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

Sabri Çelik[*] and Constantinos Maglaras[†]

June 15, 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 average 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 illustrates the value of joint pricing and leadtime control, as well as the performance of the proposed set of policies.

Keywords: Revenue management, dynamic pricing, leadtime quotation, queueing, sequencing, diffusion models.

1 Introduction

This paper considers a make-to-order production firm that offers multiple products to a market of price and delay sensitive customers. The primary goal is to develop a tractable framework for revenue optimization in 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.

---
[*]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`)

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 by the latter, a notable example of which comes from the automotive industry and their push towards producing customized cars in a make-to-order fashion.[1] 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. In addition, 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.

In more detail, the production system is modelled as a multi-class $M_q/GI/1$ queue – the first subscript indicating that the arrival rate is state-dependent. The system manager can select the product prices and quoted leadtimes dynamically, and can also choose to instantaneously expedite existing orders at a cost that may depend on the type of product. Instantaneous expediting is, of course, an idealization, which serves to model systems with significant surge capacity vis-a-vis their nominal processing capability. Methodologically, it allows us to enforce the quoted leadtimes on all accepted orders, e.g., by expediting whenever an order's age in the system reaches its leadtime, rather than having to add service level guarantees. In addition, the manager has discretion with respect to the sequencing of orders at the server. Potential customers make their purchase decisions by optimally trading off price and delay in conjunction to their private valuations for the offered products. Broadly, the firm's problem is to dynamically select its product differentiation strategy (i.e., the optimal menu of (price, leadtime) combinations at which to offer each of its goods) to maximize its profitability. In more detail, the firm should choose state-dependent pricing and leadtime quotation strategies, as well as expediting and sequencing policies to maximize the long-run average revenue minus expediting costs.

---

[1]For 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 [14].

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. In terms of modelling, this paper is one of the first to address the joint dynamic pricing and leadtime control problem in a stochastic production environment, and it 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 now be captured through a relationship that is parametrized by the given leadtime vector but only varies as a function of the price menu, and the joint pricing and leadtime control problem reduces to one of pricing subject to leadtime guarantees, which is more tractable. Developing a revenue management solution for such a manufacturer requires an accurate, data-driven customer choice model, which leads to a tractable formulation. The third modelling contribution of this paper pertains to the model of customer choice behavior outlined in §5, which builds on extensive marketing research.

The multi-dimensional 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 [18, 19] 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 [26] for such a result in the context of revenue maximization for a single-product model using static pricing. Adopting this approach, the key analytical results of this paper are the following. We propose an approximating diffusion control problem in §4.1 that is based on a novel interpretation of the system parameters that captures the tension between capacity and the potential demand, and leads to a tractable but non-trivial limiting problem. This is solved by combining and extending results by Plambeck et.al. [31] and Ata et.al. [4]. While neither of these two papers involved any consideration of revenue maximization and/or pricing capability of some sort, our problem can be treated as a combination of the underlying problems in [31, 4]. Specifically, imitating results from [31] we derive the optimal sequencing and expediting controls (Proposition 1); the term optimal is used here in the context of the approximating diffusion model. The resulting multi-dimensional drift control problem is reduced to a one-dimensional one in terms of the workload process (Proposition 2),

which is subsequently solved be specializing to our setting results from [4] (Theorem 1).

The solution of the approximating diffusion control problem leads to intuitive and practically implementable policies for the original problem (see §3), as well as to several structural insights: 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 most 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. Finally, a set of numerical results illustrates the value of joint pricing and leadtime quotation control, as well as the performance of the proposed set of policies.

The structure of the remaining paper is as follows. This section concludes with a literature survey. §2 describes the mathematical model of this paper, and §3 summarizes our proposed solution, which is then justified through the analysis of an approximating diffusion control problem in §4. §5 describes a customer choice model that is suitable for the problem under consideration. §6 summarizes a set of numerical experiments that illustrate the effects of dynamic pricing and leadtime quotation on revenue performance, and offers some concluding remarks.

Literature survey. First, our paper is related to the literature on dynamic due-date and sequencing control. Roughly, this is divided in papers that develop efficient algorithms for mathematical programming formulations of such problems, and those emphasizing the stochastic nature of the production dynamics and either evaluate heuristics via simulation or solve for optimal policies in simple settings that often include probabilistic service level guarantees. Keskinocak and Tayur [22] provide an extensive review of this literature, with more emphasis on the former, while Baker [5] and Wein [37] review the literature emphasizing the latter. The majority of the work reviewed above assumes that the demand process is independent of any pricing and/or due-date decisions. Early work that incorporated the customer response to the firm's leadtime policy in a stochastic production setting was Duenyas and Hopp [13] and Duenyas [12], while Keskinocak et.al. [21] and Charnsirisakskul et.al. [11] provide deterministic optimization models in settings with delay- and price-sensitive demand, respectively.

The second body of research related to our paper focuses on static pricing and sequencing in queueing systems. One stream of work initiated with Naor [30] that includes Mendelson [28], Mendelson and Whang [29], and Van Mieghem [36] studies problems of social welfare optimization for price and delay sensitive customers in single ([30, 28]) and multi-product settings ([29, 36]). In an important recent paper, Afeche [1] considered the problem of revenue maximization for an

$M/M/1$ queue using static pricing and sequencing control in a market with two types of price and delay sensitive customers, and as such can be viewed as an analog to [29] in the context of revenue maximization. In these papers, as well as Maglaras and Zeevi [26] that established the economic optimality of the heavy-traffic regime in a single-product system under revenue maximization, "delay" refers to the steady-state waiting time that customers experience in the equilibrium regime that emerges given a set of prices, sequencing rule and a customer choice model.

Methodologically, our work builds on the literature on fluid and diffusion approximations of queueing systems. The asymptotic approximations for queues with state-dependent parameters that underlie our work were developed in Mandelbaum and Pats [27]. State-space collapse in Brownian control problems is explained in Harrison and Van Mieghem [20], and Ata et.al [4] address a class of diffusion control problems that includes the one we analyze in §4.3. The formulation of leadtime constraints as upper bounds on the respective queue lengths is from Plambeck et.al. [31] and Maglaras and Van Mieghem [25], and builds on Reiman's "snapshot principle" [34]. Overall, our formulation extends that in [31] to incorporate dynamic pricing (i.e., drift control) capability, and is solved using results from [20, 31, 4].

Other related papers are Plambeck [32] that studied a problem of static pricing and leadtime differentiation for two partially substitutable products, Maglaras [24] that looked at dynamic pricing and sequencing for a multi-product queue with price sensitive customers and holding costs incurred by the firm, and Ata [3] that focused on admission control for a multi-class system with leadtime guarantees and thin arrival streams. While [3] does not involve any pricing decisions, similar to our work its solution builds on [31] and [4]. Papers that include expediting or dual-source modes include Plambeck and Ward [33] and Bradley [8, 7]. The demand model we propose in §5 borrows from the marketing literature, see e.g., Bucklin and Gupta [10]; for an overview of demand models for revenue management see Talluri and van Ryzin [35].

2   Model Formulation

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, combining the operational controls of sequencing and expediting with the demand controls of pricing and leadtime quotation.

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 effectively

5

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 allows us to instead impose "hard" leadtime guarantees, i.e., $\epsilon_i = 0$ for all i (explained later on).

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 at time t. 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 $\mathcal{L}(d) = \{x \geq 0 : x = \lambda(p; d), \ p \in \mathcal{P}, \ d \in \mathbb{R}_+^I\} \subseteq \mathbb{R}_+^I$ 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. 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 $p_i^\infty(p_{-i}) \in \mathcal{P}$ such that $\lim_{p_i \to p_i^\infty(p_{-i})} \lambda_i(p_i, p_{-i}; d) = 0$; 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 λ 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 $\hat{\lambda}(d) := \text{argmax } \{r(\lambda; d) : \lambda \in \mathcal{L}(d)\}$. §5 describes a choice model that satisfies our assumptions and seems suitable for the type of problem considered in this paper.

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 and 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 ξ_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 λ is defined as $\rho := m \cdot \lambda$.

In addition to pricing, the firm also controls the operational decisions of order sequencing at the server, and order expediting. 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 $(T(t) : t \geq 0)$ with $T(0) = 0$, 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

$$\sum_i T_i(t) - \sum_i T_i(s) \leq t - s \quad \text{for } 0 \leq s \leq t < \infty. \tag{1}$$

The expediting policy captures 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. It is modelled as an I-dimensional process $(B(t) : t \geq 0)$ with $B(0) = 0$, 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.[2] 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 $c_1\mu_1 \geq c_2\mu_2 \geq \cdots \geq c_I\mu_I$. Finally, a control (λ, T, B) will be admissible if in addition to all of the above conditions it is non-anticipating, i.e., decisions at time t can only use information that has been made 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:

$$Q_i(t) = Q_i(0) + N_i\left(\int_0^t \lambda_i(s)ds\right) - S_i(T_i(t)) - B_i(t) \quad \text{for } i = 1, \dots, I. \tag{2}$$

Control problem formulation. The profit maximization problem for the stochastic queueing model is the following: choose admissible demand, sequencing and expediting policies (λ, T, B),

---

respectively, to maximize the long-run average expected profit given by

$$\lim_{t \to \infty} \frac{1}{t} \ \mathbb{E}\left[\int_0^t r(\lambda(s))ds - c \cdot B(t)\right] \tag{3}$$

subject to the leadtime constraints specified above.[3]

Discussion of modelling assumptions. The Poisson nature of the demand processes is needed to be able to justify the diffusion models used in §4 as appropriate limits under a broad class of state-dependent demand policies [27]. The one on general service time distributions is innocuous, since the diffusion analysis only uses its first and second moments. As in most papers on pricing in queues and revenue management, our model assumes that self-interested customers decide whether to place an order based solely on the price (and leadtime) vector at the time of their arrival; i.e., they are strategic in making purchase selections by explicitly or implicitly optimizing some form of a personal utility function, but not strategic in selecting the timing of their arrival in response to the firm's pricing strategy. This reduces the firm's pricing problem to one of optimal intensity control not involving a game-theoretic analysis; see Lariviere and Van Mieghem [23] for a discussion of this point and a justification of the Poisson arrival process assumption as the equilibrium of such a game for a related model. Expediting control capability is a common business practice that fits naturally within our tactical profit maximization problem. Analytically, it allows the firm to satisfy its leadtime constraints, e.g., by expediting orders whenever their sojourn time reaches their quoted leadtime; of course, this need not be optimal in terms of cost. In the context of the asymptotic analysis of this paper this could also be achieved by ejecting orders from a queue, exercising "hard" admission control (i.e., turn off a demand stream), or simply by a more "aggressive" use of pricing – in all cases the goal is to maintain the queue length below a certain threshold. The latter is the most complex case, since it is likely to involve discontinuous or infinite price controls.

