



Optimal control of a production–inventory system with customer impatience

Saif Benjaafar^{a,*}, Jean-Philippe Gayon^b, Seda Tepe^a

^a Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN 55455, USA

^b Laboratoire G-scop, Grenoble INP, 46 Avenue Félix Vialet, 38031 Grenoble Cedex, France

ARTICLE INFO

Article history:

Received 10 December 2009

Accepted 16 March 2010

Available online 1 April 2010

Keywords:

Production–inventory systems

Customer impatience

Optimal control

Make-to-stock queues

Markov decision processes

ABSTRACT

We consider the control of a production–inventory system with impatient customers. We show that the optimal policy can be described using two thresholds: a production base-stock level that determines when production takes place and an admission threshold that determines when orders should be accepted. We describe an algorithm for computing the performance of the system for any choice of base-stock level and admission threshold. In a numerical study, we compare the performance of the optimal policy against several other policies.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

Inventory problems treated in the literature fall mostly into two categories. One deals with systems where customers are assumed to be infinitely patient, that is, a customer whose order is backlogged is willing to wait for that order to be fulfilled no matter how long it takes. The other deals with systems where customers have zero patience, that is, a customer whose order cannot be fulfilled immediately is considered lost. However, in practice, it is more common for customers to be willing to wait, but only up to a point. Customers whose orders are backordered eventually cancel their orders and leave if their waiting time in backlog exceeds a certain *patience time*. This patience time usually varies from one customer to another and, for the same customer, may vary from one ordering instance to the next. Despite the prevalence of such behavior in practice, there is limited literature that deals with this issue. Consequently, very little is known about optimal control policies, or even effective heuristics, for such systems. Very little is also known about the impact of not accounting for customer impatience in making inventory decisions.

In this paper, we address some of these limitations in the context of a production–inventory system with a single product. In particular, we consider an $M/M/1$ make-to-stock queue with impatient customers. At any point in time, the system manager must decide on whether or not to produce and whether or not to accept an incoming order, should one arise. We show that the

optimal policy can be described using two thresholds: a production base-stock level and an admission threshold. Using the structure of the optimal policy, we model the dynamics of the corresponding production–inventory system as a Markov chain which allows us to compute efficiently the performance of the system for any choice of base-stock level and admission threshold. Using numerical results, we compare the performance of the optimal policy against several other policies and show that those that do not account for impatience can perform poorly.

In the existing inventory literature, the issue of customer impatience has been treated mostly in the context of so called inventory systems with *partial backordering*. Under partial backordering (see for example [14,13,18,15], and the references therein), an arriving customer who faces a stockout is backordered with a certain probability and is lost otherwise. In situations where multiple orders are placed at once, this means that a fraction of customers are backordered while the remainder are lost. These models capture the simplest case of customer impatience with a mixture of only two kinds of customers: some that are infinitely patient and, therefore, can be backordered, and some that have zero patience and, therefore, are lost if they cannot be fulfilled immediately. This obviously ignores the possibility of having customers who are willing to wait but with varying degrees of patience. Posner et al. [16] and Das [5] do consider systems where customers are initially willing to wait, but if their demand is not fulfilled within their patience time, they leave the system. However, in their case, they assume a particular inventory control policy, either a (q, r) or a base-stock policy, and do not allow for the possibility of rejecting customers. To our knowledge, our paper is the first to characterize the optimal policy for an inventory system with customer impatience.

* Corresponding author.

E-mail addresses: saif@umn.edu (S. Benjaafar), jean-philippe.gayon@grenoble-inp.fr (J.-P. Gayon), tepe007@umn.edu (S. Tepe).

Although the modeling of customer impatience is surprisingly limited in the inventory literature, there is significant and growing literature that models impatience in the context of queueing systems; see for example [6,7,12,8,9,1,19] and the references therein. A queueing system can be viewed as a make-to-order version of the system that we consider in this paper, where inventory is not allowed to be held in anticipation of future demand. Much of the queueing literature that incorporates impatience is focused on performance evaluation and not optimal control. Moreover, the optimal control problem in a queueing system is simpler as there is typically only a decision about whether or not to admit a customer.

The rest of the paper is organized as follows. In Section 2, we formulate the problem. In Section 3, we characterize the structure of the optimal policy. In Section 4, we describe a performance evaluation model. In Section 5, we present numerical results. In Section 6, we offer a summary and some concluding comments.

