demands. This means that the inventory level and the inventory position increase. The waiting times for other demands do not change. The remaining times until delivery for outstanding orders also do not change, but the number of future demands that it will meet does change.

Note that in Eq. (20), we consider the cost differences between two retailer states. Therefore, we can omit from Eq. (16) the terms that are identical for both states.

Fig. 3 illustrates our decision rule.

As indicated in the chart, the retailer will first try to satisfy the demand from stock on-hand. Unmet demands will stay for $T$ time units (the committed service time) with no backorder cost charged. After $T$ units of time, if the demands have still not been satisfied, the decision rule is applied, and the retailer will make the decision on emergency orders and backordered demands, respectively.

For given reorder point and order quantity, this decision rule always makes the best choice under the assumption that there are no further emergency orders. However, it is applied repeatedly as a heuristic. Each time after using the decision rule, the firm obtains an expected improvement compared with the alternative of complete backordering. Hence, it will improve the expected performance of the retailer compared with a traditional system in which excess demands are always backordered.

However, there seems to be no easy way to obtain the optimal reorder point and order quantity. In our numerical studies, we use simulations to optimize them. We also use our numerical results to verify the improvement effect of the decision rule when using the reorder point and order quantity that are optimal in the traditional backorder system.

5. Numerical study and discussion

In this section, a numerical study is used to evaluate our decision rule and to illustrate the effects of the committed service time on retailer performance. We show that our partial backordering
(PB) system does provide a cost benefit over traditional backordering (BO) and lost sales (LS) systems. A BO system means that no demands are rejected, and in an LS system, no backorders are allowed, which is equivalent to a system that satisfies all unmet demands by emergency orders.

For comparison with Axsäter (2003), we also use a geometric distribution to generate the demand size of a customer, that is, the probability that an incoming customer orders \( j \) units is \( f_j = p(1-p)^{j-1} \).

5.1. Comparison of three systems

First, we set the committed service time \( T = 0 \); hence, the customer arrival rate, \( \lambda(T) \), is fixed at \( A \). We use the following parameters in these examples: \( L = 3 \), \( h = 1 \), \( b = 5 \), and \( K = 8 \). By considering different demand processes characterized by \( A \) and \( p \), we can compare the expected costs for the PB system with those for the BO and LS systems. The average demand per unit of time is obtained as \( A/p \), i.e., in Table 1 it is 1 for the first two setups of \( A \) and \( p \) and 5 for the remaining two. The additional cost per unit for an emergency order \( (ac) \) varies from 4 to 10. For the BO system, the expected costs and the corresponding optimal batch quantity and reorder point are easy to derive using the results from Section 4.2. Since the decision rule in the PB system is based on real time information, it is very hard to determine \( Q \) and \( R \) analytically. We do not provide such a procedure but only obtain \( Q \) and \( R \) through simulation. In Table 1, we first use the \( Q \) and \( R \) that are optimal for the BO system, and then we use the optimal \( Q \) and \( R \) obtained by simulation. Because of the optimization of \( Q \) and \( R \), the heuristic is denoted as PB*. For comparison, we also give the results of the LS system, which means a simple heuristic where all excess demands are filled by emergency orders. For the LS system, we use the \( Q \) and \( R \) that are optimal for the BO system.

It is clear that the PB system has a significant cost advantage over the BO and LS systems for a wide range of \( A, p, \) and \( ac \), especially under optimal \( Q \) and \( R \). When \( ac \) is very large, for example, when it approaches 30, PB is almost equivalent to BO (for brevity, these results are not listed in the table). In such situations, emergency orders have a very small impact on the total costs when using our decision rule. This is because there are relatively few emergency orders due to the high \( ac \). When \( ac \) is very small, PB is essentially the same as LS. In such cases, the advantages of our decision rule are also limited. This is because almost all unfilled demands are satisfied by emergency orders due to the relatively low \( ac \). But for intermediate values of \( ac \), the PB system based on our decision rule is most cost effective. The cost advantages of the PB system are due to the flexibility of partial backordering.

