Fair and Profitable: How Pricing and Lead-Time Quotation Policies Can Help

Barış Balcıoğlu, Yağız Varol,
Sabancı University, Faculty of Engineering and Natural Sciences,
Orhanlı-Tuzla, 34956 Istanbul, Turkey,
balcioglu@sabanciuniv.edu, yvarol@sabanciuniv.edu

Abstract

In this paper, we propose four policies to serve price and lead-time sensitive customers with a single type of product produced in an $M/GI/1$ type make-to-stock queueing system. The policies are developed to observe certain principles of fairness: if a customer is quoted a longer lead-time, she must be charged a lower price and a significant proportion of the deliveries have to be made during the quoted lead times. Although handling non-exponential service times in this setting presents difficulties, our analysis of the proposed policies is exact. Two of the policies operate with static prices, while one of the two dynamic pricing policies also quotes dynamic lead times. By construction of the policies, we show that the dynamic pricing policies are more profitable. Numerical examples bring additional light by showing that in small markets with oversensitive customers, dynamic policies can be profitable while static pricing policies can fail. In a larger market, a simple dynamic policy charging two prices depending on the stock availability can be a reasonable compromise. Dynamic policies tend to suffer less against high production time variability as well.

Keywords and Phrases: queueing; make-to-stock queues; pricing; lead-time quotation; service time variability

1 Introduction

In markets that are getting more competitive everyday, in order to increase profitability, companies are obliged to make offers that should adjust to changing circumstances. Dynamic pricing and lead-time quotation are well-established responsive tools to this end (see e.g., Webster, 2002, Çelik and Maglaras, 2008, Zhao, Stecke, and Prasad, 2012). Dynamic pricing has a long and rich history of application in the airline or hotel industries, or for managing inventory of items sold over a finite time horizon (see, e.g., Bitran and Caldentey, 2003, for an overview). Employing dynamic pricing policies for items that are replenished regularly is also not new (see reviews by Yano and Gilbert, 2002 and Elmaghraby and Keskinocak, 2003). While in a setting with non-renewable capacity customers would find it rational that different prices be charged at different points in time over the finite horizon, in a setting with renewable goods, customers would prefer companies that treat them fairly such as by charging customers in identical situation the same price or offering lower prices when service level gets lower for them.

In this paper, our goal is to characterize fair dynamic pricing policies and contrast them with fair static pricing policies. This helps us explore which policy is more profitable and practical to apply with different market and customer profiles. To do this, we consider a producer of a single type of item and model its production and inventory system by a make-to-stock queue where demand is generated by a Poisson arrival process that has state-dependent rates. The company can vary the price depending on the amount of work-in progress and, if backlogging is considered, can announce a lead-time as well. The demand rate is assumed to be a function of the offered price and – when quoted – the lead-time. We assume that the company pays tardiness costs when the delivery occurs after the quoted lead-time (see, Savaşaneril, Griffin, and Keskinocak, 2010, and the references therein for examples where this cost is incurred). Yet, one expects reliable quotes from a fair company so that in the long run a high proportion of deliveries is made within these quoted lead times. To consider the impact of production time variability, we assume that production times have general distributions. This gives us $M_n/GI/1$ type make-to-stock queues as the underlying building blocks. Our major contribution is to design four policies, two of them operating with static prices, and the other two varying prices dynamically. One policy in the latter group quotes lead times dynamically as well. Characterizing the optimal dynamic pricing and lead-time quotation policy in this setting is not tractable as discussed in Section 3. With numerical examples we compare these four policies and we are able to show that in smaller markets, charging prices and quoting lead times dynamically should be considered, whereas in larger markets a simpler policy charging a high/low price when there is stock/no stock can be employed as a reasonable compromise. Dynamic policies could be the only way to make profit when customers are very sensitive to price and delay. We also glean from the numerical examples that being fair does not decrease profitability significantly and the dynamic policies are, on average, more resilient against the worsening impact of the production time variability.

In the literature, Naor (1969), Mendelson (1985), and Dewan and Mendelson (1990) are examples to those who obtain the optimal static price in delay systems to control arrival rates. Low (1974) designs a dynamic pricing scheme by choosing optimal prices from a given set of prices. He shows that the optimal price to charge is nondecreasing in the number of customers in the queueing system. In make-to-stock queues using dynamic pricing schemes, in case of lost sales, Li (1988) assumes that the customer arrival/demand rate is a continuous function of the price and shows that the base-stock policy is optimal, and the optimal sequence of prices is nonincreasing in inventory level. Gayon et al. (2009) extend this model letting demand depend on the state of the external environment. They observe that if demand does not depend on the state of the external environment, the optimal dynamic pricing policy results in modest improvements against optimal static pricing. In our study, if lost sales is considered, all customers arriving when there is stock need to be charged the same price in order to be fair. The decision variables for this static pricing policy are the optimal base-stock to keep and the optimal price to charge making use of the function that relates demand rate to price.

When backlogging is permitted, Chen, Feng, and Ou (2006) in the $M/M/1$ queue, and Chen, Chen, and Pang (2011) in the $M/E_k/1$ queue employ a Markov decision process approach and show that the base-stock policy together with a price-switch policy is optimal. When the number of production orders surpasses a higher threshold a higher price is charged. This way, the customer arrival rate and the delay related penalty costs incurred get lower. Chen and Frank (2001) provide similar discussions on how listing prices non-decreasing in the length of production order queue helps lower the frequency of penalty payments. These policies are not applicable in the problem we study: if a customer arriving when the order queue is longer ends up waiting for delivery, she should be charged a lower price to compensate for the lower service level (since she has to wait) than a customer that is served immediately when there is stock. This idea is frequently used in the construction industry. A customer pays less if she is purchasing a house under construction or in the planning stage than another customer who buys a brand-new home in stock. Or a customer may have to pay more if she wants an expedited delivery in other settings. In cases of backlogging, the policies we propose quote a reliable lead-time estimate. The length of the lead-time, which may depend on the order queue length, is what lowers the demand rate when the inventory is out of stock.

The idea of using sojourn time distribution of an order in the $M/M/1$ queue to quote the lead-time is proposed by Dellaert (1991). We make use of the distribution of the sojourn time in the $M/GI/1$ setting for two policies that quote the same lead-time to all backlogged customers. Duenyas and Hopp (1995) show that the optimal lead-time to quote in the $M/M/1$ make-to-order system increases in the order queue size. Savaşaneril, Griffin, and Keskinocak consider dynamic lead-time quotation allowing order rejection when the number of pending orders gets critical in an $M/M/1$ make-to-stock queue. They show that the optimal quoted lead-time increases in the number of pending orders. Kahvecioğlu and Balcıoğlu (2016) adapt their problem setting to the $M/GI/1$ queue and propose two policies of dynamic lead-time quotation. In both studies, the price is fixed and different customer responses to the quoted-lead times change the arrival rate. Companies are not assumed to provide reliable lead-time estimates: Implicitly assuming that the customers are fine with this, companies can quote zero for lead-time if it is more profitable to sell than paying penalty costs. In our study, whether a single lead-time is quoted for all backlogged customers or the quotation is done depending on the order queue size that the arrival sees, we stipulate that a certain proportion of deliveries need be made within the quoted lead times.

Palaka, Erlebacher, and Kropp (1998) assume that the Poisson demand decreases linearly in price and lead-time in an $M/M/1$ type make-to-order queue. They design their problem to find the optimal static price and the single lead-time to announce while satisfying that a desired proportion of deliveries be made within the lead-time. While using the same linear demand function, Ray and Jewkes (2004) also consider that the price is dependent on the lead-time to capture the fact that a higher price be charged for shorter lead times. They, then, solve for the optimal lead-time to announce everyone, which in return, yields the optimal static price and the production capacity as well. Hammami, Frein, and Albana (2020) introduce the concept of rejecting customers when the number of orders reaches a critical level while optimizing the static price and lead-time in a make-to-order setting. In our study, we do not consider production capacity as a decision variable while in the literature, there are studies such as Webster, Boyacı and Ray (2003), Pekgün, Griffin, and Keskinocak (2008), Zhao, Stecke, and Prasad, that consider it to respond to market changes or meet the lead-time service constraint. Pekgün, Griffin, and Keskinocak show the inefficiencies when pricing and lead-time quotation decisions are made by separate departments. Boyacı and Ray consider a regular and express delivery option for a product and study how prices for both delivery types and lead-time for the express delivery option are determined. All these studies use the linear demand model as a model assumption, whereas we do not have that restriction although in our numerical examples we employ the linear model due to its popularity in the literature. There are also studies that consider different products allowing non-exponential distributions for the interarrival and service times. Çelik and Maglaras study a multi-class $M_n/GI/1$ type make-to-order queue allowing orders to be expedited to meet the lead times. Feng and Zhang (2017) assume renewal demand and Phase-type service times in a make-to-order setting and explore pricing and lead-time quotation considering customer differences.

