Dynamic pricing and lead-time quotation for a multi-class make-to-order queue

Sabri Çelik† and Constantinos Maglaras‡

May 31, 2005

Abstract

Consider a make-to-order manufacturer that offers multiple products to a market of price and delay sensitive users. This paper studies the problem of maximizing its long-run expected profits for a model that captures three aspects of particular interest: first, the joint use of dynamic pricing and leadtime quotation controls to manage customer demand; second, the presence of a dual sourcing mode that can be used to expedite orders at a cost; and third, the interaction of the aforementioned demand controls with the operational decisions of sequencing and expediting that the firm must employ to optimize revenues and satisfy the quoted leadtimes. Using an approximating diffusion control problem we derive near-optimal dynamic pricing, leadtime quotation, sequencing, and expediting policies that provide structural insights and lead to practically implementable recommendations. A set of numerical results illustrate the value of joint pricing and leadtime control, as well as the performance of the proposed set of policies.

1 Motivation and model formulation

Starting with the airline industry, the adoption of tactical demand management or revenue management strategies has transformed the transportation and hospitality sectors over the past couple of decades. Broadly speaking, this involves the use of sophisticated information technology systems and intense data processing to construct detailed and granular forecasts, quantitative models of consumer demand, and dynamic capacity allocation and/or pricing strategies to maximize the expected revenues from a fixed set of resources, as for example, a network of flights operated by a certain carrier. Similar approaches are now becoming increasingly important in retail, telecommunications, entertainment, financial services, health care and manufacturing. This paper is motivated from the latter, a notable example of which comes from the automotive industry and their push

---

*The full version of this paper is available at http://www.gsb.columbia.edu/faculty/cmaglaras/maglaras.htm
†IEOR Department, Columbia University, 500 W. 128th St., NY, NY 10027. (sc2190@columbia.edu)
‡Columbia Business School, 409 Uris Hall, 3022 Broadway, NY, NY 10027. (c.maglaras@gsb.columbia.edu)
towards producing customized cars in a make-to-order fashion. A revenue management strategy applied in such a setting would aim to dynamically choose the price, leadtime, rebate, etc. for a new order as a function of their book of existing orders, and simultaneously select the production schedule to optimize their profitability. Joint use of economic and operational controls allows the manufacturer to be more responsive to changes in the market conditions, as well as to fluctuations in the operating environment due to variability in the demand and production processes. And, using both price and leadtime signals to manage demand, allows the firm to achieve a form of dynamic product differentiation to exploit the customers’ heterogeneity in terms of their price and delay sensitivities and drive higher profitability.

We consider a make-to-order firm that offers multiple products, indexed by $i = 1, \ldots, I$, to a market of price and delay sensitive customers. The model described below aims to provide a tractable framework for revenue optimization of such systems, capturing three features of particular interest: first, the joint use of dynamic pricing and leadtime quotation controls to manage demand; second, the access to a dual sourcing mode that can be used to expedite orders at a cost; and third, the interaction between the demand controls with the operational ones of sequencing and expediting that the firm employs to maximize its profitability.

**Leadtime guarantees.** The firm will offer each “good” at multiple predetermined leadtimes, whereby a “product” corresponds to a (type of good, leadtime) combination. By dynamically adjusting the product prices the firm can divert demand from one leadtime to another, thus exercising dynamic leadtime control over this predetermined set of options. Specifically, product $i$ orders are quoted a leadtime “guarantee” of $d_i$ time units, which serves as a reliable upper bound for the time it takes from when the order is placed until its production is completed. In a stochastic production setting such guarantees are typically stated as $\mathbb{P}(\text{delay for a class } i \text{ order } > d_i) \leq \epsilon_i$, where $\epsilon_i \in (0, 1)$ is a desired service level, but the capability for instantaneous expediting (explained later on) allows us to circumvent such probabilistic constraints, and instead impose “hard” leadtime guarantees, i.e., $\epsilon_i = 0$ for all $i$.