3   The proposed solution

This section summarizes our proposed set of pricing, expediting and sequencing policies for the problem described above, which is based on a direct interpretation of the solution to an approximating diffusion control problem that we derive and analyze in §4.

---

[3]The formulation in (3) is derived using a standard result for intensity control problems (see Brémaud [9, §II.2]) from the primitive problem: choose p, T, B to maximize $\lim_{t \uparrow \infty} \frac{1}{t} \mathbb{E}[\int_0^t p(s) \cdot dA(s) - c \cdot B(t)]$, where $A_i(t) = N_i(\int_0^t \lambda_i(s)ds)$. We will not justify this point, since our subsequent analysis will not address (3) directly.

Given the original problem parameters, we first define

$$\bar{\lambda}(d) := \text{argmax}\left\{r(\lambda;d) \ : \ \sum_i \lambda_i/\mu_i = 1, \ \lambda \in \mathcal{L}(d)\right\} \tag{4}$$

to be the demand rate vector that maximizes the instantaneous revenue rate subject to the constraint that the server is fully utilized.[4] We assume that $\bar{\lambda}_i(d) > 0$ for all i, i.e., that it is optimal to produce all products in this deterministic planning problem. Also, recall that $\hat{\lambda}(d) := \text{argmax}\ \{r(\lambda;d) : \lambda \in \mathcal{L}(d)\}$ maximizes the instantaneous revenue rate with unconstrained capacity, and let $\hat{\Lambda} = \sum_i \hat{\lambda}_i(d)$ act as proxy of the total market potential for our problem. The vectors $\bar{\lambda}(d)$ and $\hat{\lambda}(d)$ play an important role in our subsequent analysis in a way that is reminiscent of other papers in revenue management; c.f., Gallego and van Ryzin [15] that show that the optimal demand control in a deterministic single-product model is static and given by the minimum between the rate that depletes capacity at the end of the planning horizon and the rate that maximizes revenues in the absence of the capacity constraint ($\bar{\lambda}(d)$ and $\hat{\lambda}(d)$, respectively). Our analysis focuses on optimizing the behavior of the second order stochastic fluctuations around the deterministic and static solution in (4) (much like the Central Limit Theorem characterizes the error term around the mean of i.i.d. random variables). Given that both $\bar{\lambda}(d)$ and $\hat{\lambda}(d)$ are of order $\hat{\Lambda}$, the magnitude of these fluctuations is proportional to $\sqrt{\hat{\Lambda}}$ (cf. (5) and $\$$ 4.1).

<u>Pricing</u>: 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

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

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

$$\lambda^*(w;d) \ = \ \text{argmax}\left\{r(\lambda;d) \ : \ \lambda \cdot m \leq \rho^*(w), \ \lambda \in \mathcal{L}(d)\right\} \tag{6}$$

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

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

$$b_i' = \bar{\lambda}_i(d)d_i - \delta_i,$$

---

[4]We assume that this problem is feasible for the choice of leadtime vector d under consideration, i.e., that there exists a vector of non-negative prices (including $p = 0$) for which the resulting demand will utilize all the capacity.

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

$$i^* = \operatorname*{argmax}_i \frac{Q_i(t)}{b_i'};$$

this is the "least relative slack" policy of Plambeck et.al. [31].

Expediting: Orders are expedited when the total workload reaches $\tilde{W} = m \cdot b'$ according to a rule that gives priority to class I then class $I-1$ (if no class I orders are in queue) and so on. This is the natural interpretation of the diffusion policy derived in §4.2 in the context of the original problem at hand, baring in mind that products are labelled in a way that $c_1\mu_1 \geq c_2\mu_2 \geq \cdots \geq c_I\mu_I$.

Before proceeding with the diffusion analysis of §4 to justify the above policy recommendation, we conclude this section with a few comments on its structure and general properties.

i. *Leadtime constraint formulation & sequencing:* The parameter b_i' serves as a proxy for the number of class i arrivals in d_i time units, and thus maintaining the queue lengths below their respective thresholds would tend to imply that the corresponding leadtime guarantees are met with high probability; c.f., [25, 31]. The threshold b_i' is derived from the diffusion model (see (19)), appropriately adjusted through δ_i to correct for the stochastic variability of the arrival and service time processes, and the state-dependent nature of the demand rate; i.e., since $\lambda_i^*(w; d)$ is decreasing in the workload w, $\bar{\lambda}_i(d)d_i$ is an overestimate of the number of arrivals in d_i time units.

ii. *Workload dependence and time-scale of price changes:* The proposed sequencing policy tries to distribute the workload in fixed proportions across the various queues, therefore making $W(t)$ an accurate proxy for the system state; this equivalence is exact in the diffusion model. It therefore suffices to restrict attention to pricing and expediting policies that are functions of the workload. This simplifies analysis and has an important implication on the relative time-scale for these decisions. Specifically, known results for queues operating in heavy-traffic predict that order interarrival and service times are much shorter than the typical queueing times encountered in the system, which in turn are much shorter than the time required for the workload (and the respective queue lengths) to experience significant fluctuations. Pricing changes and expediting decisions occur on the slowest time-scale on which the system workload evolves, which is practically appealing. As an example, orders may be arriving every 30 minutes, queueing delays and leadtimes may be of order of a week, and the workload (and the prices) may fluctuate on a monthly basis.

iii. *Leadtime control:* The dynamic leadtime control decisions are effectively captured in the demand control $\lambda(W(t); d)$ that specifies how to optimally divert demand from one leadtime class to another. This will become clearer in §5 where we study in more detail a particular demand model that is suitable for the problem under consideration.

4   Justification of proposed policy via analysis of a diffusion model

We start by briefly developing the approximating diffusion control problem, referring the reader to [18, 19] and Mandelbaum and Pats [27] for background and more detailed expositions. Next, we show that this multi-dimensional formulation can be reduced to a one-dimensional problem in terms of the aggregate system workload, which we solve in closed form.

4.1   Formulation of an approximating diffusion model

Parameter regime: In constructing the approximating control problem, we start by identifying the appropriate parameter regime that gives rise to the diffusion model under the appropriate heavy-traffic scaling. The first step is to extract a set of normalized parameters from the original problem data (in (7)-(8)) that are then used in defining a sequence of systems in (9)-(10) that gives rise to our diffusion approximation and is closely related to the original problem specified in §2.

Recall the definitions of $\hat{\lambda}(d)$, $\hat{\Lambda}$ and $\bar{\lambda}(d)$ from §3. Let $\bar{\rho}_i = \bar{\lambda}_i(d)/\mu_i$ and define $\hat{\mu}_i := \hat{\lambda}_i(d)/\bar{\rho}_i$ for all i to be the "nominal" processing rates required to serve the revenue maximizing demand rate vector $\hat{\lambda}(d)$ according to the fractional allocations defined through $\bar{\rho}$. The difference between μ and $\hat{\mu}$ measures the "imbalance" between the firm's capacity and the processing rate vector that would be needed to maximize the instantaneous revenue rate. This is expressed in the form

$$\mu = \hat{\mu} - \beta\sqrt{\hat{\Lambda}} \quad \text{where} \quad \beta_i := \frac{\hat{\lambda}_i(d) - \bar{\lambda}_i(d)}{\bar{\rho}_i\sqrt{\hat{\Lambda}}} \quad \forall \, i, \tag{7}$$

where $\hat{\Lambda}$ is used as a proxy for the "size" of the system.

The next step is to embed the original system in an appropriate sequence of systems whose asymptotic behavior we will henceforth analyze. This is done as follows. First, we normalize the demand function, leadtime bound, and the expediting cost according to

$$\tilde{\lambda}(\cdot; \cdot) = \lambda(\cdot; \cdot)/\hat{\Lambda}, \quad \tilde{d} = d\sqrt{\hat{\Lambda}} \quad \text{and} \quad \tilde{c} = c\sqrt{\hat{\Lambda}}. \tag{8}$$

Second, we define a sequence of systems indexed by n with parameters scaling according to

$$\lambda^n(\cdot; \cdot) = n\tilde{\lambda}(\cdot; \cdot), \quad d^n = \tilde{d}/\sqrt{n}, \quad c^n = \tilde{c}/\sqrt{n}, \quad \text{and} \quad \mu^n = \hat{\mu}^n - \beta\sqrt{\hat{\Lambda}^n}, \tag{9}$$

where $r^n(\cdot; \cdot)$ is the revenue function associated with the demand relation $\lambda^n(\cdot; \cdot)$, $\mathcal{L}^n(d^n) = n\tilde{\mathcal{L}}(d^n)$

and $\tilde{\mathcal{L}}(\cdot) = \{x : x\hat{\Lambda} \in \mathcal{L}(\cdot)\}$, and

$$\hat{\lambda}^n(d^n) := \operatorname{argmax}\{r^n(\lambda; d^n) : \lambda \in \mathcal{L}^n(d^n)\}, \quad \hat{\mu}_i^n = \hat{\lambda}_i^n(d^n)/\bar{\rho}_i \quad \text{and} \quad \hat{\Lambda}^n = \sum_i \hat{\lambda}_i^n(d^n). \tag{10}$$

[A superscript n is attached to all quantities associated with the nth system; this notation is only used in §4.1 in constructing the approximating diffusion model.] The above scaling is interpreted as follows: we will consider a sequence of problems of increasing market size served by systems with appropriately increasing processing capacity, where the latter is selected so as to keep the imbalance between the processing capability vector and the (capacity unconstrained) revenue maximizing demand rate vector constant as measured by the vector β. In addition, the leadtimes d^n and expediting costs c^n are scaled in a way that is consistent with known insights from the behavior of queues in heavy-traffic; see Plambeck et.al. [31] for a rigorous justification of these scalings, and the first part of the appendix to this paper for some related commentary. Note that for $n = \hat{\Lambda}$ we recover the parameters of the system of original interest; i.e, we have embedded the problem described in the previous section to a sequence of problems described via (9)-(10). Finally, for future reference, we define

$$\tilde{\lambda} := \lim_{n\uparrow\infty} \frac{1}{n}\hat{\lambda}^n(\tilde{d}/\sqrt{n}) = \lim_{n\uparrow\infty} \hat{\tilde{\lambda}}(\tilde{d}/\sqrt{n}), \tag{11}$$

which is assumed to exist and be componentwise strictly positive, and set $\tilde{\mu}_i = \tilde{\lambda}_i/\bar{\rho}_i$ for all i. We also let $\kappa_i^n = \beta_i\bar{\rho}_i\sqrt{\hat{\Lambda}^n/n}$ and set

$$\lim_{n\uparrow\infty} \kappa_i^n = \kappa_i \quad \text{and} \quad \lim_{n\uparrow\infty} -\frac{1}{2}\nabla^2 \tilde{r}(\hat{\tilde{\lambda}}(d^n); d^n) = A, \tag{12}$$

where A is assumed to exist and be a positive definite matrix (this follows from the strict concavity assumption of the revenue function).