Zhao, Stecke, and Prasad compare two modes, one offering a single price and lead-time, and another mode with a menu of lead times and prices and observe that customer and production characteristics are influential to decide on which mode should be preferred.

In our study, we consider a single type of customer whose demand for a single type of item is sensitive to both price and delay. There is a fixed unit time cost that is assumed to capture all the facility running, machinery maintaining, labor, and raw material inventory costs. This cost, together with the inventory holding cost for finished items and the penalty cost for late deliveries, makes the major contribution to the possibility that the proposed policies can turn unprofitable when revenues fall short of costs. All the proposed policies operate in the steady-state, thus, we need to compute the steady-state distribution of the number of pending orders in the $M_n/GI/1/$ queue for which we employ the algorithm by Yang (1994) for its speed. We refer the interested reader also to Abouee-Mehrizi and Baron (2016) and Economou and Manou (2015) for other alternatives. According to the fairness principles outlined in Section 2, if a static price is charged and inventory is kept, one should operate in a make-to-stock regime with lost sales as we do in Section 3.1. A make-to-order regime with static price studied in Section 3.2 stipulates a single lead-time to be quoted everyone. The two dynamic pricing policies operate on two arrival rates, one for when there is stock, the other for when customers are backlogged. The simplest dynamic pricing policy, in Section 3.3, announces one lead-time to all backlogged customers and charges two prices: one to customers who are served directly from stock, a lower price to the backlogged customers. A more refined policy, in Section 3.4, announces the lead-time based on the number of orders a backlogged customer sees upon arrival. This policy also limits the total number of backlogged customers to increase the profit. The optimization for the base-stock level, the maximum number of backlogged customers, and prices is performed by basically searching over arrival rate(s). With the arrival rate(s) in hand, by numerically inverting the Laplace transform (LT) of the sojourn time of an order, we first search for the lead-time(s) that satisfy a given probability of delivery during the quoted lead-time(s). With the lead-time(s) determined, one obtains the corresponding price(s).

The numerical examples presented in Section 4 following this optimization procedure show the superiority of the refined dynamic pricing policy, especially in a smaller market with customers very sensitive to price and delay. These cases could be examples for when a company can survive only with dynamic policies. In fact, we see that make-to-order regimes may not be a viable alternative, especially if production time variability cannot be reduced. In addition to decreasing production time variability, if customer sensitivity to delay can be decreased by producing high quality items and sustaining good customer relations, we see that proposed policies perform much better.

The rest of the paper is organized as follows. In Section 2, we present the problem analyzed. We discuss the proposed policies in Section 3 and present our numerical examples in Section 4. Section 5 is for the concluding remarks and possible future research questions.

2 The $M_n/GI/1$ Queue with Price and Lead-Time Quotations

In this section, we consider a manufacturer producing a single type of item whose underlying production system is modeled as a make-to-stock queue. Production is controlled by a base-stock policy: it stops when the continuously reviewed inventory level reaches the base-stock level S and starts as soon as the inventory level decreases to $S-1$. We assume that customers arrive one at a time. If there is stock, depending on the selling price an arriving customer may buy one item right away. If there is no stock, considering also the lead-time quoted, she may place an order, and consequently becomes a backlogged customer. Those who do not place an order – whether when there is stock or not – are simply lost. We assume that after placing it a backlogged customer never cancels her order and she eventually receives the item produced for her. A backlogged customer can be monetarily compensated for if the item is delivered beyond the quoted lead-time. The system does not incur any cost due to lost customers, however, there is an expected cost of K per unit time, which possibly includes the labor, the facility maintenance, and the material costs, that has to be paid for independent of the production status. To cover K, one would expect that the company cannot simply reject arriving customers (by posting extremely high prices or tendering unacceptably long lead times) and would try to attract customers while trading off the inventory holding and the late delivery/tardiness costs against the revenue to be accrued from a new customer.

Thus, for each item sold or ordered, a production order is created. In the rest of the paper, we refer to customers buying directly from the stock or placing orders as "customers" and production orders as "orders". Let $N(t)$, denote the number of (production) orders present at time t in the single server queueing system modeling the production facility. $N(t)$ gives the shortfall from the base-stock level S. This implies that when $N(t) \leq S$, the inventory carries $S - N(t)$ units and when $N(t) > S$, the system has $N(t) - S$ backlogged customers. In this setting, we focus on policies under which the price to charge and the lead-time to quote depend on the number of orders present, n, when a new customer arrives. We assume that customers (and consequently orders) arrive according to a Poisson process with a state-dependent arrival rate λ_n when there are n orders in the queueing system. Such a customer generates R_n as the revenue. If there is no stock, a lead-time d_n is quoted to such a customer. If the produced item cannot be delivered during the quoted lead-time, a tardiness cost, l, is incurred/paid to the customer per unit time for her waiting time in excess of d_n. Additionally, the system incurs a holding cost of h per unit inventory per unit time. The production/service times are assumed to be independent and identically distributed (i.i.d) random variables (r.v.s) with an LT denoted by $\widetilde{b}(\theta)$, a mean and second moment of $\beta_1 = 1/\mu$ and β_2, respectively and a variance of $\sigma^2 = \beta_2 - \beta_1^2$. Let also $c^2 = \sigma^2/\beta_1^2$ denote its squared-coefficient of variation.

The company maintains certain *principles of fairness* in alternative policies that can be implemented because we assume customers to be homogeneous in their response to the charged price and quoted lead times. Accordingly, the policies should satisfy the following principles of fairness:

1. The company charges the same price to customers if the same lead-time is quoted to them.

2. The company does not charge a higher price to a customer that is quoted a longer lead-time than the price charged to a customer to whom a shorter lead-time is quoted.

3. The company is socially responsible to deliver a disclosed proportion of deliveries within the quoted lead-time.

As a result of Principle 1, for instance, the company does not charge different prices to customers arriving when there is stock. These are the customers served with the highest service level (experiencing zero delay) who would pay the highest price. Even if the non-zero inventory level could be different when different customers arrive at different times, from the perspectives of all these customers the conditions are the same because they are to receive the item right away if they decide to buy it. Thus, the company can not risk losing customer confidence by charging different prices, which would make the company appear as exploiting some customers from time to time.

Therefore, the company decides on $S \geq 0$ and the vectors $\mathbf{d} = [d_0, d_1, \ldots, d_S, d_{S+1}, \ldots]/\mathbf{R} = [R_0, R_1, \ldots, R_S, R_{S+1}, \ldots]$ where d_n/R_n is the announced lead-time/charged price to a customer when there are n orders in the system with $d_n = 0$ for $n = 0, 1, \ldots, S-1$ and $R_0 = R_1 = \ldots = R_{S-1} > R_S \geq R_{S+1} \geq \ldots$. Assuming that the system is stable, for given S, $\mathbf{d}$, and $\mathbf{R}$ with the steady-state probability of having n orders in the system, namely $p(n) = P(N = n)$, the expected profit per unit time is

$$
\begin{aligned}
P(S, \mathbf{R}, \mathbf{d}) &= E[RV] - E[C_H] - E[C_D] - K, \\
&= \sum_{n=0}^{\infty} \lambda_n R_n p(n) - h \sum_{n=0}^{S-1} (S-n) p(n) - l \sum_{n=S}^{\infty} \lambda_n p(n) L_n(d_n) - K, \qquad (1)
\end{aligned}
$$

subject to

$$P(T_{n+1} \leq d_n) \approx \alpha, \quad n = S, \ldots. \qquad (2)$$

In Eq. (1), $L_n(d_n)$ is the expected waiting time in excess of d_n of a customer that accepts the quoted lead-time d_n. In Eq. (2), T_{n+1} is the r.v. showing the elapsed time from the moment she places this order until she receives the finished item (the subscript referring to the $(n+1)$st order that will be sent to the make-to-stock queue due to this customer) and α is the proportion of deliveries that should be done within the quoted lead times. Then, with $g_{n+1}(\cdot)$, the probability density function (PDF) of T_{n+1}, we have

$$L_n(d_n) = \int_{d_n}^{\infty} (x - d_n) g_{n+1}(x) dx. \tag{3}$$

Observe that the first term on the RHS of Eq. (1) is the expected revenue per unit time ($E[RV]$) whereas the second and third terms are the expected inventory holding ($E[C_H]$) and delay penalty cost rates ($E[C_D]$), respectively, while the last item is the expected cost rate that has to be incurred even if no item is produced.