2. Problem formulation

We consider a system where a single product is produced at a single facility to fulfill demand from customers who place orders continuously over time according to a Poisson process with rate λ . Items are produced one unit at a time with exponentially distributed production times with mean $1/\mu$. The production facility can produce ahead of demand in a make-to-stock fashion. However, items in inventory incur a holding cost h per unit per unit time. Upon arrival, an order is either fulfilled from inventory, if any is available, backordered, or rejected. If an order is rejected, the system incurs a rejection cost r . If an order is backordered, the system incurs no immediate cost. However, customers are impatient and may decide to cancel their orders if their waiting time in backlog exceeds a patience time. If a customer cancels her order, the system incurs a cancellation cost c . We assume that the rejecting cost is smaller than the cancellation cost ($r \leq c$). Otherwise, it is optimal for the customers to accept all orders. The rejection cost can be viewed as a lost sale cost (e.g., the opportunity cost of generating revenue from the sale of one unit), while the cancellation cost can be viewed as the sum of a lost sale cost and a penalty for backlogging the order and not fulfilling it within the customer's patience time. We assume that there is no other cost to backordering, although it is possible to impose an additional cost that increases with the amount of time an order stays in backlog. Customer patience times are independent and exponentially distributed with mean $1/\gamma$. This means that customers are willing to wait for an amount of time that is exponentially distributed for their orders to be fulfilled; otherwise, they cancel their orders. The assumption of exponentially distributed patience times has been widely used in modelling customer impatience in queueing systems; see for example [7,12]. We assume that there is a finite upper bound M on the number of orders that can be on backorder at any time. This assumption, which is made for mathematical tractability, is however not restrictive as we allow this upper bound to be arbitrarily large.

At any point in time, the system manager must decide whether or not to produce an item. We assume that preemption is possible at no cost. This assumption is not restrictive since, as we show in Theorem 1, it turns out that, generally, it is not optimal to interrupt production of an item once it has been initiated. At any point in time, the system manager must also decide on how to handle incoming orders. In particular, should an order arise and there is no inventory on-hand, a decision must be made on whether to backorder it or to reject it.

The state of the system at time t can be described by net inventory $X(t)$, where $X(t)^+ = \max[0, X(t)]$ corresponds to on-hand inventory, and $X(t)^- = -\min[0, X(t)]$ to backorder level (the number of orders that are still waiting to be fulfilled). The

memoryless property allows us to formulate the problem as a Markov Decision Process (MDP) and to restrict our attention to the class of Markovian policies for which actions taken at a particular decision epoch depend only on the current state of the system. A policy d specifies for each state x whether production should be initiated or not and should an order arise, whether it should be fulfilled from on-hand inventory, backordered or rejected (if there is inventory on-hand, it is trivial to show that it is always optimal to fulfill it).

Let $v^d(x)$ denote the expected discounted cost (the sum of inventory holding, order cancellation and rejecting costs) over an infinite planning horizon obtained under a policy d and a starting state x . We denote by $\alpha > 0$ the discount rate. Our objective is to choose a policy d^* that minimizes the expected discounted cost over an infinite horizon. We refer to the optimal cost function as v^* where $v^* = v^{d^*}$. Following Lippman [11], we work with a uniformized version of the problem in which the transition rate in each state under any action is $\beta = \lambda + \mu + M\gamma$ so that the transition times between decision epochs form a sequence of i.i.d. exponential random variables, each with mean $1/\beta$. The introduction of the uniform transition rate allows us to transform the continuous time decision process into a discrete time decision process. Without loss of generality, we also rescale time by letting $\alpha + \beta = 1$. The optimal cost function can now be shown (see, e.g., [17]) to satisfy the following optimality equations:

$$v^*(x) = hx^+ + \lambda T_{arr} v^*(x) + \mu T_{prod} v^*(x) + \gamma T_{imp} v^*(x),$$

where the operators T_{prod} , T_{arr} , and T_{imp} are defined as follows:

$$T_{prod} v(x) = \min(v(x), v(x+1)),$$

$$T_{arr} v(x) = \begin{cases} \min(v(x-1), v(x)+r) & \text{if } x > -M \\ v(x)+r, & \text{if } x = -M, \text{ and} \end{cases}$$

$$T_{imp} v(x) = \begin{cases} -x[v(x+1)+c] + (M+x)v(x) & \text{if } -M \leq x \leq -1 \\ Mv(x) & \text{if } x \geq 0. \end{cases}$$

Operator T_{prod} is associated with the production decision. Operator T_{arr} is associated with the handling of the arrival of an order. Note that when the backorder level reaches M an incoming order is always rejected and the cost r is incurred. Operator T_{imp} is associated with customers canceling their orders due to impatience.

3. The structure of the optimal policy

In this section, we characterize the structure of the optimal policy. In order to do so, we show that the optimal value function $v^*(x)$ for all states x satisfies certain properties as specified in Definition 1 below. We then show that these properties imply a specific rule for the optimal action in each state.