Formulation: The approximating diffusion model for our problem is that of Plambeck et.al. [31] with two modifications to account for the dynamic pricing aspect of our model that makes the demand rate state-dependent, and the fact that orders are not blocked but are expedited. The former affects the dynamics of the diffusion model and the objective function of the associated control problem, while the latter turns out to be inconsequential.

Focusing on the heavy-traffic regime outlined earlier, we consider demand controls of the form

$$\lambda_i^n(t) = \left[\bar{\rho}_i\mu_i^n - \sqrt{n}\theta_i(t)\right]^+ \tag{13}$$

where $\theta_i(t) \in [-K, K]$ is the dynamic demand rate adjustment at time t, which to minimize

12

technical complexity and comply later on with the restrictions imposed in Ata et.al. [4] is assumed to be bounded by a large constant K. Plugging into (2) we get that

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

where $S_i^n(t)$ is the cumulative number of service completions if the server spends t time units processing product i orders with a nominal service rate μ_i^n. The scaled service times can be obtained from the original sequence of service time random variables with rescaling by the appropriate factor implied by (9). For each product i, define $V_i^n(t) = \bar{\rho}_i t - T_i^n(t)$ to measure the deviation between the cumulative time allocated into processing class i orders up to time t and the nominal service requirement for that product, and note that the cumulative idleness up to time t is $I^n(t) = \sum_i V_i^n(t)$. Building on knwon insights for the heavy-traffic behavior of the $M/M/1$ queue (see the appendix), we define scaled copies of the system processes (Q^n, V^n, B^n) according to

$$Z^n(t) := \frac{Q^n(t)}{\sqrt{n}}, \quad Y^n(t) := \sqrt{n}V^n(t) \quad \text{and} \quad D^n(t) := \frac{B^n(t)}{\sqrt{n}}. \tag{14}$$

For large n, the Strong Approximation Theorem for the cumulative arrival and service completion processes [16, Theorem 5] gives that for all i

$$N_i\left(\int_0^t \lambda_i^n(s)ds\right) = \mu_i^n \bar{\rho}_i t - \sqrt{n}\int_0^t \theta_i(s)ds + \sqrt{n}\sigma_{a,i}X_{a,i}(t) + o(\sqrt{n}),$$

and

$$S_i^n(T_i^n(t)) = \mu_i^n \bar{\rho}_i t - \mu_i^n V_i^n(t) + \sqrt{n}\sigma_{s,i}X_{s,i}(t) + o(\sqrt{n}),$$

where the notation $f(x) = o(g(x))$ denotes that $f(x)/g(x) \to 0$ as $x \uparrow \infty$, $X_{a,i}$ and $X_{s,i}$ are independent standard Brownian motions, $\sigma_{a,i}^2 = \tilde{\lambda}_i$ and $\sigma_{s,i}^2 = \tilde{\lambda}_i\xi_i$. (Recall that the ξ_i's are the squared coefficients of variation of the service time random variables.) Adding terms and using the definitions of (Z^n, Y^n, D^n) suggests the following diffusion model:

$$dZ(t) = -\theta(t)dt + \Sigma dX(t) + MdY(t) - dD(t), \quad Z(0) = z \tag{15}$$

$$L(t) = \sum_i Y_i(t), \quad L(\cdot) \text{ is continuous, non-decreasing with } L(0) = 0 \tag{16}$$

$$D(\cdot) \text{ is continuous, non-decreasing with } D(0) = 0 \tag{17}$$

$$Z(t) \geq 0, \ \forall t \geq 0 \quad \text{and } Y, D \text{ are non-anticipating with respect to X} \tag{18}$$

where $M = \mathbf{diag}(\tilde{\mu}_1, \ldots, \tilde{\mu}_I)$, X is an I-dimensional standard Brownian motion, $\Sigma^2 = \mathbf{diag}((1 + \xi_1)\tilde{\lambda}_1, \ldots, (1+\xi_I)\tilde{\lambda}_I)$, and $X(0) = 0$ almost surely on some filtered probability space $(\Omega, \mathcal{F}, \mathbb{P}; \mathcal{F}_t, t \geq$

0). As in [31], the leadtime constraints take the form

$$Z(t) \le b \text{ for } t \ge 0, \quad \text{where } b_i = \tilde{\lambda}_i \tilde{d}_i \ \forall \ i. \tag{19}$$

The interpretation of the various processes follows from the scalings in (14): Z represents the queue length process, D is the expediting policy, Y is the allocation control (measuring deviations from the nominal allocation), and L represents the scaled cumulative idleness.

To derive the appropriate performance criterion for this model, we note that the expediting cost is given by $c^n \cdot B^n(t) = \tilde{c} \cdot D^n(t)$, while to express the revenue term we rewrite $\lambda_i^n(\cdot)$ as

$$\lambda_i^n(t) = \left[\bar{\rho}_i \mu_i^n - \sqrt{n} \theta_i(t) \right]^+ = \left[\hat{\lambda}_i^n(d^n) - \sqrt{n} \kappa_i^n - \sqrt{n} \theta_i(t) \right]^+.$$

Let $\tilde{r}(\cdot; \cdot)$ be the revenue function associated with the demand function $\tilde{\lambda}(\cdot; \cdot)$. Then,

$$r^n(\lambda^n(t); d^n) = n\tilde{r}\left(\left[\tilde{\lambda}^*(d^n) - \frac{\kappa^n + \theta(t)}{\sqrt{n}} \right]^+ ; d^n \right)$$

$$= n\tilde{r}(\tilde{\lambda}^*(d^n); d^n) + \frac{1}{2}[\kappa^n + \theta(t)] \cdot \nabla^2 \tilde{r}(\tilde{\lambda}^*(d^n); d^n)[\kappa^n + \theta(t)] + o(1),$$

where the second expression is obtained via a Taylor expansion of $\tilde{r}(\cdot; d^n)$ around $\tilde{\lambda}^*(d^n)$, and the first order term is missing since $\nabla \tilde{r}(\tilde{\lambda}^*(d^n); d^n) = 0$ by the definition of $\tilde{\lambda}^*(d^n)$. Using the definition of κ, A in (12), and restricting attention to Markovian, stationary, bounded drift controls, the preceding analysis suggests the following diffusion control problem: choose a non-anticipating measurable drift function $\theta(t) \in [-K, K]^I$ for all $t \ge 0$, and non-anticipating RCLL allocation and expediting policies Y and D to minimize

$$\limsup_{t \to \infty} \frac{1}{t} \mathbb{E}\left[\int_0^t \{2\kappa \cdot A\theta(s) + \theta(s) \cdot A\theta(s)\} \, ds + \tilde{c} \cdot D(t) \right] \tag{20}$$

subject to (15)-(19).

4.2 Reduction to the equivalent workload formulation

The first step in analyzing (15)-(20) establishes that the optimal pair of allocation and expediting policies (Y, D) derived in Plambeck et.al. [31] is also optimal for our problem that incorporates dynamic drift rate control capability. [The model analyzed in [31] involved admission rather than expediting decisions, but the two are analytically equivalent.] The optimal (Y, D) yield a one-dimensional equivalent workload formulation for our problem (see Harrison and Van Mieghem [20] for background), which will be used to derive the optimal dynamic drift control $\theta(\cdot)$.

14

We start with some background material. The system workload process is defined by $W(t) := \tilde{m} \cdot Z(t)$, where $\tilde{m}_i = 1/\tilde{\mu}_i$. The workload dynamics are given by:

$$dW(t) = -\tilde{m} \cdot \theta(t)dt + \sigma_w dX_w(t) + dL(t) - dU(t) \tag{21}$$

where L satisfies (16),

$$U(t) = \tilde{m} \cdot D(t), \quad U(\cdot) \text{ is continuous and non-decreasing, } U(0) = 0, \tag{22}$$

D satisfies (17), $X_w(t)$ is a standard Brownian motion, and $\sigma_w^2 = \sum_i (1 + \xi_i)\tilde{m}_i^2 \tilde{\lambda}_i$. Note that $Z(t) \leq b$ implies that

$$W(t) \in [0, \tilde{w}] \ \ t \geq 0 \quad \text{for} \quad \tilde{w} := \tilde{m} \cdot b. \tag{23}$$

The next result specializes some of the results of Plambeck et.al [31] to our model. Specifically, we show that it is optimal to (i) only expedite orders of the cheapest class I when the workload $W(t)$ reaches its upper bound $\tilde{w}$ imposed by the leadtime constraints, and (ii) schedule orders according to the "least slack policy" that maintains the relative backlogs defined as

$$\eta_i(t) := \frac{Z_i(t)}{b_i} \ \ t \geq 0, \ \forall \ i, \tag{24}$$

the same for all classes; i.e., this policy gives priority to the class that is closest to violating its leadtime constraint (which in the diffusion model corresponds to $Z_i(t) > b_i$).