Alternative policies introduced in Section 3 could yield different $E[RV]$ values. To be able to compare their profitability, we employ the following profit margin as the criterion in the numerical study in Section 4:

$$PM = \frac{P(S, \mathbf{R}, \mathbf{d})}{E[RV]}. \tag{4}$$

Let $\widetilde{g}_{n+1}(\theta)$ denote the LT of T_{n+1} which changes according to the policy implemented as discussed in Section 3. In the remainder of the paper, for various computations, we need to numerically invert a given LT $\widetilde{k}(\theta)$ and evaluate at d which will be denoted by $\mathcal{L}^{-1}\{\widetilde{k}(\theta)\}(d)$. Following Kahvecioğlu and Balcıoğlu (2016), Eq. (3) can be rewritten as

$$L_n(d_n) = \int_{d_n}^{\infty} x g_{n+1}(x) dx - d_n \overline{G}_{n+1}(d_n), \tag{5}$$

where $\overline{G}_{n+1}(\cdot)$ is the complementary distribution function of T_{n+1} with $\overline{G}_{n+1}(d_n) = \mathcal{L}^{-1}\{(1 - \widetilde{g}_{n+1}(\theta))/\theta)\}(d_n) \approx (1-\alpha)$ and finally we arrive at

$$L_n(d_n) = E[T_{n+1}] + \mathcal{L}^{-1}\{\frac{\widetilde{g}'_{n+1}(\theta)}{\theta}\}(d_n) - d_n \mathcal{L}^{-1}\{\frac{1 - \widetilde{g}_{n+1}(\theta)}{\theta}\}(d_n), \tag{6}$$

where $\widetilde{g}'_{n+1}(\theta)$ is the derivative of $\widetilde{g}_{n+1}(\theta)$.

In the next section, we discuss four fair policies that such a company can consider for profit maximization.

3 Alternative Fair Policies

In this section, we propose four fair policies, the first two operating under the static pricing scheme, and the other two with dynamic pricing strategies. We assume that the customer arrival rate is a continuous function of the charged price and the quoted lead-time. Here we have to distinguish the different natures of these two variables: While the price is an independent decision variable, which affects the arrival rate, the quoted lead-time depends on the arrival process. Consider an $M/GI/1$ queue where the linear $\lambda(R,d) = \lambda_0 - aR - bd_\alpha$ function is capturing the demand as a function of the price and the lead-time, where λ_0, a, and b are some constants. Choosing the price R and the lead-time d_α at our will gives us (as long as it is non-negative) an arrival rate but when the statistics with this arrival rate are computed, we may not see that α portion of the customers would really spend less than d_α time units in this $M/GI/1$ queueing system. Thus, under each policy, for each n (the number of orders) in the underlying queueing system, first an arrival process is considered, which consists of λ_k, $k = 0, \ldots, n$, and the corresponding lead-time d_n is computed. With λ_n and d_n for n, now in hand, the price R_n (which has to be non-negative) is determined as $R_n = (\lambda_0 - \lambda_n - bd_\alpha)/a$.

This is how a decision maker tries to determine the optimal prices to charge together with the optimal base-stock level: In Section 2, we discuss that, to observe fairness, there has to be a single price to charge when there is stock, which results in the same customer arrival rate for all $n < S$. In all policies, the price charged when there is stock should be the highest since the highest service level (zero delay in product delivery) is offered to customers. Thus, we denote the price and arrival rate when there is stock by R_H and λ_H, respectively, where the subscript H indicates the high price charged (or the high service level provided). Observe that we do not have a closed-form, differentiable function of profit margin in Eq. (4) (neither do we have a differentiable profit rate function in Eq. 1) from which we can obtain the optimal parameters. Instead, these are "approximately" found via searching over profit margin values computed for some discrete values of parameters such as the arrival rate and the base-stock level. For instance, the objective can be maximizing the profit margin in Eq. (8) of a system that does not allow any backlogs. Then, for each (λ_H, S) couple to consider, profit margins are computed as outlined in the SMTS Algorithm in Section 3.1. Then, among the computed values we identify the maximum profit margin and designate the parameters yielding it as the optimal λ_H^*, R_H^*, S^*.

The problem becomes more challenging when backlogging is allowed in the absence of stock: we end up facing practically an endless list of possible arrival rates to choose from for each n, $n \geq S$ if the quoted d_n depends on n and all λ_k's, $k = 0, \ldots, n-1$ as well. For instance, if there are M possible values to consider for each λ_n, for a system allowing at most N backlogs, we have to search for the "approximate" optimal solution over M^{N+1} arrival rate configurations of $(\lambda_0, \lambda_1, \ldots, \lambda_N)$. Thus, we cannot determine the optimal parameters of a policy allowing different arrival rates for each n, $n \geq S$. Consequently, we cannot characterize the optimal policy or determine its parameters for our problem. Therefore, we restrict our attention to policies that would keep the same arrival rate λ_L for all n when there is no stock, which would lead us to take into account fewer arrival rate configurations of the form (λ_H, λ_L) while searching for the optimum. As we demonstrate later on, with fixed λ_L, it is still possible to quote different d_n and consequently charge different R_n for different $n \geq S$.

3.1 The SMTS Policy: The Static Pricing Policy in the $M_n/GI/1$ Make-To-Stock Queue

If all customers are going to be charged the same price in a system that keeps stock, according to the fairness principles outlined in Section 2, the system should not be quoting lead times, which would otherwise necessitate charging lower prices to customers arriving when there is no stock. This implies that customers arriving when out of stock are lost. Given S, λ_H, and R_H, Eqs. (1) and (4) become

$$P(S) = \lambda_H R_H \sum_{n=0}^{S-1} p(n) - h \sum_{n=0}^{S-1} (S-n)p(n) - K, \quad (7)$$

$$PM_{SMTS} = \frac{E[RV] - E[C_H] - K}{E[RV]} = \frac{P(S)}{\lambda_H R_H \sum_{n=0}^{S-1} p(n)}, \quad (8)$$

respectively, with no costs arising due to tardiness. As a service level measure, we can also compute the proportion of customers that can be served:

$$\zeta = \sum_{n=0}^{S-1} p(n).$$

We employ the following SMTS (**S**tatic price policy in a **M**ake-**T**o-**S**tock system: the capital letters in bold yield the acronym) Algorithm in Section 4 to optimize the base-stock level and the arrival rate (and the price to charge). In this algorithm and the ones to be presented in Sections 3.2-3.4, the parameter values to consider are varied over appropriately chosen ranges. For instance, the STMS Algorithm uses two loops: starting from a minimum $\lambda^H_{\min}$ the external loop increments λ_H by some Δ in each round until a maximum $\lambda^H_{\max}$ is attained. For a λ_H provided by the external loop, starting from a base-stock level of 1, the internal loop increments S by 1 until an $S_{\max}$ is reached. To shorten the description of the algorithms, therefore, we skip outlining the basic loops and instead, we present what the algorithm does for a given parameter instance. At the end, each algorithm identifies the instance that maximizes the profit margin which in return yield the optimal parameters.

The SMTS Algorithm: This algorithm explains how the optimal SMTS policy parameter, λ^*_H (with the corresponding R^*_H) and S^* are found.