The results in Table 1 are similar to those in Zhang et al. (2003), but our decision rule is easier to use. In Zhang et al. (2003), an additional decision variable, the cutoff time, is incorporated in the traditional \((Q, R)\) policy to control the PB system. Hence, the optimization is more complex. We do not need to add any decision variables. Instead, we use only the real-time information concerning the remaining delivery times of outstanding orders and the waiting times of the customers in the waiting line. This information is easy to obtain through current information systems. Consequently, our decision rule seems to be more practical.

Another major advantage of our technique is that it can be used in a more general setting. Zhang et al. (2003) assume that there is at most one outstanding order, whereas we can have several outstanding orders. Thus, our method may be more suitable when the fixed ordering cost is relatively low. In such situations, the order quantity may be smaller than the expected lead-time demand, which increases the likelihood of having several outstanding orders. Furthermore, when using the \( Q \) and \( R \) that are optimal in the BO system in the PB system, we have a performance guarantee, the expected outcome is always better for the PB system.

5.2. The effect of committed service time

Let us now turn to the case where the online retailer has committed service time. (The methods of Zhang et al. (2003) are not applicable in this case.) We first consider the case where the demand arrival rate is independent of the committed service time, \( T \). Thus, we can compare our results with those in Axsäter (2007) to illustrate the influence of \( T \) on retailer performance (we refer to Axsäter (2007) for a comparison between his real-time-based decision rule and other heuristic rules for emergency orders when no committed service time is allowed). Table 2 shows the results for different values of the committed service time, \( T \). We use the following parameters in Table 2: \( L = 5 \), \( h = 1 \), \( b = 10 \), \( K = 5 \), and \( ac = 10 \). We set \( \lambda(T) = 1 \) and \( p = 1 \), that is, we consider simple Poisson demand.

As in Table 1, the expected cost for the BO system, and the corresponding optimal batch quantity and reorder point, are derived analytically. Then we use the same \( Q \) and \( R \) to obtain the cost for the PB system. For the purpose of comparison, we also derive the optimal \( Q \) and \( R \) in the PB system by simulation. Again, these optimized results are denoted as PB*.

<table>
<thead>
<tr>
<th>( A )</th>
<th>( p )</th>
<th>( ac )</th>
<th>BO Q</th>
<th>BO R</th>
<th>BO Cost</th>
<th>LS Q</th>
<th>LS R</th>
<th>LS Cost</th>
<th>PB Q</th>
<th>PB R</th>
<th>PB Cost</th>
<th>PB* Q</th>
<th>PB* R</th>
<th>PB* Cost</th>
<th>% Saving over BO</th>
<th>% Saving over LS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>4</td>
<td>0.10 (0.01)</td>
<td>4.88 (0.01)</td>
<td>4</td>
<td>1.43 (0.01)</td>
<td>21.03</td>
<td>13.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>5.51 (0.01)</td>
<td>5.12 (0.01)</td>
<td>5</td>
<td>1.89 (0.01)</td>
<td>12.83</td>
<td>11.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>5.92 (0.01)</td>
<td>5.30 (0.01)</td>
<td>5</td>
<td>1.72 (0.01)</td>
<td>14.42</td>
<td>11.32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>6.33 (0.01)</td>
<td>5.44 (0.01)</td>
<td>5</td>
<td>1.61 (0.01)</td>
<td>13.31</td>
<td>11.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>0.6</td>
<td>4</td>
<td>0.10 (0.01)</td>
<td>5.45 (0.01)</td>
<td>4</td>
<td>1.72 (0.01)</td>
<td>27.29</td>
<td>16.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6.12 (0.02)</td>
<td>5.73 (0.01)</td>
<td>5</td>
<td>1.83 (0.01)</td>
<td>18.34</td>
<td>13.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>6.58 (0.02)</td>
<td>5.98 (0.01)</td>
<td>5</td>
<td>1.77 (0.02)</td>
<td>11.92</td>
<td>12.31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>7.01 (0.01)</td>
<td>6.17 (0.02)</td>
<td>6</td>
<td>1.68 (0.01)</td>
<td>6.32</td>
<td>13.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.8</td>
<td>4</td>
<td>13</td>
<td>14</td>
<td>12.37</td>
<td>13.49 (0.05)</td>
<td>11.91 (0.03)</td>
<td>13.16</td>
<td>12.83 (0.04)</td>
<td>1.77</td>
<td>12.23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>14.69 (0.05)</td>
<td>12.16 (0.04)</td>
<td>11</td>
<td>12.11 (0.03)</td>
<td>2.10</td>
<td>17.56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>15.90 (0.04)</td>
<td>12.29 (0.04)</td>
<td>12</td>
<td>12.29 (0.04)</td>
<td>0.65</td>
<td>22.70</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>17.10 (0.07)</td>
<td>12.35 (0.04)</td>
<td>13</td>
<td>12.35 (0.04)</td>
<td>0.16</td>
<td>27.78</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.6</td>
<td>4</td>
<td>14</td>
<td>14</td>
<td>14.07</td>
<td>14.77 (0.03)</td>
<td>13.19 (0.04)</td>
<td>13</td>
<td>12.68 (0.03)</td>
<td>9.88</td>
<td>14.15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>16.19 (0.07)</td>
<td>13.36 (0.04)</td>
<td>12</td>
<td>13.36 (0.05)</td>
<td>4.12</td>
<td>16.68</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>17.56 (0.06)</td>
<td>13.83 (0.04)</td>
<td>14</td>
<td>13.83 (0.04)</td>
<td>1.71</td>
<td>21.24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>18.95 (0.09)</td>
<td>13.98 (0.03)</td>
<td>14</td>
<td>13.98 (0.03)</td>
<td>0.64</td>
<td>26.23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
is a decreasing function of $T$, i.e., $\lambda = \lambda = A - BT$, where the maximum customer arrival rate $A$ is for $T = 0$, and $B$ represents the committed service time sensitivity of the arrival rate. According to Table 3, it is obvious that the benefit of our decision rule is decreasing with $A$, but increasing with $B$. More specifically, the decision rule suggested in this paper seems to be more beneficial when the potential market is smaller and when the customers are more sensitive to their waiting times.