Definition 1. Let \mathcal{U} be a set of real valued functions defined on the set of integers \mathbf{Z} , such that if $v \in \mathcal{U}$, then:

Property P1 $\Delta v(x) \geq -c$, for all x ,

Property P2 $\Delta^2 v(x) \geq 0$, for all x ,

Property P3 $\Delta v(x) \geq -r$, for all $x \geq 0$, and

Property P4 $\Delta v(x) \leq 0$, for all $x < 0$,

where $\Delta v(x) = v(x+1) - v(x)$ and $\Delta^2 v(x) = \Delta v(x+1) - \Delta v(x)$. Therefore, convexity of $v(x)$ is equivalent to $\Delta^2 v(x) \geq 0$.

Lemma 1. If $v \in \mathcal{U}$, then $Tv \in \mathcal{U}$ where $Tv(x) = hx^+ + \lambda T_{arr} v(x) + \mu T_{prod} v(x) + \gamma T_{imp} v(x)$. Furthermore, the optimal cost function $v^* \in \mathcal{U}$.

The detailed proof of Lemma 1 can be found in [2]. In this proof, we show that the operator T preserves properties P1–P4; see [10,4]

for a unified treatment of similar proofs. As any sequence of value functions $v_{n+1} = Tv_n$ converges to the optimal value function v^* (see, e.g., [17]), we conclude that v^* satisfies properties P1–P4.

In order to describe the optimal policy implied by the above properties of the value function, we first define the following two threshold parameters:

$$s^* = \min(x : \Delta v^*(x) \geq 0),$$

and

$$w^* = \max\{-M, \min(x : \Delta v^*(x) + r \geq 0)\}.$$

Theorem 1. *There exists an optimal policy that can be specified using thresholds s^* and w^* as follows. The optimal production policy is a base-stock policy with base-stock level s^* , such that it is optimal to produce if $x < s^*$ and not to produce otherwise. The optimal order fulfillment policy is a limited admission policy with admission threshold w^* , such that it is optimal to accept an order if $x > w^*$ and to reject it otherwise. An admitted order is fulfilled from on-hand inventory if there is any and is backordered otherwise. Moreover, we have the following:*

- It is always optimal to produce if there are any backorders; that is, $s^* \geq 0$.
- It is always optimal to accept orders if there is on-hand inventory; that is, $w^* \leq 0$.
- If $s^* > 0$, then it is never optimal to preempt production once it has been initiated.

Proof. From Lemma 1, we know that $v^* \in \mathcal{U}$. Property P2 guarantees the existence of threshold levels s^* and w^* . Furthermore it implies that it is optimal to produce if $x < s^*$ and not to produce otherwise and to accept an order if $x > w^*$ and to reject it otherwise. Property P3 states that it is optimal to accept orders if there is on-hand inventory and implies that $w^* \leq 0$. Property P4 states that it is optimal to produce if $x < 0$ and implies that $s^* \geq 0$. Finally, note that, as long as $s^* > 0$, it is never optimal to preempt production once it is initiated. In particular, if it is optimal to produce in state $0 \leq x < s^*$, then it continues, by virtue of the fact that the production policy is a base-stock policy with base-stock level s^* , to be optimal to produce in state $x - 1$ if an order arrives and we decide not to reject it (of course it continues to be optimal if an order arrives and we decide to reject it, leaving the system in state x). Similarly, if it is optimal to produce in state $w^* \leq x < 0$, then it continues to be optimal to produce in both state $x + 1$, corresponding to an order cancellation, and state $x - 1$, corresponding to an order arrival. The only scenario under which preemption is possible is when $s^* = 0$ (i.e., inventory is never held and we produce only if there is a backorder). However, even in this case, preemption is optimal only if the system is in state $x = -1$ and an order is cancelled, moving the system to state $x = 0$. \square

In contrast to common pure lost sales and pure backorder policies, the optimal policy allows for both backordering and order rejection. In doing so, the policy limits both inventory and backorder levels (inventory is costly because of holding costs and backordering is costly because it increases the chance of customers canceling their orders due to impatience). The structure of the optimal policy in Theorem 1 can be shown to continue to hold for several variants of the problem, including systems where there is a linear production cost and a convex holding cost. It also continues to hold in the case where the optimization criterion is the average cost per unit time instead of the expected discounted cost. The existence of an optimal policy for the average cost, and for this average cost to be finite and independent of the starting state, can be proven via an argument involving taking the limit as $\alpha \rightarrow 0$ in the discounted cost problem (see for example [3,20]).

In order to show monotonicity properties for s^* and w^* with respect to various system parameters, we compare the optimal

value functions of two systems that are identical except for the value of one system parameter, denoted by p ; see [4] for a unified treatment of similar proofs. For short, we write $p = \lambda$ when demand rate is varied, $p = r$ when rejecting cost is varied and so on. The optimal base-stock level, admission threshold and value function corresponding to a given system parameter p will be represented by s_p^* , w_p^* and $v_p^*(x)$ respectively, where p belongs to the set of system parameters $\{\lambda, \mu, h, c, r, M\}$.