Proposition 1 *Fix any admissible drift control $(\theta(t), t \geq 0)$ and consider the problem of choosing (Y, D) to minimize*

$$\limsup_{t \to \infty} \frac{1}{t} \ \mathbb{E}\left[\tilde{c} \cdot D(t)\right], \tag{25}$$

subject to the constraints (15)-(19). Then, the following policy is optimal:

$$Y_i(t) := \frac{\tilde{m}_i b_i}{\tilde{w}} L(t), \ \forall \ i, \tag{26}$$

and

$$D_i(t) = 0 \ \ \forall \ i \neq I, \quad D_I(t) = \tilde{\mu}_I U(t), \tag{27}$$

where L, U are continuous, non-decreasing processes, with $L(0) = U(0) = 0$ such that

$$\int_0^t 1_{\{W(s)>0\}} dL(s) = 0 \quad and \quad \int_0^t 1_{\{W(s)<\tilde{w}\}} dU(s) = 0 \ \ t \geq 0, \tag{28}$$

where $W(t)$ satisfies condition (21) and the constraint (23).

Proof: We first establish the feasibility of Y, D. Straightforward algebraic manipulations involving (15), (16) and (21) and (22) show that for the control $Y(t)$ given in (26), $\eta_i(t) = \nu(t)$ for $t \geq 0$ and all products i, where

$$W(t) = \tilde{m} \cdot Z(t) = \nu(t)(\tilde{m} \cdot b) = \nu(t)\tilde{w} \quad \Rightarrow \quad \nu(t) = \frac{W(t)}{\tilde{w}} \quad t \geq 0. \tag{29}$$

Given (24) and (29) and the fact that $U(t)$ keeps $W(t) \leq \tilde{w}$ it follows that $Z(t) \leq b$ for $t \geq 0$. Also, by construction of $L(t)$ it follows that $W(t) \geq 0$, and therefore that $Z(t) \geq 0$ for $t \geq 0$. Therefore, Y, D are feasible for (15)-(19). Second, the minimality of $D(t)$ follows from the properties of the two-sided regulator (see Harrison [17, §2.4]). $\blacksquare$

Note that with $\theta(t)$ fixed, the first term in (20) is a constant, and thus the performance criterion reduces to (25). Under the control (Y, D), $Z(t) = \frac{b}{\tilde{w}}W(t)$ and $\tilde{c} \cdot D(t) = \tilde{c}_I\tilde{\mu}_I U(t)$, and the approximating diffusion control problem reduces to one selecting the drift control θ to minimize

$$\limsup_{t\to\infty} \frac{1}{t} \, \mathbb{E}\left[\int_0^t \{2\kappa \cdot A\theta(s) + \theta(s) \cdot A\theta(s)\} \, ds + \tilde{c}_I\tilde{\mu}_I U(t)\right] \tag{30}$$

subject to (21), the workload constraint (23), conditions (28) that identify the processes L, U as the associated unique two-sided regulator (see Harrison [17, §2.4]) and (26), (27) and (29) that specify Y, D, Z, respectively.

Since the drift control θ only affects the system dynamics through its aggregate value $\psi := \tilde{m} \cdot \theta$, the above problem can be further simplified to one expressed in terms of the one-dimensional control $\psi \in [-K', K']$. Specifically, letting

$$\theta^* = \operatorname{argmin}\{2\kappa \cdot A\theta + \theta \cdot A\theta \; : \; \tilde{m} \cdot \theta = \psi\} = \tilde{f}(\psi) := \frac{A^{-1}\tilde{m}}{\tilde{m} \cdot A^{-1}\tilde{m}}(\psi + \tilde{m}\kappa) - \kappa, \tag{31}$$

the revenue loss at θ^* becomes

$$2\kappa \cdot A\theta^* + \theta^* \cdot A\theta^* = \frac{(\psi + \tilde{m} \cdot \kappa)^2}{\tilde{m} \cdot A^{-1}\tilde{m}} + \kappa \cdot A\kappa. \tag{32}$$

Using these definitions and removing the constant term $\kappa \cdot A\kappa$ of (32) from the objective function, we can rewrite the diffusion control problem in the form:

Proposition 2 *The following control problem is equivalent to (15)-(20): choose a non-anticipating measurable function $\psi(t) \in [-K', K']$ for $t \geq 0$ to minimize*

$$\limsup_{t\to\infty} \frac{1}{t} \, \mathbb{E}\left[\int_0^t \left\{\frac{[\psi(s) + \tilde{m} \cdot \kappa]^2}{\tilde{m} \cdot A^{-1}\tilde{m}}\right\} ds + \tilde{c}_I\tilde{\mu}_I U(t)\right]; \tag{33}$$

subject to

$$dW(t) = -\psi(t)dt + \sigma_w dX_w(t) + dL(t) - dU(t), \tag{34}$$

and (23) and (28). Conditions (26), (27), (29) and (31) define the optimal (Y, D, Z, θ), respectively.

4.3 Solution of the equivalent workload formulation

The equivalent workload formulation is a one-dimensional drift control problem for a diffusion that is constrained to lie in the interval $[0, \tilde{w}]$. We will solve the problem described immediately above using results derived in Ata et.al. [4], for which we need to restrict attention to Markovian, stationary, and bounded controls of the form $\psi : [0, \tilde{w}] \to [-K', K']$. To simplify notation in the remainder of this section we let $\alpha_w = (\tilde{m} \cdot A^{-1}\tilde{m})^{-1} > 0$ and $\kappa_w = \tilde{m} \cdot \kappa$, which together with the Markovian structure of the drift controls allows us to rewrite the objective as

$$\limsup_{t \to \infty} \frac{1}{t} \left[\mathbb{E} \int_0^t \left\{ \alpha_w [\psi(W(s)) + \kappa_w]^2 \right\} ds + \tilde{c}_I \tilde{\mu}_I U(t) \right]. \tag{35}$$

Let

$$h(\tilde{c}_I \tilde{\mu}_I, \alpha_w, \kappa_w, \tilde{w}, \sigma_w) := \tilde{c}_I \tilde{\mu}_I - \left[2\alpha_w \kappa_w - \left(\frac{\tilde{w}}{2\alpha_w \sigma_w^2} + \frac{1}{2\alpha_w \kappa_w} \right)^{-1} \right].$$

Theorem 1 *Consider the problem of selecting a non-anticipating measurable function $\psi : [0, \tilde{w}] \to [-K', K']$ to minimize (35) subject to (34), (23) and (28). Then, if $h(\tilde{c}_I \tilde{\mu}_I, \alpha_w, \kappa_w, \tilde{w}, \sigma_w) = 0$,*

$$\psi^*(w) = - \left(\frac{w}{\sigma_w^2} + \frac{1}{\kappa_w} \right)^{-1}, \tag{36}$$

if $h(\tilde{c}_I \tilde{\mu}_I, \alpha_w, \kappa_w, \tilde{w}, \sigma_w) > 0$,

$$\psi^*(w) = \sqrt{\frac{\zeta_1}{\alpha_w}} \tan \left[\frac{w}{\sigma_w^2} \sqrt{\frac{\zeta_1}{\alpha_w}} - \arctan \left(\kappa_w \sqrt{\frac{\alpha_w}{\zeta_1}} \right) \right], \tag{37}$$

where ζ_1 is the unique positive solution of (50), and otherwise if $h(\tilde{c}_I \tilde{\mu}_I, \alpha_w, \kappa_w, \tilde{w}, \sigma_w) < 0$,

$$\psi^*(w) = \sqrt{\frac{\zeta_2}{\alpha_w}} - 2\sqrt{\frac{\zeta_2}{\alpha_w}} \left[1 - \exp \left\{ -\frac{2w}{\sigma^2} \sqrt{\frac{\zeta_2}{\alpha_w}} + C \right\} \right]^{-1}, \tag{38}$$

where $C = \ln \left(\frac{\sqrt{\zeta_2/\alpha_w} - \kappa_w}{\sqrt{\zeta_2/\alpha_w} + \kappa_w} \right)$ and ζ_2 is the unique solution of (53) that lies in $(0, \alpha_w \kappa_w^2)$. The optimal product-level drift rate is given by

$$\theta^*(w) = \tilde{f}(\psi^*(w)) = \alpha_w A^{-1}\tilde{m}(\psi^*(w) + \tilde{m}\kappa) - \kappa. \tag{39}$$

17

Remark: We note that ψ^* is monotonically increasing in w in all three cases above. This implies that $\theta(w)$ is increasing in w, and in light of (13) that the proposed demand control $\lambda(Q(t))$ is a decreasing function of the system workload. We also note that the expression in (37) corresponds to the case where the expediting cost $\tilde{c}_I \tilde{\mu}_I$ is large, whereas expression (38) is for the case where expediting is cheap. Finally, it is now easy to connect the the optimal ψ, Y, D extracted above with the proposed policy of §3. [The proof of the Theorem is relegated to the appendix.]

4.4   Optimal static pricing solution

The numerical experiments in §6 will contrast the proposed solution against one that uses static pricing. This amounts to selecting the constant vector $\theta \in \mathbb{R}^I$ to minimize

$$\limsup_{t \to \infty} \frac{1}{t} \left[\mathbb{E} \int_0^t \{2\kappa \cdot A\theta + \theta \cdot A\theta\} \, ds + \tilde{c} \cdot D(t) \right] \tag{40}$$

subject to (15)-(19). Using Propositions 1 and 2, this is reduced to the problem

$$\min_{\psi \in \mathbb{R}} \left\{ \alpha_w [\psi + \kappa_w]^2 + \tilde{c}_I \tilde{\mu}_I \limsup_{t \to \infty} \mathbb{E} \left[\frac{1}{t} U(t) \right] \right\} \tag{41}$$

subject to (34), (23) and (28). That is, the optimal allocation and expediting policies Y, D are the same with those for the dynamic drift control problem. The workload process evolves like a Brownian motion with infinitesimal drift ψ and infinitesimal variance σ_w^2 in the interval $[0, \tilde{w}]$, with exponential steady-state distribution with mean $\sigma_w^2/(2\psi)$. This leads to the optimization problem

$$\min_{\psi \in \mathbb{R}} \left\{ \alpha_w [\psi + \kappa_w]^2 + \frac{\tilde{c}_I \tilde{\mu}_I \psi}{e^{2\psi\tilde{w}/\sigma_w^2} - 1} \right\}, \tag{42}$$

where the expression for the second term above is given in Harrison [17, pg.88-90]. This result is summarized in the following theorem.