Main Step For the (S, λ_H) values considered: Employ the algorithm provided by Yang to obtain the steady-state probabilities of having n production orders ($p(n)$) in the underlying $M_n/GI/1$ queue. Using λ_H obtain the corresponding R_H from the $\lambda(R, d)$ function. Compute and record PM_{SMTS} from Eqs.(7) and (8) to be used in the Final Step.

Final Step Among all the instances with positive $P(S)$ values coming from the Main Step, the one with the highest PM_{SMTS} value gives the optimal instance and its parameters are the optimal λ_H^*, R_H^*, and S^*. If none of the instances yields a positive profit, the SMTS policy is deemed not profitable/feasible.

3.2 The SMTO Policy: The Static Pricing Policy in the $M/GI/1$ Make-To-Order Queue

If no stock is kept, the system is a make-to-order system. If a single price is going to be charged, according to the fairness principles, each customer should be quoted the same lead-time d_α. The random delivery time for an arbitrary customer is the system time of a production order in the $M/GI/1$ queue denoted by W. Then, Eq. (2) becomes

$$P(W \leq d_\alpha) \approx \alpha, \tag{9}$$

for all customers. At the end of this section, we provide the SMTO Algorithm to obtain d_α satisfying the constraint provided above.

After d_α is computed for λ_L, from the function relating the arrival rate to the price and the lead-time, the corresponding price denoted by R_L is computed. And Eqs. (1) and (4) become

$$P(d_\alpha) = \lambda_L(R_L - lL(d_\alpha)) - K, \tag{10}$$

$$PM_{SMTO} = \frac{E[RV] - E[C_D] - K}{E[RV]} = \frac{P(d_\alpha)}{\lambda_L R_L}, \tag{11}$$

respectively, with no costs arising due to holding stock. Eq. (3) turns into $L(d_\alpha) = \int_d^\infty (x - d_\alpha) w(x) dx$ with $w(x)$ denoting the probability density function of W for which the LT is (e.g., Gross and Harris, 1998, p. 226)

$$\widetilde{w}(\theta) = \frac{(1 - \lambda_L \beta_1)\theta \widetilde{b}(\theta)}{\theta - \lambda_L(1 - \widetilde{b}(\theta))}.$$

Consequently, from Eq. (6) we have

$$L(d_\alpha) = E[W] + \mathcal{L}^{-1}\{\frac{\widetilde{w}'(\theta)}{\theta}\}(d_\alpha) - d_\alpha \mathcal{L}^{-1}\{\frac{1 - \widetilde{w}(\theta)}{\theta}\}(d_\alpha), \tag{12}$$

where $\widetilde{w}'(\theta)$ is the first derivative of $\widetilde{w}(\theta)$ and the mean delivery time $E[W]$, that is, the mean sojourn time of an order in the $M/GI/1$ queue is (e.g., Kleinrock, 1975, p. 190)

$$E[W] = \beta_1 + \frac{\lambda_L(1+c^2)\beta_1^2}{2(1-\lambda_L\beta_1)}.$$

Recall that $\beta_1 = 1/\mu$ denotes the mean production time. Note that in the $M/M/1$ case, W is exponentially distributed with rate $\mu - \lambda_L$ (e.g., Gross and Harris, 1998, p. 68). Thus, Eq. (9) can be solved as a strict equality from which we obtain

$$d_\alpha = \frac{-\ln(1-\alpha)}{\mu - \lambda_L}, \tag{13}$$

and from Eq. (5), one arrives at

$$L(d_\alpha) = \frac{1-\alpha}{\mu - \lambda_L}. \tag{14}$$

We employ the following SMTO (**S**tatic price policy in a **M**ake-**T**o-**O**rder system: the capital letters in bold are used to obtain the acronym) Algorithm in Section 4 to optimize the arrival rate (with the due-date to quote and the price to charge):

The SMTO Algorithm: This algorithm explains how the optimal SMTO policy parameter λ_L^* and the corresponding d_α^* and R_L^* are found.

Main Step For the λ_L considered: Employ Sanajian and Balcıoğlu (2009) to obtain the steady-state probabilities of having n production orders ($p(n)$) in the underlying $M/GI/1$ queue. Set LB=0 and UB=$d_{\max}$, respectively, as the lower and upper limits for the interval over which the following binary search is conducted to determine the d_α value:

Step 1a Set $d_\alpha = (LB + UB)/2$. Using the numerical LT inversion technique of Abate and Valkó (2004), invert $L^{-1}\{\widetilde{w}(\theta)/\theta\}(d_\alpha) = P(W \leq d_\alpha)$. If $P(W \leq d_\alpha) = \alpha \pm \epsilon_\alpha$ for some tolerance ϵ_α chosen, then d_α is the lead-time to announce (Instead of numerically inverting the LT, one can use Eq. 13 if the production time is exponentially distributed). Go to Step 1c. Else go to Step 1b.

Step 1b If $L^{-1}\{\widetilde{w}(\theta)/\theta\}(d_n) = P(W \leq d_\alpha) < \alpha$ (implying that a longer lead-time is needed), then set LB=d_α and go to Step 1a. If $L^{-1}\{\widetilde{w}(\theta)/\theta\}(d_n) = P(W \leq d_\alpha) > \alpha$ (implying that a shorter lead-time is needed), set UB=d_α and go to Step 1a.

Step 1c Using λ_L and d_α coming from Step 1a, obtain R_L from the $\lambda(R,d)$ function and go to Step 2.

Step 2 Compute and store PM_{SMTO} using Eqs.(12) (or Eq. 14 for the exponential production times), (10) and (11) to be compared in the Final Step (Employ the numerical LT inversion technique of Abate and Valkó for computing Eq.12).

Final Step Among all the instances with positive $P(d_\alpha)$ values coming from the Main Step, the one with the highest PM_{SMTO} value gives the optimal instance and its parameters are the optimal λ_L^*, R_L^*, and d_α^*. If none of the instances yields a positive profit, the SMTO policy is deemed not feasible/profitable.

3.3 The SDP Policy: The Simple Dynamic Pricing Policy in the $M_n/GI/1$ Make-To-Stock Queue

If a dynamic pricing policy is to be implemented in a make-to-stock system, the simplest policy would be charging two prices: a high price R_H, yielding an arrival rate of λ_H, when there is stock and a low price R_L, yielding an arrival rate of λ_L, when there is no stock. If the same R_L is to be charged when there is no stock, all backlogged customers must be quoted the same lead-time d_α. Given S, the two prices (s.t. $R_H > R_L$), and the corresponding arrival rates (λ_H and λ_L), the following result from Eqs. (1) and (4):

$$\begin{aligned} P(S,\mathbf{R},d_\alpha) &= \lambda_H R_H \sum_{n=0}^{S-1} p(n) + \lambda_L R_L \sum_{n=S}^{\infty} p(n) \\ &\quad -h\sum_{n=0}^{S-1}(S-n)p(n) - l\lambda_L \sum_{n=S}^{\infty} p(n)L(d_\alpha) - K, \end{aligned} \tag{15}$$

$$\begin{aligned} PM_{SDP} &= \frac{E[RV]-E[C_H]-E[C_D]-K}{E[RV]}, \\ &= \frac{P(S,\mathbf{R},d_\alpha)}{\lambda_H R_H \sum_{n=0}^{S-1} p(n) + \lambda_L R_L \sum_{n=S}^{\infty} p(n)}. \end{aligned} \tag{16}$$

Observe that a backlogged customer waits only for customers backlogged earlier to be served. When all the backlogged customers are cleared, the system produces to stock. A customer (to be referred to as the *exceptional* backlogged customer whereas the others as the regular backlogged customers) that arrives when there are exactly $S(>0)$ production orders waits only for the ongoing production to finish for a product to be handed over to her (since $S = 0$ reduces the SDP Policy to the SMTO Policy, we do not consider this case again). If production times are not exponential, this residual production time for an exceptional backlogged customer is different in distribution from the production times for the regular backlogged customers.