**Economic structure and demand model.** The firm operates in a market with imperfect competition, and has power to influence its vector of demand rates by varying its prices $p$; $p_i(t)$ denotes the per-unit price for product $i$. Potential customers arriving at the system at time $t$ observe the current menu of products, which is summarized by the pair $(p(t), d)$ and make their decision of which product to buy, if any. Given the vector of leadtime guarantees $d$ and the selected prices $p$, the resulting demand is assumed to be an $I$-dimensional non-homogeneous Poisson process with instantaneous rate vector $\lambda(p(t); d)$ determined through a *demand function* that maps a price vector $p \in \mathcal{P}$ into a vector of demand rates $\lambda \in \mathcal{L}(d)$, where $\mathcal{P} \subseteq \mathbb{R}^I$ is the set of feasible price vectors, and

---

1For example, BMW claims that 80% of the cars sold in Europe and 30% of those sold in the US are built to order. When a dealer inputs a potential order to BMW’s web ordering service, a target leadtime is generated within five seconds. This is typically 11 to 12 days in Europe and about double that amount in the US [5].
\( \mathcal{L}(d) = \{ x \geq 0 : x = \lambda(p; d), \ p \in \mathcal{P}, \ d \in \mathbb{R}_+^I \} \subseteq \mathbb{R}_+^I \) is the set of achievable demand rate vectors. We assume that \( \mathcal{L}(d) \) is a convex set for all \( d \in \mathbb{R}_+^I \), the demand function \( \lambda(p; d) \) is stationary, continuously differentiable in both \( p \) and \( d \), and bounded. And, in addition: (a) for each product \( i \), \( \lambda_i(p; d) \) is strictly decreasing in \( p_i \); (b) for each feasible \( p-i = (p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_I) \) and leadtime vector \( d \), there exists a null price \( p^\infty_i(p-i) \in \mathcal{P} \) such that \( \lim_{p_i \to p^\infty_i(p-i)} \lambda_i(p_i, p-i; d) = 0; \) and (c) the revenue rate \( p \cdot \lambda(p; d) = \sum_i p_i \lambda_i(p; d) \) is bounded for all \( p \in \mathcal{P} \) and has a finite maximizer. (For any two \( n \)-vectors, \( x \cdot y \) will denote their inner product.)

Under these assumptions, there exists an inverse demand function \( p(\lambda; d), p : \mathcal{L}(d) \to \mathcal{P}, \) that maps an achievable vector of demand rates \( \lambda \) into a corresponding vector of prices \( p(\lambda; d) \). Following a standard practice from revenue management, we may then view the demand rate vector as the firm’s control, and once this is determined derive the corresponding prices using the inverse demand function. In this case, the expected revenue rate will be denoted by \( r(\lambda; d), \) where \( r(\lambda; d) = \lambda \cdot p(\lambda; d) \). We will assume that \( r(\lambda; d) \) is bounded, strictly concave and twice continuously differentiable, and denote its maximizer by \( \lambda^*(d) := \text{argmax} \{ r(\lambda; d) : \lambda \in \mathcal{L}(d) \} \).

**The system model.** The production facility is modelled as a multi-product (or multi-class) single-server queue. Orders for each product arrive according to non-homogeneous Poisson processes upon arrival join dedicated, infinite capacity buffers associated with each product. For each product \( i \) the number of orders that were placed in \([0, t] \) is given by

\[
N_i \left( \int_0^t \lambda_i(s) ds \right),
\]

where \( N_i(t) \) is a unit rate Poisson process. Service time requirements for product \( i \) orders are independent identically distributed (i.i.d.), drawn from some general distribution with mean \( m_i \) (rate \( \mu_i = 1/m_i \)) and squared coefficient of variation \( \xi_i \), and \( S_i(t) \) will denote the number of class \( i \) service completions if the server dedicates \( t \) time units in processing class \( i \) orders. The processes \( N_i, S_i \) are independent of each other and across products.

The *load* or *traffic intensity* of the system when the demand vector is \( \lambda \) is defined as \( \rho := m \cdot \lambda \). For future use, we define the *aggregate* revenue function as the maximum achievable revenue rate when all products jointly consume capacity at rate \( \rho \),

\[
R(\rho; d) = \max \{ r(\lambda; d) : \lambda \in \mathcal{L}(d), \ \sum_i \lambda_i / \mu_i = \rho \},
\]