Theorem 2 *Consider the problem of selecting a constant vector $\theta \in \mathbb{R}^I$ to minimize (42) subject to (15)-(19). Let ψ^* be the minimizer of (42). The optimal drift vector is given by*

$$\theta^* = \alpha_w A^{-1} \tilde{m}(\psi^* + \tilde{m}\kappa) - \kappa.$$

5 A choice model for joint pricing and leadtime control

This section proposes a customer choice model that satisfies the assumptions imposed in §2, but more importantly seems suitable for the problem considered in this paper. Specifically, the model described below builds on a framework that has been used extensively in the marketing literature (see Bucklin and Gupta [10]), and postulates that customers make their purchase selection in two stages: first, they decide whether to buy any of the products within a category or segment, and second, if they choose to buy in the first stage, they then proceed to select a product from that category – these are referred to as "purchase incidence" and "product or brand choice," respectively; see [10] for a discussion of such models and their practical use. From our viewpoint, this model has the essential property to be able to capture the substitution effects among otherwise identical products offered at different price and leadtime combinations, while maintaining analytical and numerical tractability and being suitable for calibration using real data as indicated by the associated voluminous marketing literature.

As explained in §2, a product corresponds to a (type of good, leadtime) combination. To simplify the exposition we first describe the choice model for the case of one good offered at multiple leadtimes, and then extend it to consider many goods offered at multiple leadtimes each.

One good offered at multiple (price, leadtime) combinations: Following Bucklin and Gupta [10], we model the probability that a customer will buy one of the products (the "purchase incidence") using a binary logit function

$$P(\text{inc}) = \frac{e^{V(p,d)}}{1 + e^{V(p,d)}}, \quad \text{where} \quad V(p,d) = \gamma_0 + \gamma_1 \log\left(\sum_i e^{-b_1 p_i - b_2 d_i}\right). \tag{43}$$

$V(p,d)$ corresponds to the deterministic component of the purchase utility from all offered products specified through p, d. The constants $\gamma_0, \gamma_1, b_1, b_2$ are meant to be calibrated from observed data (see [10]). This purchase incidence probability is equivalent to saying that each arriving customer assigns a utility $V(p,d) + \epsilon$ to the offered group of products, where ϵ is an i.i.d. random component that differentiates potential customers, and follows a logistic distribution with shape parameter equal to one; the effect of different shape parameters can be rolled into $\gamma_0, \gamma_1, b_1, b_2$.

Each arriving customer also has a random delay sensitivity parameter χ for the offered good, which is assumed to be drawn from a continuous distribution with finite support, is independent of ϵ and i.i.d. across customers. Given that an arriving customer decides to purchase a product, she makes her selection to minimize their cost given by $p_i + \chi d_i$, i.e.,

$$\mathbb{P}(i \mid \text{inc}) = \mathbb{P}\left(p_i + \chi d_i \leq p_j + \chi d_j, \ \forall j \neq i\right); \tag{44}$$

the form of (44) captures the price-delay trade-off faced by each customer, and differs from the Multinomial Logit model used in [10]. Assuming that potential customers arrive according to a Poisson process with rate Λ, products are labelled in such a way that $d_1 < d_2 < \cdots < d_I$, and that prices are ordered in reverse, i.e., $p_1 \geq p_2 \geq \cdots \geq p_I$, we get that

$$\lambda_i(p;d) \;=\; \Lambda \cdot \mathbb{P}(\text{inc}) \cdot \mathbb{P}(i \mid \text{inc}) \;=\; \Lambda \cdot \frac{e^{V(p,d)}}{1 + e^{V(p,d)}} \cdot \mathbb{P}\left(\max_{j>i} \frac{p_i - p_j}{d_j - d_i} \;\leq\; \chi \;\leq\; \min_{k<i} \frac{p_k - p_i}{d_i - d_k} \right).$$

Many goods offered at multiple (price, leadtime) combinations: The above model can be extended to allow for many goods offered at potentially multiple (price, leadtime) combinations by incorporating the decision of which good to purchase in the incidence probability, leaving unchanged the second decision stage where a customer selects which product option of a particular good to purchase. Specifically, suppose that there are K goods, with $K < I$. Define a $K \times I$ constituency matrix C such that $C(k,i) = 1$ if product i corresponds to good k, and $C(k,i) = 0$ otherwise, and let $\mathcal{C}(k) = \{i : C(k,i) = 1\}$ to be the set of products that correspond to good k. Let

$$V^k(p,d) = \gamma_0^k + \gamma_1^k \log\left(\sum_{i \in \mathcal{C}(k)} e^{-b_1^k p_i - b_2^k d_i}\right),$$

denote the purchase utility from good k products, and the constants $\gamma_0^k, \gamma_i^k, b_1^k, b_2^k$ are meant to have been calibrated from observed data. Assume that a customer's net purchase utility for good k products is $V^k(p,d) + \epsilon^k$, where the ϵ^k's are i.i.d. across goods, Gumbell distributed random variables with shape parameter one. The incidence probability, which now reduces to the decision of which good to purchase, if any, is computed using the Multinomial Logit Model [35, §7.2]:

$$P(\text{select good } k) = \frac{e^{V^k(p,d)}}{1 + \sum_j e^{V^j(p,d)}}.$$

The product choice is done according to (44), specialized only to products in the set $\mathcal{C}(k)$.

A structural property of this demand model and its impact on leadtime control: An important step in implementing the policy described in §3 is the computation of the optimal demand vector given a target aggregate traffic intensity $\rho^* = 1 - \psi^*/\sqrt{\hat{\Lambda}}$; see §3 and Theorem 1. Simple but long algebraic manipulations show that in the parameter regime of interest in this paper, i.e., where Λ and μ are large, the solution to the problem for the single good case

$$\max_p \; \left\{ \sum_i p_i \lambda_i(p;d) \; : \; \sum_i \lambda_i(p;d) = \mu(1 - \psi/\sqrt{\hat{\Lambda}}) \right\},$$

is of the form

$$p_i = \bar{p} + \frac{\pi_i}{\sqrt{\hat{\Lambda}}} + \frac{z(\psi)}{\sqrt{\hat{\Lambda}}} + o(1/\sqrt{\hat{\Lambda}}),$$

20

where $\bar{p} = p(\bar{\lambda}(d); d)$, and $\pi_i, z(\psi) \in \mathbb{R}$. A similar result can be proved for multiple goods offered at different price, leadtime combinations, in which case the terms $z^k(\psi)$ will depend on the good k.

This expression has an important implication on the firm's "leadtime control policy." Specifically, for such a pricing policy $(p_i - p_j) = (\pi_i - \pi_j)/\sqrt{\hat{\Lambda}}$ for all ψ, which plugging into (44) gives that $\mathbb{P}(i \mid \text{inc})$ is independent of ψ! From a managerial perspective this provides a very insightful result: the firm adjusts its nominal price level through $z(\psi)/\sqrt{\hat{\Lambda}}$ to modulate the aggregate order volume placed with the system, while keeping the fractions of the total order flow that choose each leadtime option constant. That is, it does not choose to divert demand from one leadtime to another as the system gets congested, but rather scale down the demand for all products by a common factor by adjusting its price to affect the incidence probability.

Comments on modelling price and delay sensitive demand: An alternative to the two-stage decision process of our model would consider customers arrive with a random valuation v and delay sensitivity parameter χ, and make their purchase decisions in one stage according to

$$\lambda_i(p; d) = \Lambda \mathbb{P}(v - p_i - \chi_i d_i \geq 0, \ p_i + \chi d_i \leq p_j + \chi d_j \ \forall \ j \neq i).$$

While this may appear to be a more direct and natural model of demand, it is harder to analyze because evaluating the above expression involves the joint distribution of (v, χ). Moreover, its complexity increases significantly when one considers multiple goods offered in different (price, leadtime) options, where customers arrive with different valuations for each of these goods. The hierarchical decision approach of our model assumes that problem away by restricting attention to a structure where the probability that a customer selects a particular service is given by the product of two probabilities, where one depends on the valuation and the other on the delay sensitivity. Apart from its inherent tractability, extensive studies reported in the marketing literature indicate its versatility in capturing customer demand in diverse and complicated settings.

Instead of atomistic choice models, one could employ an aggregated demand relation, such as the linear, $\lambda(p; d) = \Lambda - Hp - Fd$ or the exponential, $\lambda_i(p; d) = \Lambda_i e^{-h_i \cdot p - f_i \cdot d}$, for appropriate Λ, H, F; see Talluri and van Ryzin [35, §7.3]. In single-product settings with customers that either have a random valuation or a random delay sensitivity parameter, but not both, there is a one-to-one mapping between aggregate demand functions and distributional forms for the random customer parameter; the linear model correspond to uniformly distributed parameters, the exponential model corresponds to exponentially distributed parameters, and so on (see [35, §7.3.1.2]).

For customers that have two-dimensional types (i.e., two random parameters that characterize them, such as their valuation and delay sensitivity) and in settings with multiple products that complicate the choice behavior, such a connection is hard to establish. In part, such demand

functions that are common practice in the revenue management literature, they seem to be better suited in modelling demand for partially substitutable products, making them less suitable for our problem. As an illustration of this point consider an example with two identical products offered at different (price, leadtime) combinations as we vary the respective leadtimes. Assuming that all customers are averse to delay (potentially to different degrees), one would expect that the more expensive product option should have a shorter leadtime, and if we were to raise its leadtime to be equal or greater to that of the cheaper product, then its demand would be zero. This behavior that seems crucial in our model where price changes serve to divert demand from one leadtime to another, cannot be captured by these aggregate demand relations. The same issue arises if the product choice probability (44) was computed using the Multinomial Logit model [35, §7.2].

6 Numerical results and concluding remarks

This section reports on a set of numerical experiments that illustrate the effectiveness of dynamic over static pricing, as well as the impact of leadtime control flexibility through the offering of multiple (two in our experiments) leadtime options. We conclude with some closing remarks.

In the sequel, we adopt the following notation. $\mathbb{E}[\pi]$: expected profit, $\mathbb{E}[\rho]$: expected utilization rate, $\mathbb{E}[EC]$: expected cost of expediting, $\mathbb{P}(LT)$: probability of violating the leadtime constraint, $\mathbb{P}(exp)$: probability of expediting, $\overline{TR}$: average tardiness, and $\mathbb{E}[TPT]$: expected sojourn time.