We state that a function v_p is submodular in x and p (denoted by $SubM(x, p)$), if and only if

$$\Delta v_p(x) \geq \Delta v_{p+\epsilon}(x), \quad \forall x \geq -M, \forall p \neq M, \forall \epsilon \geq 0.$$

The supermodularity in x and p , denoted by $SuperM(x, p)$, is the opposite inequality ($\Delta v_p(x) \leq \Delta v_{p+\epsilon}(x)$). These definitions can be used when $p \in \{r, c, h, \lambda, \mu\}$. However, when $p = M$, p is discrete and the state space depends on M . In this case, we state that v is $SubM(x, M)$ if and only if the following inequality holds:

$$\Delta v_M(x) \geq \Delta v_{M+1}(x), \quad \forall x \geq -M, \forall M \in \mathbb{N}.$$

$SuperM(x, M)$ is the same inequality in the opposite direction.

Next, we define \mathcal{V} as a set of real valued functions with the following properties:

Definition 2. If $v \in \mathcal{V}$, then:

Property Q1 $v \in \mathcal{U}$,

Property Q2 $\forall p \in \{\mu, h, M\}$, v is $SuperM(x, p)$ and $\forall p \in \{\lambda, r, c\}$, v is $SubM(x, p)$,

Property Q3 $\Delta v_{r+\epsilon}(x) + \epsilon \geq \Delta v_r(x)$, $\forall r \geq 0, \forall \epsilon \geq 0$.

If we prove that $v^* \in \mathcal{V}$, then, we obtain the monotonicity results for s^* and w^* as described in Theorem 2.

The uniformization rate depends on $\{\lambda, \mu, M\}$ and needs to be constant for two systems to be comparable. We rescale the time using a uniformization rate δ which is sufficiently larger than $(\alpha + \lambda + \mu + M\gamma)$ to have the same uniformization rate with parameter values p or $p + \epsilon$. Therefore, the optimality equations are redefined by adding a new operator, T_{unif} , to maintain a constant uniformization rate. T_{unif} is a fictitious event operator that transfers the system into the same state. Optimality equations can be rewritten as follows:

$$v_p^*(x) = Tv_p^*(x), \quad \forall x,$$

with

$$Tv_p(x) = \frac{1}{\delta} [hx^+ + \lambda T_{arr} v_p(x) + \mu T_{prod} v_p(x) + \gamma T_{imp} v_p(x) + T_{unif} v_p(x)],$$

and

$$T_{unif} v_p(x) = (\delta - \alpha - \lambda - \mu - M\gamma) v_p(x).$$

Operators T_{arr} , T_{prod} and T_{imp} are defined as previously. Note that all operators may depend on p . For instance, if $p = \lambda$, we have the following optimal operator T when the arrival rate equals $\lambda + \epsilon$:

$$Tv_{\lambda+\epsilon}(x) = \frac{1}{\delta} [hx^+ + (\lambda + \epsilon)T_{arr} v_{\lambda+\epsilon}(x) + \mu T_{prod} v_{\lambda+\epsilon}(x) + \gamma T_{imp} v_{\lambda+\epsilon}(x) + (\delta - \alpha - \lambda - \epsilon - \mu - M\gamma) v_{\lambda+\epsilon}(x)].$$

In order to prove that T preserves properties of the set \mathcal{V} , we make use of the following two lemmas (proofs can be found in [2]). In Lemma 2, we show that individual operators T_{prod} , T_{imp} , T_{arr} , T_{unif} preserve some modular properties.

Lemma 2. *If $v \in \mathcal{V}$, then the preservation of submodularity and supermodularity by the operators is as given in Table 1.*

Table 1 can be interpreted as follows. For instance, if $v \in \mathcal{V}$ then $T_{imp}v$ is $SuperM(x, p)$ for all $p \in \{\mu, h\}$ and $SubM(x, p)$ for all $p \in \{\lambda, c, r\}$.

Table 1
Preservation of submodularity and supermodularity by the operators.

	SuperM(x, p)	SubM(x, p)
hx^+	$\forall p \in \{\mu, h, M\}$	$\forall p \in \{\lambda, c, r\}$
$T_{prod}v$	$\forall p \in \{\mu, h, M\}$	$\forall p \in \{\lambda, c, r\}$
$T_{imp}v$	$\forall p \in \{\mu, h\}$	$\forall p \in \{\lambda, c, r\}$
$T_{arr}v$	$\forall p \in \{\mu, h, M\}$	$\forall p \in \{\lambda, c, r\}$
$T_{unif}v$	$\forall p \in \{\mu, h\}$	$\forall p \in \{\lambda, c, r\}$

Lemma 3 establishes additional results necessary to prove that T preserves $SuperM(x, \mu)$ and $SubM(x, \lambda)$. We have suppressed p from the notation of the value function since these results hold independently of p .