and denote by \( \lambda^*_\rho(d) \) the corresponding maximizer. We will assume that \( \lambda^*_\rho \) is non-decreasing in \( \rho \) for all products \( i \). From the properties of \( r(\cdot; d) \) it follows that \( R(\cdot; d) \) is concave, bounded, twice continuously differentiable, and has a finite maximizer \( \rho^*(d) := \text{argmax} \{ R(\rho; d) : \rho = m \cdot \lambda, \ \lambda \in \mathcal{L}(d) \} = m \cdot \lambda^*(d) \). §4 describes a specific customer choice model that satisfies our assumptions and seems suitable for the type of problem considered in this paper.
In addition to pricing, the firm also controls the operational decisions of order sequencing at the server, and order expediting. The latter models actions such as the use of overtime, sub-contractors, etc. to increase the firm’s short term production capacity, whenever necessary to meet its leadtime guarantees. Within each product, orders are processed in First-In-First-Out (FIFO), the server can only work on one job at any given time, and preemptive-resume type of service is allowed. Under these assumptions, a sequencing policy takes the form of the $I$-dimensional cumulative allocation process $(T_i(t) : t \geq 0)$ with $T_i(0) = 0$, where $T_i(t)$ denotes the cumulative time that the server has allocated to class $i$ jobs up to time $t$. In addition, $T(t)$ is continuous and non-decreasing, and satisfies the capacity constraint

$$\sum_i T_i(t) - \sum_i T_i(s) \leq t - s \quad \text{for } 0 \leq s \leq t < \infty.$$  

The expediting policy is modelled as an $I$-dimensional process $(B_i(t) : t \geq 0)$ with $B_i(0) = 0$, where $B_i(t)$ is the cumulative number of product $i$ orders that were expedited in $[0, t]$. We will make the simplifying assumption that expedited orders are produced (and get removed from the corresponding queue) instantaneously. The cost of expediting a class $i$ order is $c_i$, and without loss of generality we will assume that products are so labelled that $c_1 \mu_1 \geq c_2 \mu_2 \geq \cdots \geq c_I \mu_I$. Controls $(\lambda, T, B)$ that satisfy the above conditions are said to be admissible if they are also non-anticipating, i.e., decisions at time $t$ only use information that is available up to that time.

Let $Q_i(t)$ denote the number of product $i$ jobs in the system (i.e., in queue or in service) at time $t$. The queue length dynamics are described through the following equations:

$$Q_i(t) = Q_i(0) + N_i \left( \int_0^t \lambda_i(s)ds \right) - S_i(T_i(t)) - B_i(t) \quad \text{for } i = 1, \ldots, I.$$  

**Control problem formulation.** The profit maximization problem for the stochastic queueing model is the following: choose admissible demand, sequencing and expediting policies $(\lambda, T, B)$, respectively, to maximize the long-run expected average profit rate given by

$$\text{maximize } \lim_{t \to \infty} \frac{1}{t} \mathbb{E} \left[ \int_0^t r(\lambda(s))ds - c \cdot B(t) \right].$$

subject to the leadtime constraints specified above.

## 2 Summary of main results

This paper strives to contribute in terms of modelling, analysis, and the derivation of structural insights that may be useful in practical revenue management solutions for such systems.
**Modelling:** The system analyzed in this paper combines two novel features: first, the incorporation of expediting decisions that both enriches the class of systems under consideration and simultaneously simplifies the analysis of leadtime guarantees; and second, the particular way in which we formulate the dynamic leadtime decisions. Specifically, instead of using dynamic leadtime control, the firm commits to offer each “good” at multiple predetermined leadtimes, and focuses on pricing for these products. Through dynamic pricing the firm can divert demand from one leadtime to another, thereby exercising dynamic leadtime control over this discrete set of options. Restricting the possible leadtime options (e.g., 1, 2 or 4 weeks) may be more practical. Also, the customer choice behavior can be captured through a relationship that is parametrized by the leadtime vector but only varies as a function of the price menu, and the joint pricing and leadtime control problems reduces to one of pricing subject to leadtime guarantees, which is more tractable.

**Proposed solution:** The control problem described above could be tackled within the context of Markov Decision Processes but this is analytically and numerically intractable. This paper follows the general methodology proposed by Harrison [10, 11] that suggests studying the underlying control problem in an operating regime where the processing resources are almost fully utilized. The resulting formulation involves the control of a Brownian motion or a diffusion, and is often simpler than the original problem at hand. Apart from this analytical simplification, this operating regime can -at least, in some cases- be justified economically; see Maglaras and Zeevi [14] for such a result in the context of revenue maximization for a single-product model using static pricing.