From Kerner (2008) we obtain the LT $\widetilde{h}_n(\theta)$ of the residual service time experienced by a customer finding n orders upon arrival in the $M_n/GI/1$ make-to-stock queue operating under the SDP Policy as

$$\widetilde{h}_n(\theta) = \frac{\lambda_n}{\theta - \lambda_n}\left(\widetilde{b}(\lambda_n)\frac{1-\widetilde{h}_{n-1}(\theta)}{1-\widetilde{h}_{n-1}(\lambda_n)} - \widetilde{b}(\theta)\right), \quad n = 1, \dots, \tag{17}$$

with $\widetilde{h}_0(\theta) = \widetilde{b}(\theta)$ and

$$\lambda_n = \begin{cases} \lambda_H, & \text{for } n = 1, \dots, S-1, \\ \lambda_L, & \text{for } n \geq S. \end{cases}$$

Then, the exceptional backlogged customer finding S production orders upon arrival has $\widetilde{h}_S(\theta)$ as the LT for the exceptional service time she experiences.

Following Kerner again, we employ the following recursive method

$$\begin{aligned} E[H_1] &= \frac{\beta_1}{(1-\widetilde{b}(\lambda_1))} - \frac{1}{\lambda_1}, \\ E[H_n] &= \frac{\widetilde{b}(\lambda_n)}{1-\widetilde{h}_{n-1}(\lambda_n)}E[H_{n-1}] - \frac{1}{\lambda_n} + \beta_1, \ \ n \geq 2, \end{aligned} \tag{18}$$

from which we obtain $E[H_S]$, namely, the mean production time for the exceptional backlogged customers.

By taking the second derivative of $\widetilde{h}_n(\theta)$ in Eq. (17) and substituting 0 for θ, we obtain the following recursive formulae to compute the second moment of the residual service times as

$$E[H_1^2] = \frac{\beta_2}{(1-\widetilde{b}(\lambda_1))} - \frac{2\beta_1}{\lambda_1(1-\widetilde{b}(\lambda_1))} + \frac{2}{\lambda_1^2},$$

$$E[H_n^2] = \beta_2 + \frac{\widetilde{b}(\lambda_n)}{1-\widetilde{h}_{n-1}(\lambda_n)}\left(E[H_{n-1}^2] - \frac{2E[H_{n-1}]}{\lambda_n}\right) - \frac{2\beta_1}{\lambda_n} + \frac{2}{\lambda_n^2},\ \ n \geq 2, \tag{19}$$

to compute the second moment, $E[H_S^2]$, of the exceptional service time. Recall that β_2 in Eq. (19) denotes the second moment of the regular production time.

Now with $\widetilde{h}_S(\theta)$, $E[H_S]$, and $E[H_S^2]$ from Eqs. (17)-(19) in hand, characterizing the exceptional first service time, we can consider an "exceptional" $M/GI/1$ queue with arrival rate λ_L and a service time LT of $\widetilde{b}(\theta)$ except for the customers initiating the busy cycle who have an exceptional service time with an LT of $\widetilde{h}_S(\theta)$. Observe that the customers of this queue are probabilistically equivalent to the backlogged customers in a system operating under the SDP Policy. Then, the system time r.v. W_E in the exceptional $M/GI/1$ queue is the random delivery time of a product to a backlogged customer (exceptional or not) in the SDP system. With this, Eq. (2) becomes

$$P(W_E \leq d) \geq \alpha, \tag{20}$$

for all backlogged customers and Eq. (3) becomes $L(d) = \int_d^{\infty}(x-d)w_E(x)dx$ where $w_E(x)$ denotes the probability density function of W_E for which the LT is provided by Welch (1964) as

$$\widetilde{w}_E(\theta) = \frac{1-\beta_1\lambda_L}{1-\lambda_L(\beta_1 - E[H_S])}\frac{\lambda_L(\widetilde{h}_S(\theta) - \widetilde{b}(\theta)) - \theta\widetilde{h}_S(\theta)}{\lambda_L(1-\widetilde{b}(\theta)) - \theta}.$$

Consequently, from Eq. (6) we have

$$L(d) = E[W_E] + \mathcal{L}^{-1}\{\frac{\widetilde{w}'_E(\theta)}{\theta}\}(d) - d\mathcal{L}^{-1}\{\frac{1-\widetilde{w}_E(\theta)}{\theta}\}(d)$$

where $\widetilde{w}'_E(\theta)$ is the first derivative of $\widetilde{w}_E(\theta)$ and the mean delivery time $E[W_E]$ is again provided by Welch as

$$E[W_E] = \frac{\lambda_L(E[H_S^2] - \beta_2) + 2E[H_S]}{2(1-\lambda_L\beta_1 + \lambda_L E[H_S])} + \frac{\lambda_L\beta_2}{2(1-\lambda_L\beta_1)}.$$

Note that when production times are memoryless, that is, exponentially distributed, d_α and $L(d)$ can be computed using Eqs. (13)-(14), respectively.

We employ the following SDP (**S**imple **D**ynamic **P**ricing policy in a make-to-stock system: the capital letters in bold are used to obtain the acronym) Algorithm in Section 4 to optimize the base-stock level and the arrival rates (with the due-date to quote and the prices to charge):

The SDP Algorithm: This algorithm explains how the optimal SDP policy parameters S^*, λ_H^*, λ_L^*, and the corresponding d_α^*, R_H^*, R_L^* are found.

Main Step For the $(S, \lambda_H, \lambda_L)$ values considered: Employ the algorithm provided by Yang to obtain the steady-state probabilities of having n production orders ($p(n)$) in the underlying $M_n/GI/1$ queue. Using λ_H obtain the corresponding R_H from the $\lambda(R, d)$ function. Use Steps 1 and 2 of the SMTO Algorithm replacing $\widetilde{w}(\theta)$ by $\widetilde{w}_E(\theta)$ to obtain d_α, R_L. With all the parameters determined compute PM_{SDP} in Eq. (16).

Final Step Among all the instances with $R_H > R_L$ and positive $P(S, \mathbf{R}, d_\alpha)$ values coming from the Main Step, the one with the highest PM_{SDP} value gives the optimal instance and its parameters are the optimal S^*, λ_H^*, λ_L^*, R_H^*, R_L^*, and d_α^*. If none of the instances yields a positive profit, the SDP policy is deemed not feasible/profitable.

3.4 The RDP Policy: The Refined Dynamic Pricing Policy in the $M_n/GI/1$ Make-To-Stock Queue

This policy classifies backlogged customers in different groups according to how many production orders (n) they see upon arrival and announces a different lead-time d_n to each group with $d_n < d_{n+1}$ satisfying Eq. (2). According to the fairness principles, this also stipulates $R_H > R_S > \ldots > R_{S+N-1}$ where N is the maximum number of customers to backlog. Then, with the vectors $\mathbf{d} = [d_0, d_1, \ldots, d_S, d_{S+1}, \ldots, d_{S+N-1}]/\mathbf{R} = [R_0, R_1, \ldots, R_S, R_{S+1}, \ldots, R_{S+N-1}]$

where $d_n = 0/R_n = R_H$ for $n = 0, 1, \ldots, S-1$, Eq. (1) becomes

$$\begin{aligned} P(S, N; \mathbf{R}, \mathbf{d}) &= \lambda_H R_H \sum_{n=0}^{S-1} p(n) + \lambda_L \sum_{n=S}^{S+N-1} R_n p(n) \\ &\quad -h \sum_{n=0}^{S-1} (S-n) p(n) - l \sum_{n=S}^{S+N-1} \lambda_n p(n) L_n(d_n) - K, \qquad (21) \\ PM_{RDP} &= \frac{E[RV] - E[C_H] - E[C_D] - K}{E[RV]}, \\ &= \frac{P(S, N; \mathbf{R}, \mathbf{d})}{\lambda_H R_H \sum_{n=0}^{S-1} p(n) + \lambda_L \sum_{n=S}^{N+S-1} R_n p(n)}. \qquad (22) \end{aligned}$$

Using the notation introduced in Section 2, for a customer finding $n(= S, S+1, \ldots, S+N-1)$ production orders in the system, the random time for this customer to receive her product, T_{n+1}, has the LT of $\widetilde{g}_{n+1}(\theta)$ given as

$$\widetilde{g}_{n+1}(\theta) = \widetilde{h}_n(\theta) \widetilde{b}(\theta)^{n-S}, \qquad (23)$$

where $\widetilde{h}_n(\theta)$ is given in Eq. (17). Moreover, $E[T_{n+1}] = E[H_n] + (n-S)\beta_1$ to be used in Eq. (6) to compute $L_n(d_n)$ with $E[H_n]$ given in Eq. (18).