Lemma 3. If $v \in \mathcal{U}$, then

- $\Delta[T_{prod}v(x) - v(x)] \geq 0$, for all $x \geq -M$, and
- $\Delta[T_{arr}v(x) - v(x)] \leq 0$, for all $x \geq -M$.

Using **Lemmas 2** and **3**, we can then show (see [2]) that T preserves properties of \mathcal{V} (that is, if $v \in \mathcal{V}$ then $Tv \in \mathcal{V}$). In turn, this implies, by value iteration, that v^* belongs to \mathcal{V} and thus the results in **Theorem 2**.

Theorem 2. The optimal base-stock level s^* is non-increasing in h, μ , and M and is non-decreasing in c, r , and λ . The optimal admission threshold w^* is non-increasing in h, μ, r , and M and is non-decreasing in c and λ .

Theorem 2 also applies to the average-cost case. Numerical results obtained (for details see [2]) show that both s^* and w^* can be quite sensitive to changes in system parameter values. For example, when $c \rightarrow r, w^* \rightarrow -\infty$ and it becomes optimal to never reject orders. When $c \rightarrow \infty, w^* \rightarrow 0$ and it becomes optimal to always reject orders when there is no inventory on-hand. When λ is much larger than μ , it also becomes optimal to always reject when there is no on-hand inventory. In this case, there is not sufficient capacity to fulfill demand and a fraction of total demand must always be rejected. Note that we were not able to establish monotonicity results with respect to the impatience parameter γ . In Section 5, we present numerical results regarding the effect of γ and examine the impact of customer impatience on system performance.

4. A performance evaluation model

In this section, we use knowledge of the structure of the optimal policy to construct a performance evaluation model for computing efficiently the optimal base-stock level and the optimal admission threshold under the average-cost criterion. Having such a model eliminates the need to use dynamic programming to carry out computations. Moreover, a dynamic programming algorithm, depending on problem parameter values, may require truncation of the state space, making the corresponding results approximate.

The approach that we take follows from the recognition that a system operating under a control policy specified by a fixed base-stock level s and an admission threshold w can be modeled as a Markov chain. In particular, the net inventory level, $X(t)$, evolves as a continuous time Markov chain with rates of transition from state j to state k, q_{jk} , given by

$$q_{jk} = \begin{cases} \lambda & \text{if } k = j - 1, w < j \leq s, \\ \mu + \gamma j^- & \text{if } k = j + 1, w \leq j < s, \\ 0 & \text{otherwise,} \end{cases}$$

where $j^- = -\min[0, j]$. Define $\rho = \frac{\lambda}{\mu}$. The stationary probabilities $\pi_j = \lim_{t \rightarrow \infty} P(X(t) = j)$ can be shown to be given by the

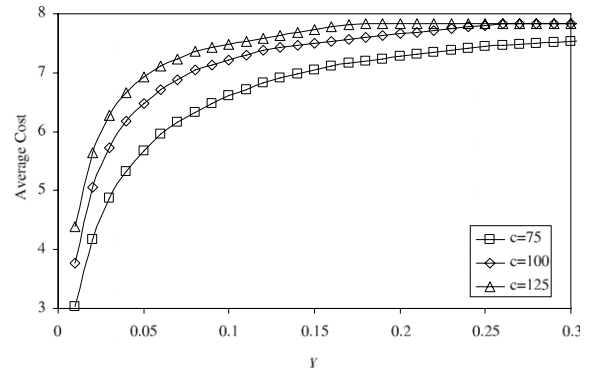


Fig. 1. Impact of impatience rate on the average cost when $\lambda = 0.9, \mu = 1, r = 50$, and $h = 1$.

following:

$$\pi_j = \begin{cases} \rho^{s-j} \pi_s & \text{if } 0 \leq j \leq s, \\ \left(\prod_{k=1}^{-j} \frac{\lambda}{\mu + \gamma k} \right) \rho^s \pi_s & \text{if } w \leq j < 0, \\ 0 & \text{otherwise,} \end{cases}$$

$$\pi_s = \left[\sum_{j=0}^s \rho^j + \sum_{j=w}^{-1} \left(\prod_{k=1}^{-j} \frac{\lambda}{\mu + \gamma k} \right) \rho^s \right]^{-1}.$$

For given s and w , the expected cost, which we denote by $V(s, w)$, can now be obtained as

$$\begin{aligned} V(s, w) &= hE(X^+) + c\gamma E(X^-) + r\lambda\pi_w \\ &= h \sum_{j=1}^s j\pi_j + c\gamma \sum_{j=w}^{-1} -j\pi_j + r\lambda\pi_w. \end{aligned}$$

The above expression involves the sum of finite terms and, therefore, can be computed efficiently. The optimal values for s and w can be obtained via an exhaustive search over a large enough range of s and w (unfortunately, the function $V(s, w)$ is not jointly convex in s and w). The computational effort for carrying out this search is generally modest. For example, a search over a 1000 by 1000 grid takes only few seconds on a standard personal computer. The computations can be further expedited by noting that the optimal base-stock level, s^* , has an upper bound given by the optimal base-stock level, \hat{s}^* , of a system with $M = 0$ where items are always rejected when they cannot be fulfilled from on-hand inventory. In the following theorem, we show how an upper bound on \hat{s}^* and, therefore, also on s^* can be obtained in closed form (for a proof see [2]).