The proposed set of controls is outlined below. First, we introduce some notation. Given the original problem parameters, define \( \lambda(d) := \text{argmax} \{ r(\lambda; d) : \sum_i \lambda_i/\mu_i = 1, \lambda \in \mathcal{L}(d) \} \) to be the demand rate vector that maximizes the instantaneous revenue rate subject to the constraint that the server is fully utilized. We assume that \( \lambda_i(d) > 0 \) for all \( i \), i.e., that it is optimal to produce all products (and use all of the available leadtime options) in this deterministic planning problem. Also, recall that \( \lambda^*(d) := \text{argmax} \{ r(\lambda; d) : \lambda \in \mathcal{L}(d) \} \), maximizes the instantaneous revenue rate without any consideration on the capacity constraint, and let \( \Lambda^*(d) = \sum_i \lambda^*_i(d) \) act as proxy of the total market potential for our problem. As in other papers in the revenue management literature that use approximations to compute pricing heuristics, the quantities \( \lambda(d) \) and \( \lambda^*(d) \) will play an important role in our analysis; c.f., Gallego and van Ryzin [7] show that the optimal demand strategy in a deterministic single-product model is static and is equal to the minimum between the rate that depletes capacity at the end of the planning horizon (cf. \( \lambda(d) \)) and the rate that maximizes revenues in the absence of the capacity constraint (cf. \( \lambda^*(d) \)). Our analysis focuses on optimizing the behavior of the second order stochastic fluctuations around the deterministic solution (much like the Central Limit Theorem characterizes the error term around the mean of i.i.d. random variables), and in that places particular emphasis in the difference between \( \lambda(d) \) and \( \lambda^*(d) \), which is measured in multiples of \( \sqrt{\Lambda^*(d)} \)’s, the natural system scale parameter on which to measure its second order fluctuations.
Pricing: Our analysis shows that the optimal demand control in the approximating diffusion model is a function of the aggregate workload in the system $W(t) = m \cdot Q(t)$. Specifically, given the workload position $w$, the manager computes the target resource utilization $\rho^*(w)$ as

$$\rho^*(w) := \left[ 1 - \left( \frac{1}{\sqrt{\Lambda^*(d)}} \right) \cdot \psi \left( w \sqrt{\Lambda^*(d)} \right) \right]^+, \quad (5)$$

where $\psi(\cdot)$ is a monotonically increasing function specified in Theorem 1, and then selects the demand rate vector

$$\lambda^*(w; d) = \arg\max \{ r(\lambda; d) : \lambda \cdot m = \rho^*(w), \lambda \in L(d) \} = \lambda^*(\rho^*(w); d). \quad (6)$$

The corresponding pricing strategy can be inferred via the inverse demand relation $p(\lambda; d)$.

Sequencing: Note that $\lambda^*(0; d) = \bar{\lambda}(d)$. For each product $i$, define a threshold

$$b'_i = \hat{\lambda}_i(d) \cdot d_i - \delta_i,$$

for some $\delta_i \in \mathbb{R}$, and sequence orders according to the policy that gives priority to class

$$i^* = \arg\max \frac{Q_i(t)}{b'_i};$$

this is the “least relative slack” policy of Plambeck et al. [17].

Expediting: Jobs are expedited when the total workload reaches $\tilde{W} = m \cdot b'$ according to a rule that gives priority to class $I$ then class $I - 1$ and so on. [This priority derives from the fact that products are labelled in a way that $c_1 \theta_1 \geq c_2 \theta_2 \geq \cdots \geq c_I \theta_I$.]

Structural insights: We conclude with a short list of some of the main insights gleaned from our analysis: i. Pricing and expediting decisions depend on the aggregate system workload and not the product-level queue lengths, and thus change on the slower time-scale on which the aggregate workload evolves, which is practically appealing. ii. The system expedites according to a greedy priority rule (from cheapest to more expensive) in order to keep the total workload below a certain level that is selected in accordance to the predetermined leadtime bounds. iii. Sequencing is done according to a dynamic rule that roughly speaking serves the order that is “closest” to violating its leadtime; $b'_i$ is a proxy for the number of class $i$ orders placed in $d_i$ time units. iv. The system regulates in an operating regime where its resource is almost fully utilized. v. Developing a revenue management solution for such a manufacturer requires an accurate, data-driven customer demand model, which leads to a tractable formulation. §4 outlines such a proposal that builds on extensive marketing research, and contrasts it against other natural demand model alternatives. The leadtime control induced by (5) and (6) depends on the structure of $\lambda(p; d)$. 