6.1 Single lead time

In order to isolate the effect of dynamic over static pricing, this subsection focuses on problems of pricing and expediting for a single product offered with a leadtime guarantee. We consider the following setup for these experiments. The demand model is that of §5, and unless otherwise stated, its parameters will be as follows: the market potential is $\Lambda = 10$ and $b_1 = 1$, $b_2 = 0.15$, $\gamma_0 = 2$, and $\gamma_1 = 0.4$ (cf. §5). Service times are i.i.d. exponentially distributed with rate μ. The expediting cost is $c = \$5$ per order, and expediting is used whenever the queue length reaches the threshold $\mu \cdot d - \delta$, where $\delta \geq 0$ is a "tune"-parameter the effect of which will be studied in Table 1. In Tables 1 and 2, the offered leadtime is $d = 4$, which optimizes the profit rate under static pricing. For these parameters, the capacity unconstrained revenue maximizing demand rate is $\hat{\lambda}(d) = 5$.

Table 1: The effect of the expediting "tune"-parameter δ. (Typical standard deviations for the various estimated performance measures were of the order of .1% of the estimated parameter value.)

The effect of the expediting "tune"-parameter δ: The first set of results focuses on the

	Dynamic					Static				
δ	$\mathbb{E}[\pi]$	$\mathbb{E}[\rho]$	$\mathbb{P}(exp)$	$\mathbb{P}(LT)$	$\overline{TR}$	$\mathbb{E}[\pi]$	$\mathbb{E}[\rho]$	$\mathbb{P}(exp)$	$\mathbb{P}(LT)$	$\overline{TR}$
0	20.66	.98	.021	.097	.57	19.24	.93	.060	.0041	.29
1	20.40	.97	.030	.070	.48	18.99	.92	.068	.0018	.27
2	20.10	.97	.039	.046	.43	18.78	.92	.074	.0005	.22
3	19.83	.97	.046	.027	.38	18.53	.91	.080	.0002	.21
4	19.43	.97	.058	.014	.34	18.20	.91	.090	.0000	.10

efficiency of the proposed expediting policy. In these experiments, the service rate is $\mu = 4$, which gives that $\rho^*(d) = \hat{\lambda}(d)/\mu = 1.25$, or equivalently that $\kappa = 0.45$ (c.f., equations (7) and (12)). We note that the gap between static and dynamic pricing was around 6.5%, providing an illustration of the conservative nature of static pricing policies, which, in turn, lead to higher probabilities of expediting (since price increases cannot be used to turn away orders) but lower probabilities of leadtime violation. The average tardiness is also shorter under static pricing (recall that the target leadtime is $d = 4$). The effect of the tune parameter δ on the probability of violating the leadtime guarantee was as expected, and hereafter this parameter will be selected so that the probability of an order violating its leadtime guarantee is no greater than 3%. (In almost all tests the static pricing policy had a smaller parameter δ than the dynamic pricing one.)

The effect of capacity imbalance or load factor (ρ^*): The accuracy of the approximations used in §4 is higher in systems where the capacity unconstrained revenue maximizing demand rate $\hat{\lambda}(d)$ is close to the available capacity μ. This is best described by the load factor $\rho^*(d) = \hat{\lambda}(d)/\mu$, although in the context of our analysis this distance is captured via the parameter κ which measures the distance $\hat{\lambda}(d) - \mu$ relative to $\sqrt{\hat{\lambda}(d)}$, the natural scale on which to study and control the second order behavior of the system. Table 2 explores the dependence of our results with respect to this parameter, and illustrates that the value of dynamic pricing is more pronounced for nominal loads $\rho^*(d)$ that are between .8 and 1.25. If the system is either significantly over- or under-capacitated then the gap between dynamic and static pricing policies decreases. This agrees with what one would expect from studying the functional forms for $\psi^*(w)$ in Theorem 1 as $|\kappa|$ grows large. It is also consistent with the optimal policy that would emerge from a diffusion model formulation with a linear profit loss function in (20), which would apply if $\hat{\lambda}(d) - \mu$ was large and a first order rather than second order Taylor expansion was more applicable.

Table 2: The effect of capacity imbalance (κ) or load factor (ρ^*).

The effect of the leadtime (d): Table 3 studies the system behavior under the dynamic and static pricing heuristics derived in §4 as a function of the quoted leadtime. The main observation is that the impact of dynamic pricing is more pronounced when leadtimes are shorter, since in such

μ	$\rho^*(d)$	κ	Dynamic			Static		
			$\mathbb{E}[\pi]$	$\mathbb{E}[\rho]$	$\mathbb{P}(exp)$	Gap (%)	$\mathbb{E}[\rho]$	$\mathbb{P}(exp)$
3.5	1.43	.67	17.76	.99	.092	.26	.94	.090
4	1.25	.45	19.83	.97	.046	2.97	.93	.060
4.5	1.11	.22	21.30	.95	.014	4.52	.91	.039
5	1.00	.00	21.65	.92	.009	2.72	.88	.025
5.5	0.91	-.22	21.85	.88	.005	1.35	.85	.013
6	0.83	-.45	21.93	.82	.002	.53	.81	.006
6.5	0.77	-.67	21.96	.77	.001	.19	.76	.002
7	0.71	-.89	21.99	.71	.0005	.12	.71	.001

cases static prices have to be selected conservatively to avoid excessive use of expediting.

We also experimented with varying other parameters such as the market potential Λ and the expediting cost c. The former has a similar effect to decreasing the capacity μ (c.f., Table 2), while the latter had the expected effect that as c increases, prices increase so as to lower the probability of expediting, and in such cases the benefits from dynamic pricing are more important.

Table 3: The effect of different leadtime guarantees.

d	Dynamic				Static			
	$\mathbb{E}[\pi]$	$\mathbb{P}(exp)$	$\mathbb{P}(LT)$	$\mathbb{E}[TPT]$	$\mathbb{E}[\pi]$	$\mathbb{P}(exp)$	$\mathbb{P}(LT)$	$\mathbb{E}[TPT]$
2	18.63	.115	.030	.93	17.20	.152	.002	.62
3	19.65	.068	.028	1.46	18.83	.087	.004	1.10
4	19.83	.046	.027	2.00	19.24	.060	.004	1.56
5	19.62	.037	.026	2.59	19.21	.046	.004	2.02
6	19.30	.029	.024	3.15	18.97	.037	.005	2.51

6.2 Effect of leadtime flexibility: single good offered at two leadtimes

We adopt the same model parameters as for the experiments reported above, setting the service rate at $\mu = 4$, together with the specification that the delay sensitivity parameter χ used in selecting a product in (44) is uniformly distributed in an interval $[0, \chi_m]$, and $\chi_m = 2$ unless otherwise specified. Since both products correspond to the same good, we will assume that $c_i = 5$ for $i = 1, 2$.

Table 4: Performance measures for different leadtime combinations.

Our first set of results looks at the impact of leadtime flexibility on a baseline example that was already analyzed in the previous subsection, for which $\mu = 4$ and a single leadtime option

d_1, d_2	Dynamic		Static	
	$\Delta(\pi)$ (%)	$\mathbb{P}(exp)$	$\Delta(\pi)$ (%)	$\mathbb{P}(exp)$
3, 4	9.65	.073	7.95	.085
3.5, 4	10.34	.065	8.88	.075
3.5, 4.5	12.00	.056	8.77	.078
3.5, 5	10.61	.066	9.02	.078
4, 4.5	12.71	.047	9.84	.067
4, 5	11.74	.056	10.21	.067

was offered at $d = 4$ (that corresponds to the second row in Table 2). The results in this table study the performance of dynamic and static pricing for various pairs of leadtimes (d_1, d_2). The profit gain reported in the table is in comparison to the system with the single leadtime $d = 4$ under static pricing. Recall that from Table 2 we know that even with a single leadtime the use of dynamic pricing leads to a 3% profit gain. The results of this table offer a representative picture of the type of performance gains observed in a variety of other examples we tested, as well as the differences between dynamic and static pricing. [The latter, of course, depends on the parameter κ_w as illustrated in Table 2, and are more significant whenever the system's capacity is close to the revenue maximizing demand rate $\hat{\lambda}(d)$.

We complement this table by reporting average results from a larger set of test problems where we varied some of the demand model parameters as follows: $\Lambda \in \{8, 10, 12\}$, $b_1 \in \{.8, 1, 1.2\}$, and $b_2 \in \{.1, .15, .2\}$. For all possible parameter combinations, we first considered the problem of offering one product option and searched for the optimal leadtime d^* under the static pricing policy. We then tested the performance of the dynamic and static pricing policies for the case where the firm offered two leadtime options defined as $d_1 = (.8)d^*$ and $d_2 = (1.2)d^*$. Once again we report % profit gains over the single-leadtime system under static pricing. We observed the following results:

i. Dynamic pricing: average profit gain was 13.97% with a standard deviation of 2.34%.

ii. Static pricing: average profit gain was 10.69% with a standard deviation of 2.99%.

iii. Dynamic versus static pricing: average gap 3.28% with a standard deviation of 2.03%. Actual differences ranged in $[1.06\%, 10.30\%]$.

6.3   Concluding remarks

This paper developed a framework for studying problems of joint dynamic pricing and leadtime quotation in make-to-order queues. A key feature of our model was that it assumed that the firm commits to a set of predetermined leadtime options at which to offer each good, and then focuses

on dynamic pricing decisions to both manage the overall demand into the system as well as its split across the different leadtime options. This reduced the joint control problem to one of dynamic pricing subject to leadtime guarantees, which is more tractable. In parallel, we proposed a model of purchase behavior for price and delay sensitive demand that seems suitable for problems with perfectly substitutable products that are differentiated in terms of their prices and leadtimes. Our analysis based on an approximating diffusion control problem led to several insights regarding the structure of near-optimal sequencing, expediting and pricing policies, as well as to the value of dynamic pricing and leadtime control in make-to-order environments.

Perhaps the most important issue for future work that is motivated by this paper would be to study the extent to which the proposed demand model offers an accurate representation for actual purchase behavior from price and delay sensitive customers selecting among products that are differentiated in these dimensions. This practical concern lies at the core of revenue optimization in make-to-order or stochastic service systems, and has important implications on the nature of the decisions that one would make, the type of product differentiation that a firm may want to achieve, and the market equilibrium that would emerge in a competitive environment. A related aspect of demand modelling that needs to be studied is the long-run strategic interaction between the firm and its customers when the firm adopts tactical demand management techniques.