Theorem 3.

$$0 \leq s^* \leq s_u \quad \text{with } s_u = \begin{cases} \sqrt{2\lambda r/h} & \text{if } \rho \leq 1 \\ (\rho - 1)/(h'\rho) + 1/\ln \rho & \text{if } \rho > 1 \end{cases}$$

and $h' = h/(\lambda r)$.

5. Some numerical results

In this section, we briefly provide some numerical results that illustrate the impact of customer impatience on optimal average cost and that examine the sensitivity of the base-stock level and admission threshold to system parameters. We also provide numerical results that compare the optimal policy to other commonly used policies.

In **Fig. 1**, we present results that show the impact of varying the patience time parameter γ on optimal average cost. As we can see, customer impatience can have a significant impact on the cost. Cost is increasing in a roughly concave fashion in the impatience parameter, γ . These results highlight the importance of carefully accounting for customer impatience, as underestimating

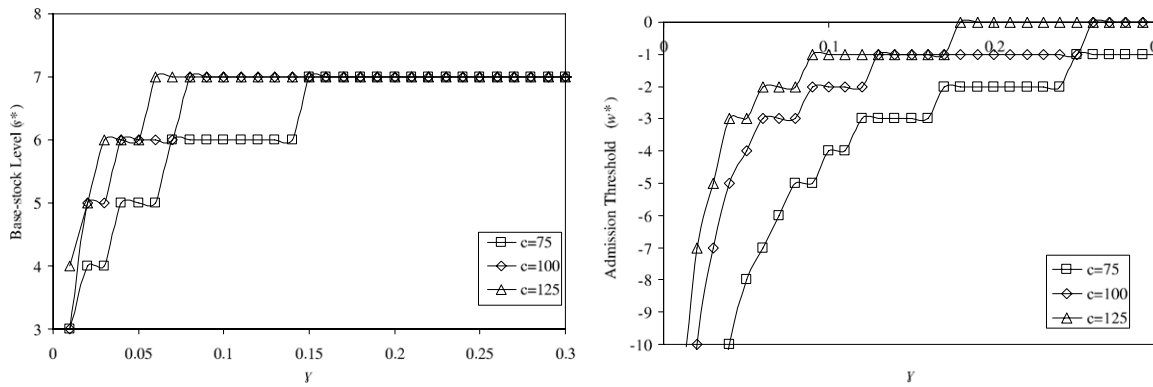


Fig. 2. Impact of impatience rate on s^* and w^* when $\lambda = 0.9$, $\mu = 1$, $r = 50$, and $h = 1$.