6. Conclusions

In this paper, we have derived a decision rule for emergency ordering in a continuous review inventory model with a committed service time. The inventory system consists of an online retailer and its customers, where the mean customer arrival rate is a linear decreasing function of the committed service time. The assumption of a committed service time means a generalization of the decision rule in Assåter (2007). The suggested decision rule uses real-time information, which is easy to obtain in practice by using modern information technology. Hence, this decision rule is easy to apply. Furthermore, a committed service time is quite common in online businesses, which means that our decision rule can be used in many practical cases.

The decision rule has been evaluated in a simulation study, and the results indicate that it performs well. The expected costs, when applying our decision rule to choose between backordering and emergency ordering, are significantly lower than the corresponding costs in the case of complete backordering or when we do

5.3. When demand is related to the committed service time

Since the competitiveness of a retailer is closely related to the length of $T$, we now assume that the customer arrival rate is a decreasing function of $T$. In this way, we can analyze how the system performance is affected by changes in the demand-related parameters. Table 3 illustrates the effects of market characteristics on the optimal profits. The optimal $Q$ and $R$ in the BO case are again derived using the results in Section 4.2, while the corresponding optimal parameters in the PB case are obtained through simulation. Apart from the demand-related parameters, all of the other parameters are the same as in Table 2. We assume in Table 3 that $T = 0.5$. Again, we use a geometric distribution to generate the demand size of a customer and set $p = 0.8$. Because the profits are affected by the size of the demand, we now consider total profits rather than costs. As in Table 2, we assume that the additional cost per unit for emergency orders $ac = 10$, which equals the profit per unit.

Remember that we also assume that the customer arrival rate depends linearly on $T$, i.e., $\lambda = A - BT$, where the maximum customer arrival rate $A$ is for $T = 0$, and $B$ represents the committed service time sensitivity of the arrival rate. According to Table 3, it is obvious that the benefit of our decision rule is decreasing with $A$, but increasing with $B$. More specifically, the decision rule suggested in this paper seems to be more beneficial when the potential market is smaller and when the customers are more sensitive to their waiting times.