Analytically, the model analyzed in this paper is fairly stylized and needs to be extended in order to offer a more accurate representation of a production system, as well as to systems that either exclusively or partly operate in a make-to-stock fashion. Another interesting problem would be to optimize over the vector of leadtime options that the firm will choose to offer; this would use the demand model proposed in §5 to optimally trade off revenues with expediting costs. Finally, from a purely analytical viewpoint, our paper left unanswered the question of whether the proposed set of policies is asymptotically optimal in the regime identified in §4.1. Addressing this point would combine elements from Plambeck et.al [31] and Mandelbaum and Pats [27], with a significant extension, however, to deal with the fact that the proposed policy is not pathwise optimal (which is a property that simplifies such proofs allowing one to use sample path arguments).

A Appendix

Background on queues in heavy-traffic. We list below a few known results from heavy-traffic theory for an $M/M/1$ queue that play a role in setting up our diffusion approximation. Specifically, consider a sequence of single-product queues with no pricing or expediting control capability that are indexed by n, where the nth system has Poisson demand with rate $\lambda^n = \mu(n - \theta\sqrt{n})$ for $\theta > 0$ and service rate $\mu^n = n\mu$. This set of parameters gives rise to the so-called heavy-traffic regime,

where $\rho^n = 1 - \theta/\sqrt{n}$. Let $Q^n(t)$ denote the queue length process associated with that system. Then, the following are true (see [27] for a derivation):

i. $Q^n(t)/\sqrt{n}$ converges weakly[5] to a limit process $Z(t)$ (i.e., $Q^n(t)$ is of order $\sqrt{n}$).

ii. Let $w^n(t)$ denote the virtual waiting time for an order arriving at time t. Then, $\sqrt{n}w^n(t)$ converges weakly to $Z(t)/\mu$ (i.e., waiting times $w^n(t)$ are of order $1/\sqrt{n}$).

From ii. we see that leadtime constraints of the form $w^n(t) \leq d^n$, asymptotically reduce to an upper bound on the queue length of the form $Z(t) \leq \mu d$, where $d = \lim_n \sqrt{n}d^n$. With that in mind, consider the same $M/M/1$ queue with order expediting whenever the queue reaches the threshold $K^n = \mu^n d^n$, noting that the magnitude of $K^n = \sqrt{n}(\mu d)$ is consistent with point i.. This system behaves like an $M/M/1/K^n$ queue, for which one can show (see Plambeck et.al. [31] for an analysis of a multi-product version of such a system) that i. - ii. above continue to hold and moreover:

iii. $B^n(t)/\sqrt{n}$ converges weakly to a well defined continuous, non-decreasing limit process (i.e., the number of orders expedited in $[0, t]$ is of order $\sqrt{n}$).

Sketch of proof of Theorem 1. In the sequel we will only provide a skeleton of the steps required in the proof of this result with appropriate references to [4] for detailed arguments.

Step 1 – Characterization of the solution using results from [4]: The arguments of Propositions 3 and 4 in [4] suggest that the optimal policy can be characterized as a solution to the following Bellman equation by finding a constant γ and a continuous, twice differentiable function V that satisfy:

$$\gamma = \min_\psi \left\{ \frac{\sigma_w^2}{2} V''(w) - \psi V'(w) + \alpha_w (\psi + \kappa_w)^2 \right\} \quad \text{for } w \in (0, \tilde{w}) \tag{45}$$

with boundary conditions

$$V'(0) = 0 \quad \text{and} \quad V'(\tilde{w}) = \tilde{c}_I \tilde{\mu}_I. \tag{46}$$

The first order optimality conditions for (45) give that $\psi^*(w) = V'(w)/(2\alpha_w) - \kappa_w$ and that

$$\gamma = \frac{\sigma_w^2}{2} V''(w) - \frac{1}{4\alpha_w}(V'(w) - 2\alpha_w \kappa_w)^2 + \alpha_w \kappa_w^2. \tag{47}$$

This is, of course, a first-order differential equation, which can be summarized in the form:

$$\frac{\sigma_w^2}{2} \left(\frac{v'(w)}{\gamma - \alpha_w \kappa_w^2 + \frac{1}{4\alpha_w}(v(w) - 2\alpha_w \kappa_w)^2} \right) = 1 \quad \text{for } w \in (0, \tilde{w}), \tag{48}$$

---

[5]If $(X^n(t) : t \geq 0)$ and $(X(t) : t \geq 0)$ are continuous time stochastic processes with sample paths in $D^m[0, \infty)$, the space of right-continuous functions with left limits (RCLL), then $X^n(\cdot) \Rightarrow X(\cdot)$ denotes weak convergence in $D^m[0, \infty)$ with respect to the Skorohod topology, see, e.g., [6, §3].

for $v(w) = V'(w)$. The boundary conditions are $v(0) = 0$ and $v(\tilde{w}) = \tilde{c}_I \tilde{\mu}_I$.

The solution to (48) can be obtained using the steps outlined in [4, §3.3-3.4]. Specifically, in the notation of [4] we have that $\phi(v(w)) = 1/(4\alpha_w)(v(w) - 2\alpha_w\kappa_w)^2 - \alpha_w\kappa_w^2$, $b = \tilde{w}$ and $p = \tilde{c}_I\tilde{\mu}_I$. Corollary 1 from [4] ensures that this equation admits a unique solution (γ, v), with $\gamma > 0$, that jointly satisfy (48) together with the two boundary conditions. In addition, the long-run profit loss γ is strictly increasing in the expediting cost parameter $\tilde{c}_I\tilde{\mu}_I$ (see Ata [2, Proposition 12]).

For the particular cost structure of our problem, one can derive a closed form characterization of the optimal (γ, v), and therefore of the optimal drift function itself, $\psi^*(w)$. This is done in the remainder of this proof that is divided in three cases covering the possible parameter combinations where $\tilde{c}_I\tilde{\mu}_I - h(\alpha_w, \kappa_w, \tilde{w}, \sigma_w)$ is zero, positive, or negative, respectively.

Step 2 – Closed form characterization of the optimal drift function:

Case i. $(h(\tilde{c}_I\tilde{\mu}_I, \alpha_w, \kappa_w, \tilde{w}, \sigma_w) = 0)$: Let's assume optimistically that $\gamma = \alpha_w\kappa_w^2$, in which case (48) reduces to

$$2\alpha_w\sigma_w^2 \frac{v'(w)}{(v(w) - 2\alpha_w\kappa_w)^2} = 1 \quad \text{for } w \in (0, \tilde{w}),$$

with the boundary conditions $v(0) = 0$ and $v(\tilde{w}) = \tilde{c}_I\tilde{\mu}_I$. Making the change of variable $u := (v(w) - 2\alpha_w\kappa_w)$, and integrating from $w = 0$ to $w = y$ we transform the above equation to

$$2\alpha_w\sigma_w^2 \int_{u_1}^{u_2} \frac{du}{u^2} = y \quad \text{for } y \in (0, \tilde{w}),$$

where $u_1 = -2\alpha_w\kappa_w$ and $u_2 = (v(y) - 2\alpha_w\kappa_w)$, from which we get that

$$2\alpha_w\sigma_w^2 \left[\frac{1}{2\alpha_w\kappa_w - v(y)} - \frac{1}{2\alpha_w\kappa_w} \right] = y \quad \text{for } y \in (0, \tilde{w}).$$

and that

$$v(w) = 2\alpha_w\kappa_w - \left(\frac{w}{2\alpha_w\sigma_w^2} + \frac{1}{2\alpha_w\kappa_w} \right)^{-1}.$$

Note that $v(0) = 0$ and $v(\tilde{w}) = \tilde{c}_I\tilde{\mu}_I$ if and only if $h(\tilde{c}_I\tilde{\mu}_I, \alpha_w, \kappa_w, \tilde{w}, \sigma_w) = 0$. From $\psi^*(w) = v(w)/(2\alpha_w) - \kappa_w$ we derive (36). Brute force verification shows that $\gamma = \alpha_w\kappa_w^2$ and $v(w)$ defined above satisfy (48), and an application of Theorem 1 [4] establishes the optimality of (36).

Case ii. $(h(\tilde{c}_I\tilde{\mu}_I, \alpha_w, \kappa_w, \tilde{w}, \sigma_w) > 0)$: Given the monotonicity of γ as a function of $\tilde{c}_I\tilde{\mu}_I$ [2, Proposition 12] and the result of case i., the parameter combinations considered in this case will have $\gamma > \alpha_w\kappa_w^2$. With that in mind, we let $\zeta_1 = \gamma - \alpha_w\kappa^2$, and proceed under the assumption that $\zeta_1 > 0$, making the change of variable $u := (v(w) - 2\alpha_w\kappa_w)/(2\sqrt{\alpha_w\zeta_1})$, and integrating from $w = 0$

to $w = y$ we transform the above differential equation to

$$\sigma_w^2 \sqrt{\frac{\alpha_w}{\zeta_1}} \int_{u_1}^{u_2} \frac{du}{1 + u^2} = y \quad \text{for } y \in (0, \tilde{w}),$$

where $u_1 = -\kappa_w \sqrt{\frac{\alpha_w}{\zeta_1}}$ and $u_2 = (v(y) - 2\alpha_w \kappa_w)/(2\sqrt{\alpha_w \zeta_1})$. Recalling from calculus that $\int_t^w \frac{du}{1+u^2} = \arctan w - \arctan t$, leads to the following solution for the above equation:

$$\sigma_w^2 \sqrt{\frac{\alpha_w}{\zeta_1}} \left[\arctan\left(\frac{v(y) - 2\alpha_w \kappa_w}{2\sqrt{\alpha_w \zeta_1}} \right) + \arctan\left(\kappa_w \sqrt{\frac{\alpha_w}{\zeta_1}} \right) \right] = y \quad \text{for } y \in (0, \tilde{w}) \tag{49}$$

which should satisfy the boundary conditions at $y = 0$ and $y = \tilde{w}$. The condition at $y = 0$ is satisfied since $v(0) = 0$. Using $v(\tilde{w}) = \tilde{c}_I \tilde{\mu}_I$ at $y = \tilde{w}$, the second boundary condition reduces to

$$\sigma_w^2 \sqrt{\frac{\alpha_w}{\zeta_1}} \left[\arctan\left(\frac{\tilde{c}_I \tilde{\mu}_I - 2\alpha_w \kappa_w}{2\sqrt{\alpha_w \zeta_1}} \right) + \arctan\left(\kappa_w \sqrt{\frac{\alpha_w}{\zeta_1}} \right) \right] = \tilde{w}, \tag{50}$$

which can be used to specify the positive constant ζ_1, which is guaranteed to exist and is unique [4, Corollary 1]. Using (49) we get that

$$v(w) = 2\sqrt{\alpha_w \zeta_1} \cdot \tan\left[\frac{w}{\sigma_w^2} \sqrt{\frac{\zeta_1}{\alpha_w}} - \arctan\left(\kappa_w \sqrt{\frac{\alpha_w}{\zeta_1}} \right) \right] + 2\alpha_w \kappa_w, \tag{51}$$

from which we derive the control ψ given in (37). Brute force analysis shows that (γ, v) with $\gamma = \zeta_1 + \alpha_w \kappa_w^2$ solve (48), and an application of Theorem 1 in [4] establishes the optimality of (37).