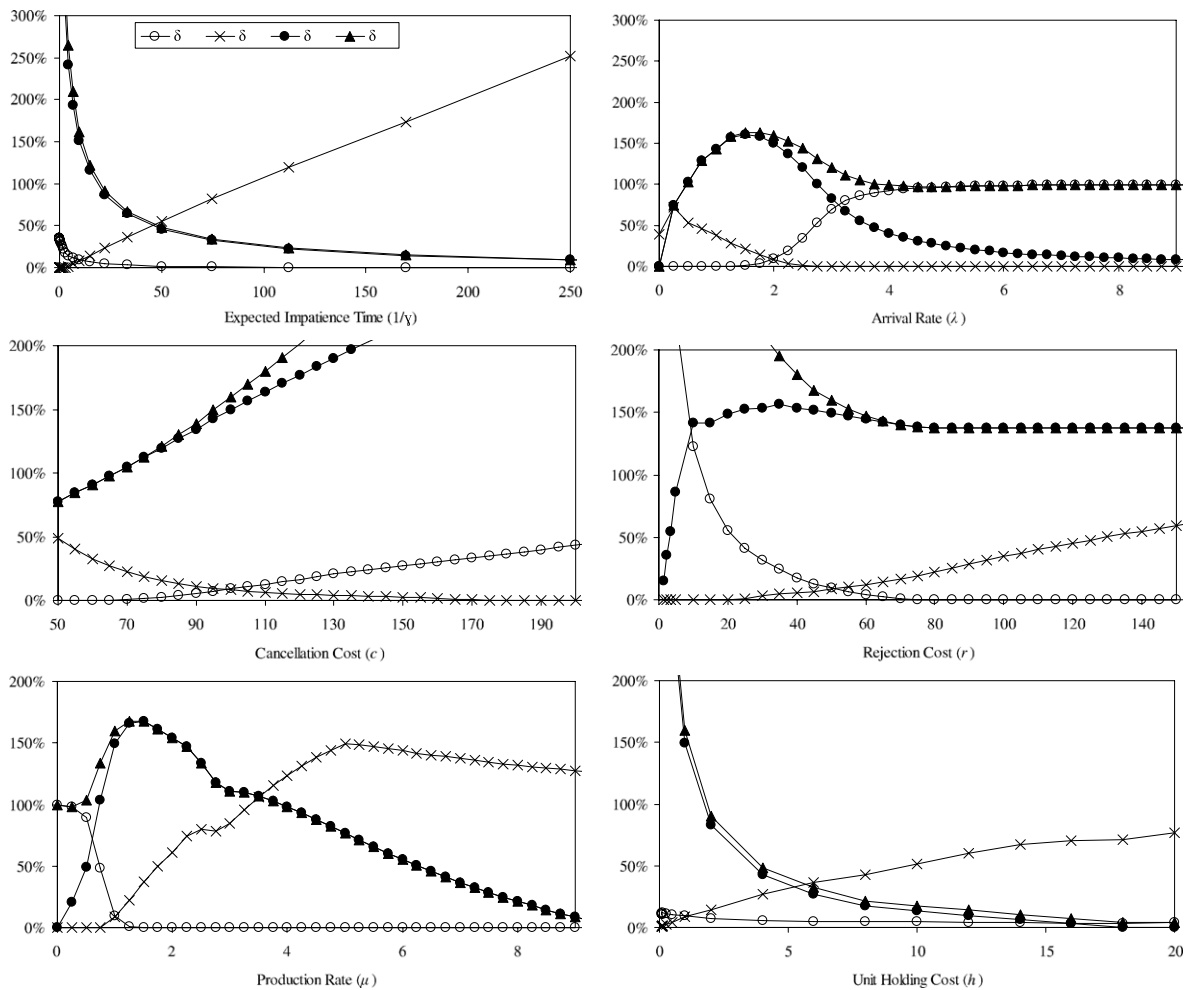


Fig. 3. Impact of system parameters on δ_i . Unless they are being varied, the following parameter values are used: $\gamma = 0.1$, $\lambda = 0.9$, $\mu = 1$, $c = 100$, $r = 50$, and $h = 1$.

or overestimating customers' willingness to wait can lead to significantly underestimating or overestimating the true cost. They can also lead to significant errors in selecting values for the base-stock level and the admission thresholds (Fig. 2).

In Fig. 3, we compare the performance of the optimal policy with four other policies, that are perhaps simpler to implement as they all involve a single control parameter, but that either ignore or do not fully account for the impact of customer impatience. The policies are as follows:

Policy H_1 : Orders are never rejected and are always backordered. Production is managed according to a base-stock policy with a fixed base-stock level.

Policy H_2 : Orders that cannot be fulfilled from on-hand inventory are always rejected. Production is managed according to a base-stock policy with a fixed base-stock level.

Policy H_3 : No inventory is held in anticipation of future demand and orders are always backordered as long as the backorder level does not exceed a specified threshold.

Policy H_4 : No inventory is held in anticipation of future demand and orders are always backordered when they arrive.

The above policies can all be viewed as special cases of an (s, w) policy, where production is managed using a base-stock policy with base-stock level s and order fulfillment is managed using an admission policy with an admission threshold w . In the case of

H_1 , $s \geq 0$ and $w = -\infty$; for H_2 , $s \geq 0$ and $w = 0$; for H_3 , $s = 0$ and $w \leq 0$; and for H_4 , $s = 0$ and $w = -\infty$. To allow for a fair comparison against the optimal policy, the parameters of the four policies are always chosen optimally. Fig. 3 shows the percentage cost difference between the cost of the optimal policy and the optimal cost of each policy. The percentage cost difference, δ_i , for policy H_i is computed as $\delta_i = (C_i^* - C^*)/C^* \times 100\%$, where C_i^* is the optimal average cost under policy H_i and C^* is the average cost under the optimal policy.

As we can see from Fig. 3, all four policies can perform poorly. In general, policies that do not allow for rejection ($w = -\infty$) perform poorly when customers are very impatient (γ is high), cancellation cost is high (c is high), rejection cost is low (r is low), or when the utilization of the production facility is high (the ratio λ/μ is high). On the other hand, policies that always reject orders when they cannot be fulfilled from on-hand inventory ($w = 0$) perform poorly when customers are patient (γ is low), impatience cost is low (c is low), rejection cost is high (r is high), or when the utilization of the production facility is low (the ratio λ/μ is low). There are of course settings where each of the four policies performs reasonably well. However, in most settings when impatience matters, either because of a low customer patience time or a high cost of cancellation, there are significant benefits to using the optimal policy or, to a lesser degree, a policy that limits the number of backorders, such as policy H_2 or H_3 .