We employ the following RDP (**R**efined **D**ynamic **P**ricing policy in a make-to-stock system: the capital letters in bold are used to obtain the acronym) Algorithm in Section 4 to optimize the base-stock level, the maximum number of customers to backlog and the arrival rates (with the due-dates to quote and the prices to charge):

The RDP Algorithm: This algorithm explains how the optimal RDP policy parameters S^*, λ_H^*, λ_L^*, and the corresponding vectors $\mathbf{d}^*$, $\mathbf{R}^*$ are found.

Main Step For the $(S, N, \lambda_H, \lambda_L)$ values considered: Employ the algorithm provided by Yang to obtain the steady-state probabilities of having n production orders ($p(n)$) in the underlying $M_n/GI/1$ queue. Using λ_H obtain the corresponding R_H from the $\lambda(R, d)$ function. For each $n = S, \ldots, S+N-1$ use Step 1 of the SMTO Algorithm replacing $\widetilde{w}(\theta)$ by $\widetilde{g}_{n+1}(\theta)$ given in Eq. (23) to obtain d_n and R_n. With the due-date and price vectors now in hand, compute PM_{RDP}.

Final Step Among all the instances with $R_H > R_S > \ldots > R_{S+N-1}$ and positive $P(S, N; \mathbf{R}, \mathbf{d})$ values coming from the Main Step, the one with the highest PM_{RDP} value gives the optimal instance and its parameters are the optimal S^*, N^* λ_H^*, λ_L^*, $\mathbf{d}^*$, and $\mathbf{R}^*$. If none of the instances yields a positive profit, the RDP policy is deemed not feasible/profitable.

3.5 A Brief Comparison of the Proposed Policies

Before presenting the numerical study in the next section, what can we say about the relative performances of the proposed policies? When it comes to the static pricing policies, in the case of having extremely high holding cost rate together with a low tardiness penalty cost rate, we can foresee that the SMTS policy cannot have a chance of feasibility whereas the SMTO policy can turn out to be profitable. However, other than visualizing such extreme cases, without conducting computations, we cannot foretell which policy yields a higher profit margin.

We cannot also tell, without computations, which of the two dynamic pricing policies is superior. Due to its ability of reducing the number of backlogs we may think that the RDP policy can be superior, but we cannot show this merely with arguments. The different lead times and prices determined for each backlogged customer that the RDP charges make an easy comparison out of reach. We only have two results that can make pairwise comparison between a dynamic pricing and a static pricing policy, which lead to a third result showing that dynamic pricing policies are superior to static pricing policies.

Result 1 *The SDP policy is superior to the SMTO policy. In other words,*

$$PM_{SDP} \geq PM_{SMTO}.$$

Result 1 follows from the construction of the policies: the SDP policy can always increase R_H to make $\lambda_H = 0$ and $S = 0$ if that leads to the best solution, which would reduce it to the SMTO policy. That is to say the SDP policy cannot perform worse than the SMTO policy. On the other hand, due to the fairness principle, the SDP cannot increase R_L beyond R_H, which would prevent it to yield $\lambda_L = 0$. This implies that the SDP policy cannot reduce to the SMTS policy in the limit. Thus, we need computations to see which of the SDP and SMTS policies is better.

The RDP policy can also charge an infinite R_H to yield $\lambda_H = 0$ and $S = 0$ if it is more profitable. However, it quotes different lead times for backlogged customers seeing different number of orders upon arrival whereas the SMTO quotes one for all backlogged customers. This prevents us to foresee how it can perform with respect to the SMTO policy without making computations. Yet, the following result holds because the RDP can can choose to have no backlog if that is more profitable, which in turn reduces it to the STMS policy. Hence, the RDP cannot perform worse than the SMTS policy.

Result 2 *The RDP policy is superior to the SMTS policy. In other words,*

$$PM_{RDP} \geq PM_{SMTS}.$$

As a direct consequence of Results 1 and 2, we have the following result:

Result 3 *One of the dynamic pricing policies gives the highest profit margin. In other words,*

$$\max\{PM_{SDP}, PM_{RDP}\} \geq \max\{PM_{SMTO}, PM_{SMTS}\}.$$

4 Numerical Experiment

In this section, we primarily investigate the relative performances of the policies proposed in Section 3 via a numerical study. Recall that as stated in Result 3, one of the dynamic pricing policies would give us the highest profit margin. However, we do not know how much these profit margins would differ from one another, or in which settings the dynamic policies could be more profitable. The numerical study can shed some light on answers for these questions. It also helps us explore the impact of production time variability on the profitability of these policies. Finally, we make a note of whether the profitability would change significantly if fairness principles were ignored and one could charge a higher price to a customer that would wait a longer delivery time than a customer who could reach the product immediately.

To this end, we consider the linear model $\lambda(R, d) = \lambda_0 - aR - bd$ to capture different demand behaviors. Here $\lambda_0 > 0$ shows the potential market size whereas coefficients $a > 0$ and $b > 0$ capture the customer demand sensitivity to price and delay in delivery. Recalling that make-to-stock queues are just abstract representations of production/inventory systems, the values we assign to λ_0, a, and b (and the values for other parameters to be presented shortly) do not correspond to factual data. Instead, we assign a high and a low value for each parameter that appears in the linear demand function. Thus, higher λ_0 implies a bigger market whereas higher a/b indicates a higher customer sensitivity to increase in price/delay in delivery. In the first four columns of Table 1, we list the eight different demand functions considered for eight sets of numerical experiments. According to this, sets 1-4 (5-8) cover the smaller (larger) market examples. In their own market segment, sets 1 and 5 (4 and 8) are to capture the behavior of the customers who are the least (the most) sensitive to both price and delay, etc.

To introduce the impact of production time variability, we consider different service time distributions with unit mean ($\beta_1 = \mu = 1$), but different variances leading to different squared-coefficient of variation, c^2. For our numerical examples, we consider the following three service time distributions, each presented with its density function LT:

1. The deterministic service time with $c^2 = 0$ and the density function LT $\widetilde{b}(\theta) = e^{-\theta}$.

2. The exponential distribution with $\mu = 1$, $c^2 = 1$, and the density function LT

$$\widetilde{b}(\theta) = \frac{\mu}{\mu + \theta}.$$

3. The 2-stage Hyperexponential (H2) distribution with $\mu_1 = 4, \mu_2 = 0.6, p = 0.47$, $c^2 = 2$, and the density function LT

$$\widetilde{b}(\theta) = p\frac{\mu_1}{\mu_1 + \theta} + (1-p)\frac{\mu_2}{\mu_2 + \theta}.$$

Recall that an H2 distribution is an exponential distribution with rate μ_1 (μ_2) with probability p (1-p). Since higher c^2 indicates a more variable service time, the cases with H2 distribution would correspond to the most chaotic production facilities whereas the cases with deterministic service times to the most organized and smooth ones.

In total with 8 different demand functions, 3 service time distributions, and 4 policies, we have determined the optimal control parameters and PMs of 96 examples. These PMs are listed in Table 1. In all examples, the proportion of backlogged customers receiving their orders within the quoted lead-time is $\alpha = 0.9$. The holding cost, penalty cost rates and K are set as $h = 4$, $l = 4$ and $K = 20$, respectively. In all examples the highest average holding cost per unit time is 6.77 ($E[C_H]$ of the SMTS policy when service times are H2 r.v.s) while the highest average tardiness cost rate is 0.61 ($E[C_D]$ of the SMTO policy when service times are H2 r.v.s). These cost parameters prevent $S + N$ from assuming large values that would, otherwise, make the numerical LT inversion fail in the Main Step of The RDP Algorithm while inverting $\widetilde{g}_{n+1}(\theta)/\theta$ or $\widetilde{g}'_{n+1}(\theta)/\theta$ as n increases to $N + S - 1$ to compute Eq. (6).