Table 2 shows that our decision rule has a clearly better performance, even when using the same $Q$ and $R$ as in the corresponding BO cases. Its relative advantage over BO decreases a little with increasing $T$. Table 2 also illustrates that the optimal $Q$ and $R$ obtained through simulation can significantly improve the performance. Furthermore, the cost advantage is relatively stable for different values of $T$, which indicates that the benefits of partial backordering remain about the same even when the retailer is more flexible due to the committed service time.

It is obvious that given a certain type of inventory control, when a committed service time is allowed, the expected costs are always lower than those when $T = 0$. Furthermore, these costs are decreasing with $T$. The optimal $Q$ and $R$ are decreasing in both the BO and PB cases. Even a short $T$ can reduce the expected costs significantly. This, among many other reasons, enables the price advantage of many online retailers. If $T$ is longer, then the expected costs may be lower, but the retailer may be less competitive. Committed service time can be viewed as a kind of service metric of the firm, and it should be aligned with the firm's business strategy. For example, Amazon has worked long and hard to ensure that its committed service time can be viewed as a kind of service metric of the firm, and its customers, where the mean customer arrival rate $A$ is for $T = 0$, and $B$ represents the committed service time sensitivity of the arrival rate. According to Table 3, it is obvious that the benefit of our decision rule is decreasing with $A$, but increasing with $B$. More specifically, the decision rule suggested in this paper seems to be more beneficial when the potential market is smaller and when the customers are more sensitive to their waiting times.

6. Conclusions

In this paper, we have derived a decision rule for emergency ordering in a continuous review inventory model with a committed service time. The inventory system consists of an online retailer and its customers, where the mean customer arrival rate is a linear decreasing function of the committed service time. The assumption of a committed service time means a generalization of the decision rule in Assåter (2007). The suggested decision rule uses real-time information, which is easy to obtain in practice by using modern information technology. Hence, this decision rule is easy to apply. Furthermore, a committed service time is quite common in online businesses, which means that our decision rule can be used in many practical cases.

The decision rule has been evaluated in a simulation study, and the results indicate that it performs well. The expected costs, when applying our decision rule to choose between backordering and emergency ordering, are significantly lower than the corresponding costs in the case of complete backordering or when we do
not allow backorders. The impact of the size of the committed service time has been evaluated in a numerical study. By considering different demand-related parameters, we conclude that our decision rule is most suitable when demand is low and customers are more time sensitive.

An advantage of the suggested rule is that it can be used in a quite general setting. We do not need to make any assumptions concerning the number of outstanding orders. However, it seems to be difficult to determine the optimal $Q$ and $R$ when using our decision rule. But by using the $Q$ and $R$ that are optimal in the case of complete backordering, we have a performance guarantee. That is, our decision rule will always lead to reduced expected costs.

Our work can be extended in several ways, for example, by incorporating a pricing strategy. Price is usually influenced by the length of the committed service time. It seems to be straightforward to incorporate pricing, as long as the unit price is a decreasing linear function of the length of the committed service time (Ray and Jewkes, 2004). That is, based on the assumption that the mean arrival rate depends linearly on the committed service time and on the unit price, and assuming that the unit price is a linear decreasing function of $T$, the resulting mean customer arrival rate is again a linear function of $T$. Thus, it seems our analytical procedures can be used in a similar way in this new situation. Our work can also be extended to the case where the retailer has a drop-shipping option, thus when facing shortage, we can use the decision rule to determine whether or not the supplier should ship directly to the customers.

A limitation of this work is that we only consider one product. This means, for example, that we cannot handle products with dependent demands. When demands for different products are not independent, for example, if they are substitutable or complementary, we need to extend our model, which may be relatively difficult. Similar problems may occur in case of capacity constraints and joint ordering costs. For the purpose of tractability, we do not take the supplier into consideration. Nevertheless, it would be a valuable extension to consider uncertainty in supply, including the supply lead time.

Acknowledgment

Sh. Huang and J. Chen's work on this paper was supported by the National Natural Science Foundation of China, under Grant Nos. 70890082 and 71072016.

References