6. Conclusions and future research

The results of the paper highlight the importance of incorporating customer impatience in the management of inventory systems. The results also highlight the inadequacy of existing inventory models which tend to assume that orders that cannot be immediately fulfilled from on-hand inventory are either all backordered (pure backorder systems) or all rejected (pure lost sales systems) and illustrates the need for models that allow for both backordering and rejection.

This paper is obviously only a first step toward a more comprehensive modeling and analysis of inventory systems with impatience. Avenues for future research are many. It will be useful to consider systems with different demand, production time, and patience time distributions. For example, it is possible to substitute the exponential distribution by phase-type distributions which can be constructed to approximate other more general distributions. Phase-type distributions retain the Markovian property of the system and continue to allow the formulation of the problem as an MDP. It will also be useful to extend the analysis to systems with multiple demand classes with different patience time parameters

and different rejection and cancellation costs. This would give rise to new kinds of decisions regarding how on-hand inventory should be allocated and how fulfillment priorities should be assigned to orders that are in backlog.

References

- [1] M. Armony, E. Plambeck, S. Seshadri, Sensitivity of optimal capacity to customer impatience in an unobservable $M/M/S$ queue (Why you shouldn't shout at the DMV), *Manufacturing and Service Operations Management* 11 (1) (2009).
- [2] S. Benjaafar, J.P. Gayon, S. Tepe, Optimal control of a production–inventory system with customer impatience, Working paper, University of Minnesota, 2009. www.ie.umn.edu/faculty/faculty/Benjaafar.shtml.
- [3] R. Cavazos-Cadena, L. Sennott, Comparing recent assumptions for the existence of average optimal stationary policies, *Operations Research Letters* 11 (1) (1992) 33–37.
- [4] E.B. Çil, E.L. Örmeci, F. Karaesmen, Effects of system parameters on the optimal policy structure in a class of queueing control problems, *Queueing Systems* 61 (4) (2009) 273–304.
- [5] C. Das, The $(S - 1, S)$ inventory model under time limit on backorders, *Operations Research* 25 (1977) 835–850.
- [6] N. Gans, G. Koole, A. Mandelbaum, Telephone call centers: tutorial, review, and research prospects, *Manufacturing and Service Operations Management* 5 (2) (2003) 79–141.
- [7] O. Garnet, A. Mandelbaum, M. Reiman, Designing a call center with impatient customers, *Manufacturing & Service Operations Management* 4 (3) (2002) 208–227.
- [8] O. Jouini, Y. Dallery, Analysis of a multiple priority queue with impatient customers, Working Paper, Ecole Centrale Paris, 2007.
- [9] O. Jouini, A. Pot, Y. Dallery, G. Koole, Real-time dynamic scheduling policies for multiclass call centers with impatient customers, Working Paper, Ecole Centrale Paris, 2007.
- [10] G. Koole, Structural results for the control of queueing systems using event-based dynamic programming, *Queueing Systems* 30 (3) (1998) 323–339.
- [11] S. Lippman, Applying a new device in the optimization of exponential queueing systems, *Operations Research* 23 (4) (1975) 687–710.
- [12] A. Mandelbaum, S. Zeltyn, The Palm/Erlang-a queue, with applications to call centers, Working Paper, Tel Aviv University, 2005.
- [13] K. Moinzadeh, Operating characteristics of the $(S - 1, S)$ inventory system with partial backorders and constant resupply times, *Management Science* (1989) 472–477.
- [14] D.C. Montgomery, M.S. Bazaraa, A.K. Keswani, Inventory models with a mixture of backorders and lost sales, *Naval Research Logistics Quarterly* 20 (2) (1973).
- [15] S. Nahmias, S.A. Smith, Optimizing inventory levels in a two-echelon retailer system with partial lost sales, *Management Science* (1994) 582–596.
- [16] M.J.M. Posner, B. Yansouni, A class of inventory models with customer impatience, *Naval Research Logistics Quarterly* 19 (3) (1972).
- [17] M. Puterman, *Markov Decision Processes*, John Wiley and Sons Inc., New York, 1994.
- [18] E. Smeitink, A Note on “Operating characteristics of the $(S - 1, S)$ inventory system with partial backorders and constant resupply times”, *Management Science* (1990) 1413–1414.
- [19] A.R. Ward, S. Kumar, Asymptotically optimal admission control of a queue with impatient customers, *Mathematics of Operations Research* 33 (1) (2008) 167–202.
- [20] R.R. Weber, S. Stidham Jr., Optimal control of service rates in networks of queues, *Advances in Applied Probability* 19 (1) (1987) 202–218.