In the supplementary document (SD), Tables 1-8 list the optimal results for four policies for three different service time distributions for data sets 1 to 8, respectively. For the RDP policy, we present the vectors $\mathbf{d}$ and $\mathbf{R}$ in SD Table 9. Notice that the last rows in SD Tables 1-8 are what appear in the corresponding rows of Table 1, which are the optimal PMs. The empty cells, as in those for the SMTO and the SMTS policies for sets 3 an 4 when service times are H2 r.v.s, indicate that the corresponding policy did not generate a positive profit, deeming it infeasible for that setting. Based on these results, we make the following observations:

- As expected, increase in price (a) or delay sensitivity (b) decreases the PM. Higher production time variability also lowers the PM. The larger market size (λ_0), on the other hand, increases it.

- For each case (for a given demand set and a production time distribution), we see the following holds:

$$PM_{RDP} > PM_{SDP} > \max\{PM_{SMTO}, PM_{SMTS}\}.$$

Except for three cases (for sets 1, 3, and 7 when the production times are deterministic), the SMTS policy yields higher PM than the SMTO policy. H2 production times for sets 3 and 4 render both static pricing policies unprofitable. These cases also give the lowest PMs for the dynamic policies.

- We define the following to capture the relative increase in PM when the SDP is used instead of the best static pricing policy for that case (which cannot be computed for sets 3 and 4 when the production times follow H2 distribution):

$$\Delta_1 = \frac{PM_{SDP} - \max\{PM_{SMTO}, PM_{SMTS}\}}{\max\{PM_{SMTO}, PM_{SMTS}\}} \times 100.$$

Table 2 displays the statistics concerning Δ_1. Although the PMs are higher in the larger market examples, the SDP yields a larger relative increase in PM in the smaller market. The lowest increases are seen for sets 2 and 6 (lower price, higher delay sensitivity) when production times are H2 r.v.s. The highest increases are seen for sets 3 and 7 (higher price, lower delay sensitivity) when production times are Exponential and H2 r.v.s, respectively (remember that we could not compute Δ_1 for set 3 when we have H2 production times). We see that the lowest and highest increases in Δ_1 are observed in cases where production times are not deterministic. And, we do not see a regular pattern for the impact of production time variability on Δ_1. In other words, we fail to say that the SDP gets relatively more (or less) profitable when production time variability increases.

- We define the following to capture the relative increase in PM when the RDP is used instead of the SDP:

$$\Delta_2 = \frac{PM_{RDP} - PM_{SDP}}{PM_{SDP}} \times 100.$$

Table 3 displays the statistics concerning Δ_2. We see that the relative advantage of using the RDP instead of the SDP is higher in the smaller market. The lowest increases are seen for sets 2 and 6 (lower price, higher delay sensitivity) when production times are H2 r.v.s. The highest increases are seen for set 4 with determinist production times and set 7 when production times are H2 r.v.s. We again fail to say that the RDP gets relatively more (or less) profitable when production time variability increases.

- For each policy, the best performance is obtained when production times are deterministic. We compute how much the PM decreases when production times are exponential or H2 r.v.s instead of being deterministic. For instance, for set 1, the PM decreases by 42.77% (from 38.20% to 21.86%) if service times are exponentially distributed instead of being deterministic. Let $\overline{\Delta}_E$ and $\overline{\Delta}_H$ denote the mean reduction in PM for a given policy when production times follow exponential and H2 distributions, respectively, instead of having deterministic production times. Table 4 lists $\overline{\Delta}_E$ and $\overline{\Delta}_H$ for the four policies studied. The least resilient policy against increase in production time variability is the SMTO policy. Although $\overline{\Delta}_H = -22.25\%$ for the SMTS policy appears better than those for the dynamic policies, this is because we omit the infeasible cases in the calculations (for this policy, those in data sets 3 and 4). These results agree with what Kahvecioğlu and Balcıoğlu observe in another setting as dynamic policies mitigating the worsening impact of production time variability better. Thus, we also recommend dynamic policies if an immediate solution is not available to reduce the production time variability.

- From the results presented in Tables 2 and 3, we can conclude that dynamic policies are relatively more profitable in the smaller market. In the smaller market, the RDP should be implemented for sure where customized service can be offered more conveniently. Although the RDP is still the best policy, the SDP, due to its simplicity, can be considered in the larger market given that it yields PMs closer to those of the RDP.

- Let $\overline{\Delta}_1$ and $\overline{\Delta}_2$ denote the mean values of Δ_1 and Δ_2 for each set, respectively, which are presented in each row in Table 5. We see that, on average, dynamic policies yield the highest relative increase in PM, when customers are price sensitive. These relative increases are the highest when customers are also less delay sensitive. These observations roughly sketch the environment where dynamic policies become unavoidable. If a company has competitors, customers can become more price sensitive. In this environment, if the company can also offer high quality products, customers can tend to be more willing to wait for the deliveries. This is the setting in which a company would increase its relative profitability the most by employing dynamic policies.

- As a final note, violating the fairness principle by possibly charging a higher price to a customer that would wait longer for delivery does not increase the PMs of the SDP and the RDP significantly. The highest relative increases in PM would occur as follows: For the set 3 with deterministic production times, the SDP would yield 24.57% as the PM (instead of 24.00% in Table 1) by charging 36.42 to customers arriving at a rate of 0.98 when there is stock (controlled by a base-stock level of $S = 2$) and 41.21 to customers arriving at a rate of 0.6 when there is no stock.

  Again in data set 3, this time with exponential production times, the RDP would yield 17.81% as the PM (instead of 17.00% in Table 1) by keeping a base-stock level of $S = 2$ and backlogging at most $N = 4$ customers by employing the price vector $\mathbf{R} = [38.21, 38.21, 43.23, 37.56, 32.42, 27.54]$ and $\mathbf{d} = [0, 0, 2.2949, 3.8818, 5.5323, 6.6895]$ as the lead-time vector. This policy yields 0.93/0.56 as the customer arrival rate when there is stock/no stock.

Table 1: The different demand functions considered and the summary of the results

Set	λ_0	a	b	Deterministic				Exponential				Hyperexponential			
				SMTO	SMTS	SDP	RDP	SMTO	SMTS	SDP	RDP	SMTO	SMTS	SDP	RDP
1	2	0.02	0.1	38.20%	37.74%	45.72%	48.88%	21.86%	32.08%	39.50%	40.80%	6.81%	28.32%	34.65%	35.83%
2		0.02	0.2	19.21%	37.74%	43.01%	45.03%		32.08%	35.09%	36.37%		28.32%	29.90%	30.39%
3		0.028	0.1	13.48%	12.84%	24.00%	28.43%		4.92%	15.30%	17.12%			8.51%	10.16%
4		0.028	0.2		12.84%	20.22%	23.18%		4.92%	9.12%	10.92%			1.87%	2.54%
5	2.4	0.02	0.1	55.33%	56.35%	60.79%	63.14%	44.52%	51.85%	56.17%	57.36%	34.33%	48.93%	52.60%	53.64%
6		0.02	0.2	43.61%	56.35%	58.74%	60.62%	19.97%	51.85%	53.30%	54.13%		48.93%	49.20%	49.78%
7		0.028	0.1	37.46%	37.11%	45.10%	48.40%	22.33%	30.68%	38.64%	40.30%	8.06%	26.49%	33.65%	35.10%
8		0.028	0.2	21.05%	37.11%	42.23%	44.86%		30.68%	34.62%	35.78%		26.49%	28.88%	29.69%

Table 2: The increase in PM (Δ_1) the SDP generates when compared to a static pricing policy

λ_0	Min	Mean	Median	Max
2	5.58%	52.62%	22.74%	210.99%
2.4	0.56%	11.70%	8.69%	27.02%

Table 3: The increase in PM (Δ_2) the RDP generates when compared to the SDP

λ_0	Min	Mean	Median	Max
2	1.62%	11.98%	9.43%	36.19%
2.4	1.17%	3.52%	3.27%	7.31%

Table 4: The mean reduction in PM due to higher production time variability

SMTO		SMTS		SDP		RDP	
$\overline{\Delta}_E$	$\overline{\Delta}_H$	$\overline{\Delta}_E$	$\overline{\Delta}_H$	$\overline{\Delta}_E$	$\overline{\Delta}_H$	$\overline{\Delta}_E$	H2
-39.23%	-66.21%	-25.50%	-22.25%	-21.55%	-37.09%	-23.16%	-38.34%