Case iii. $(h(\tilde{c}_I \tilde{\mu}_I, \alpha_w, \kappa_w, \tilde{w}, \sigma_w) < 0)$: As before, the monotonicity of γ as a function of $\tilde{c}_I \tilde{\mu}_I$ [2, Proposition 12] and the result of case i. imply that the parameter combinations considered here will have $0 < \gamma < \alpha_w \kappa_w^2$. With that in mind, we let $\zeta_2 = -(\gamma - \alpha_w \kappa_w^2)$, and rewrite (47) as:

$$-\frac{\sigma^2}{2\zeta_2} \left(\frac{v'(w)}{1 - \frac{1}{4\alpha_w \zeta_2}(v(w) - 2\alpha_w \kappa_w)^2} \right) = 1 \quad \text{for } w \in (0, \tilde{w}).$$

Proceeding under the assumption that $\zeta_2 > 0$, making the change of variable $u := (v(w) - 2\alpha_w \kappa_w)/(2\sqrt{\alpha_w \zeta_2})$, and integrating from $w = 0$ to $w = y$ we get that

$$-\frac{\sigma^2}{2} \sqrt{\frac{\alpha_w}{\zeta_2}} \left[\ln\left| \frac{v(y) - 2\alpha_w \kappa_w + 2\sqrt{\alpha_w \zeta_2}}{-v(y) + 2\alpha_w \kappa_w + 2\sqrt{\alpha_w \zeta_2}} \right| - \ln\left| \frac{\sqrt{\zeta_2/\alpha_w} - \kappa_w}{\sqrt{\zeta_2/\alpha_w} + \kappa_w} \right| \right] = y \quad \text{for } y \in (0, \tilde{w}), \tag{52}$$

together with the boundary conditions $v(0) = 0$ and $v(\tilde{w}) = \tilde{c}_I \tilde{\mu}_I$. Using (52) we get that

$$v(w) = (2\alpha_w \kappa_w + 2\sqrt{\alpha_w \zeta_2}) - 4\sqrt{\alpha_w \zeta_2} \left[1 - \exp\left\{ -\frac{2w}{\sigma^2} \sqrt{\frac{\zeta_2}{\alpha_w}} + C \right\} \right]^{-1}$$

where $C = \ln\left(\frac{\sqrt{\zeta_2/\alpha_w}-\kappa_w}{\sqrt{\zeta_2/\alpha_w}+\kappa_w}\right)$, which in turn implies the functional form for ψ given in (38). It is easy to verify that $v(0) = 0$, while $v(\bar{w}) = \tilde{c}_I \tilde{\mu}_I$ reduces to

$$(2\alpha_w\kappa_w + 2\sqrt{\alpha_w\zeta_2}) - 4\sqrt{\alpha_w\zeta_2}\left[1 - \exp\left\{-\frac{2\tilde{w}}{\sigma^2}\sqrt{\frac{\zeta_2}{\alpha_w}} + C\right\}\right]^{-1} = \tilde{c}_I\tilde{\mu}_I, \qquad (53)$$

which can be used to define ζ_2 as its unique positive solution, which from [4, Corollary 1] and [2, Proposition 12] is guaranteed to lie in $(0, \alpha_w\kappa_w^2)$. Proceeding as for the two previous cases for $\gamma = \alpha_w\kappa_w^2 - \zeta_2$ establishes the optimality of (38). $\blacksquare$

References

[1] P. Afeche. Incentive-compatible revenue management in queueing systems: Optimal strategic idleness and other delaying tactics. 2004. Working paper, Kellogg School of Management, Northwestern University.

[2] B. Ata. *Dynamic control for stochastic networks*. PhD thesis, Graduate School of Business, Stanford University, 2003.

[3] B. Ata. Dynamic control of a mutliclass queue with thin arrival streams. 2004. Preprint, Kellogg School of Management, Northwestern University.

[4] B. Ata, J. M. Harrison, and L. A. Shepp. Drift rate control of a Brownian processing system. *Ann. Appl. Prob.*, 15(2):1145–1160, 2005.

[5] K. Baker. Sequencing rules and due-date assignmenets in a job shop. *Management Science*, 30(9):1093–1104, 1984.

[6] P. Billingsley. *Convergence of Probability Measures*. 2nd ed., Wiley, New York, 1999.

[7] J. L. Bradley. A Brownian approximation of a production-inventory system with a manufacturer that subcontracts. *Oper. Res.*, 2004. Forthcoming.

[8] J. L. Bradley. Optimal control of a dual service rate M/M/1 production-inventory model. *European Journal Operational Research*, 2004. Forthcoming.

[9] P. Brémaud. *Point Processes and Queues: Martingale Dynamics*. Springer-Verlag, 1980.

[10] R. E. Bucklin and S. Gupta. Brand choice, purchase incidence, and segmentation: an integrated modeling approach. *Journal of Marketing Research*, XXIX:201–215, 1992.

[11] K. Charnisirisakskul, P. Griffin, and P. Keskinocak. Pricing and scheduling decisions with leadtime flexibility. 2004. Preprint, Georgia Institute of Technology, GA.

[12] I. Duenyas. Single facility due date setting with multiple customer classes. *Management Science*, 41:608–619, 1995.

[13] I. Duenyas and W. J. Hopp. Quoting customer lead times. *Management Science*, 41(1):43–57, 1995.

[14] G. Edmondson. Customization – BMW. *BusinessWeek*. November 24, 2003.

[15] G. Gallego and G. van Ryzin. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Science*, 40(8):999–1020, 1994.

[16] P. W. Glynn. Diffusion approximations. In D. Heyman and M. Sobel, editors, *Stochastic Models*, volume 2 of *Handbooks in OR & MS*, pages 145–198. North-Holland, 1990.

[17] J. M. Harrison. *Brownian motion and stochastic flow systems*. John Wiley & Sons, 1985.

[18] J. M. Harrison. Brownian models of queueing networks with heterogeneous customer populations. In W. Fleming and P. L. Lions, editors, *Stochastic Differential Systems, Stochastic Control Theory and Applications*, volume 10 of *Proceedings of the IMA*, pages 147–186. Springer-Verlag, New York, 1988.

[19] J. M. Harrison. A broader view of brownian networks. *Ann. Appl. Prob.*, 13, 2003.

[20] J. M. Harrison and J. A. Van Mieghem. Dynamic control of brownian networks: State space collapse and equivalent workload formulations. *Ann. Appl. Prob.*, 7:747–771, 1996.

[21] P. Keskinocak, R. Ravi, and S. Tayur. Scheduling and reliable lead-time quotation for orders with availability intervals and lead-time sensitive returns. *Management Science*, 47(2):264–279, 2001.

[22] P. Keskinocak and S. Tayur. Due dtate management policies. In D. Simchi-Levi, S. D. Wu, and Z. M. Shen, editors, *Supply chain analysis in the e-business era*, pages 485–553. Kluwer Acedemic Publishers, Norwell, MA, 2003.

[23] M. Lariviere and J. A. V. Mieghem. Strategically seeking service: How competition can generate Poisson arrivals. *Manufacturing & Service Operations Management*, 6(1):23–40, 2004.

[24] C. Maglaras. Revenue management for a multi-class make-to-order queue. Technical report, Working paper, Columbia University, 2003.

[25] C. Maglaras and J. Van Mieghem. Queueing systems with leadtime constraints: a fluid model approach for admission and sequencing control. *European Journal Operational Research*, 167:179–207, 2005.

[26] C. Maglaras and A. Zeevi. Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Science*, 49(8):1018–1038, 2003.

[27] A. Mandelbaum and G. Pats. State-dependent queues: approximations and applications. In F. Kelly and R. Williams, editors, *Stochastic Networks*, volume 71, pages 239–282. Proceedings of the IMA, 1995.

[28] H. Mendelson. Pricing computer services: queueing effects. *Communications of the ACM*, 28(3):312–321, 1985.

[29] H. Mendelson and S. Whang. Optimal incentive-compatible priority pricing for the M/M/1 queue. *Oper. Res.*, 38(5):870–883, 1990.

[30] P. Naor. The regulation of queue size by levying tolls. *Econometrica*, 37:15–24, 1969.

[31] E. Plambeck, S. Kumar, and J. M. Harrison. Leadtime constraints in stochastic processing networks under heavy traffic conditions. *Queueing Systems*, 39:23–54, 2001.

[32] E. L. Plambeck. Optimal leadtime differentiation via diffusion approximations. *Oper. Res.*, 52(2):213–228, 2004.

[33] E. L. Plambeck and A. R. Ward. A separation principle for assemble-to-order systems with expediting. 2003. Working paper, Stanford Business School.

[34] M. I. Reiman. Open queueing networks in heavy traffic. *Math. Oper. Res.*, 9:441–458, 1984.

[35] K. Talluri and G. van Ryzin. Revenue management under a general discrete choice model of consumer behavior. *Management Science*, 50(1):15–33, 2004.

[36] J. Van Mieghem. Price and service discrimination in queueing systems: incentive compatibility of $Gc\mu$ policy. *Management Science*, 46(9):1249–1267, 2000.

[37] L. M. Wein. Due-date setting and priority sequencing in a multiclass $M/G/1$ queue. *Management Science*, 37:834–850, 1991.