5 Conclusion and Future Work

In this paper, we propose four practical and fair pricing and lead-time quotation policies for a company serving price and delay sensitive customers with a single type of product. The production facility is modeled as an $M_n/GI/1/K$ queue. Three policies quoting lead times employ numerical inversion of the LT of the sojourn time r.v. of an order to be placed. The refined dynamic policy appears as the champion among four policies. This is due to its power of limiting the number of backlogged customers and its ability to quote separate lead times

Table 5: The mean increases in PM with $\overline{\Delta}_1$ from static pricing policies to the SDP and $\overline{\Delta}_2$ from the SDP to the RDP

λ_0	(a,b)	$\overline{\Delta}_1$	$\overline{\Delta}_2$
2	(0.02,0.1)	21.72%	4.54%
	(0.02,0.2)	9.64%	3.32%
	(0.028,0.1)	144.56%	16.57%
	(0.028,0.2)	71.50%	23.51%
2.4	(0.028,0.1)	7.91%	2.66%
	(0.028,0.2)	2.53%	1.97%
	(0.028,0.1)	24.46%	5.31%
	(0.028,0.2)	11.91%	4.12%

and charge different prices depending on the number of orders a backlogged customer sees. Yet, we have a restriction for this policy: The system receives two arrivals rates, one when there is stock and one when backlogging occurs. In the framework of our research, we could not overcome this restriction. A policy operating with different arrival rates depending on the number of backlogged customers can increase the profit margins even more. Although we provide the formulation of such a policy, the difficulty would arise if the optimal policy parameters were to be searched. This has led us to focus on the restrictive form. Instead of a queueing based analysis, a simulation-optimization approach may offer a solution that can also handle larger ranges of base-stock level and the number of backlogged customers. Our results and the future efforts would help create a market environment where customers are served fairly while companies preserve their profitability.

Acknowledgements

This work was supported in part by TÜBİTAK, The Scientific and Technological Research Council of Turkey, under the grant number 213M428.

References

Abate, J., and P.P. Valkó. 2004. “Multi-precision Laplace transform inversion”, *International Journal for Numerical Methods in Engineering*, Vol. 60, No. 5, 979–993.

Abouee-Mehrizi, H., and O. Baron. 2016. “State-dependent $M/G/1$ queueing systems”, *Queueing Systems*, Vol. 82, 121–148.

Bitran, G. and R. Caldentey. 2003. “An overview of pricing models for revenue management”, *Manufacturing & Service Operations Management*, Vol. 5, No. 3, 203–229.

Boyacı, T., and S. Ray. 2003. ”Product differentiation and capacity cost interaction in time and price sensitive markets”, *Manufacturing & Service Operations Management*, Vol. 5, No. 1, 18–36.

Chen, H. and M. Z. Frank. 2001. “State dependent pricing with a queue,” *IIE Transactions*, Vol. 33, 847–860.

Chen, L., Y. Feng, and J. Ou. 2006. “Joint management of finished goods inventory and demand process for a make-to-stock product: A computational approach”, *IEEE Transactions on Automatic Control*, Vol. 51, No. 2, 258–273.

Chen, L., Y. Chen, and Z. Pang. 2011. “Dynamic pricing and inventory control in a make-to-stock queue with information on the production times”, *IEEE Transactions on Automation Science and Engineering*, Vol. 8, No. 2, 361–373.

Çelik, S. and C. Maglaras. 2008. ”Dynamic pricing and lead-time quotation for a multiclass make-to-order queue”, *Management Science*, Vol. 54 , 1132–1146.

Dewan, S. and H. Mendelson. 1990. “User delay costs and internal pricing for a service facility", *Management Science*, Vol. 36, No. 12, 1502–1517.

Dellaert, N. P. 1991. "Due-date setting and production control", *International Journal of Production Economics*, Vol. 23, 59–67.

Duenyas, I., W. J. Hopp. 1995. "Quoting customer lead times", *Management Science*, Vol. 41, No. 1, 43–57.

Economou, A., and A. Manou. 2015. "A probabilistic approach for the analysis of the $M_n/G/1$ queue", *Annals of Operations Research*, https://doi.org/10.1007/s10479-015-1943-0.

Elmaghraby, W. and P. Keskinocak. 2003. "Dynamic pricing in the presence of inventory considerations: Research overview, current practices, and future directions", *Management Science*, Vol. 49, No. 10, 1287–1309.

Feng, J., and M. Zhang. 2017. "Dynamic quotation of leadtime and price for a make-to-order system with multiple customer classes and perfect information on customer preferences", *European Journal of Operational Research*, Vol. 258, 334–342.

Gayon, J.P., I. Talay-Değirmenci, F. Karaesmen, and L. Örmeci. 2009. "Optimal pricing and production policies of a make-to-stock system with fluctuating demand", *Probability in the Engineering and Informational Sciences*, Vol. 23, 205–230.

Gross, D., C. M. Harris. 1998. *Fundamentals of Queueing Theory.* John Wiley & Sons, New York.

Hammami, R., Frein, Y., and A. S. Albana. 2020. "Customer rejection to guide lead time quotation and pricing decisions'", *Journal of the Operational Research Society*, DOI: 10.1080/01605682.2020.1718556

Kahvecioğlu, G., and B. Balcıoğlu. 2016. "Coping with production time variability via dynamic lead-time quotation", *OR Spectrum*, Vol. 38, 877–898.

Kerner, Y. 2008. "The conditional distribution of the residual service time in the $M_n/G/1$ queue", *Stochastic Models*, Vol. 24, 364–375.

Kleinrock, L. (1975). *Queueing Systems Volume I: Theory*, John Wiley & Sons, New York.

Li, L. 1988. “A Stochastic Theory of the Firm”, *Mathematics of Operations Research*, Vol. 13, No. 3, 447–466.

Low, D. 1974. “Optimal dynamic pricing policies for an $M/M/s$ queue”,*Operations Research*, Vol. 22, No. 3, 545–561.

Mendelson, H. 1985. “Pricing computer services: Queueing effects”, *Communications of the ACM*, Vol. 28, No. 3, 312–321.

Naor, P. 1969. “The regulation of queue size by levying tolls”, *Econometrica*, Vol. 37, No. 1, 15–24.

Palaka, K., S. Erlebacher, and D. Kropp. 1998. “Lead time setting, capacity utilization, and pricing decisions under lead-time dependent demand”, *IIE Transactions*, Vol. 30, No. 2, 151–162.

Pekgün, P., P. M. Griffin, and P. Keskinocak. 2008. ”Coordination of marketing and production for price and leadtime decisions”, *IIE Transactions*, Vol. 40, No. 1, 12–30.

Ray, S. and E. M. Jewkes. 2004. ”Customer lead time management when both demand and price are lead time sensitive”, *European Journal of Operational Research*, Vol. 153, No. 3, 769–781.

Sanajian, N., B. Balcıoğlu. 2009. “The impact of production time variability on make-to-stock queue performance”, *European Journal of Operational Research*, Vol. 194, 847–855.

Savaşaneril, S., P. M. Griffin, and P. Keskinocak. 2010. “Dynamic lead-time quotation for an M/M/1 base-stock inventory queue”, *Operations Research*, Vol. 58, No. 2, 383–395.

Webster, S. 2002. ”Dynamic pricing and lead time policies for make-to-order systems,” *Decision Sciences*, Vol. 33, No. 4) 579–599.

Welch, P. D. 1964. “On a generalized $M/G/1$ queuing process in which the first customer of each busy period receives exceptional service”, *Operations Research*, Vol. 12, No. 5, 736–752.

Yang, P. 1994. “A unified algorithm for computing the stationary queue length distributions in $M(k)/G/1/N$ and $GI/M(k)/1/N$ queues”, *Queueing Systems*, Vol. 17, 383–401.

Yano, C. A. and S.M. Gilbert. 2002.“ Coordinated pricing and production/procurement decisions: A review”. In Chakravarty, A. and J. Eliashbert (eds.), *Managing business interfaces: Marketing, engineering and manufacturing perspectives.* Amsterdam: Kluwer Academic.

Zhao, X., K. E. Stecke, and A. Prasad. 2012. “Lead time and price quotation mode selection: Uniform or differentiated?”, *Production and Operations Management*, Vol. 21, No. 1, 177–193.