

The Impact of Duplicate Orders on Demand Estimation and Capacity Investment

Mor Armony

Stern School of Business, New York University, New York, New York 10012, marmony@stern.nyu.edu

Erica L. Plambeck

Graduate School of Business, Stanford University, Stanford, California 94305, elp@stanford.edu

Motivated by a \$2.2 billion inventory write-off by Cisco Systems, we investigate how duplicate orders can lead a manufacturer to err in estimating the demand rate and customers' sensitivity to delay, and to make faulty decisions about capacity investment. We consider a manufacturer that sells through two distributors. If a customer finds that his distributor is out of stock, then he will sometimes seek to make a purchase from the other distributor; if the latter is also out of stock, the customer will order from both distributors. When his order is filled by one of the distributors, the customer cancels any duplicate orders. Furthermore, the customer cancels all of his outstanding orders after a random period of time.

Assuming that the manufacturer is unaware of duplicate orders, we prove that she will overestimate both the demand rate and the cancellation rate. Surprisingly, failure to account for duplicate orders can cause short-term underinvestment in capacity. However, in long-term equilibrium under stable demand conditions the manufacturer overinvests in capacity. Our results suggest that Cisco's write-off was caused by estimation errors and cannot be blamed entirely on the economic downturn. Finally, we provide some guidance on estimation in the presence of double orders.

Key words: maximum-likelihood estimation; duplicate ordering; distribution channels; queueing systems; renegeing

History: Accepted by William S. Lovejoy, operations and supply chain management; received April 19, 2002. This paper was with the authors 3½ months for 2 revisions.

1. Introduction

Amid a general economic downturn, networking titan Cisco Systems experienced a spectacular fall in market value from \$430 billion in March of 2000 to \$180 billion in March of 2001. Net income dropped from \$0.8 billion in the first quarter of 2001 to -\$2.7 billion in the third quarter of that year as Cisco wrote off \$2.2 billion worth of component inventory and laid off 8,500 workers (*Business Week* 2002a). According to the *Wall Street Journal*, "Cisco executives ignored or misread crucial warning signs that their sales forecasts were too ambitious." Because of duplicate orders, Cisco executives overestimated demand and therefore "continued to expand capacity aggressively, even after business slowed" (Thurm 2001a, p. A1). Cisco is certainly not the only technology company to have difficulties in forecasting because of duplicate orders. Intel and other semiconductor manufacturers believe that their bookings data is "irrelevant and potentially misleading" because of duplicate orders (*Business Week* 2002b, p. 28). This paper shows that even in a stable business environment, a manufacturer that fails to account for double orders will carry excess capacity.

Cisco's policy of outsourcing all of its manufacturing has been lauded in the business press. Less widely recognized is that, since 1998 (when Cisco achieved 65% of its revenues through direct sales), Cisco has sought to outsource sales and distribution. Cisco sells networking hardware to distributors (e.g., Ingram Micro) that sell to systems integrators (including IBM and a host of smaller firms) that in turn sell to Cisco's end customers and provide ongoing support and maintenance. By 2001, the number of Cisco-qualified distributors and resellers (systems integrators) had increased to 20,000 for the United States alone (Kothari 2001). Only 14% of Cisco's sales were direct; 86% were through channels. Cisco was using the Internet to share real-time information about inventory and production schedules with its component suppliers and contact manufacturers (*Business Week* 2001). However, on the demand side, Cisco's information systems were relatively weak. In particular, Cisco had limited visibility of distributors' inventory and order backlog (Kothari 2001).

In the summer of 2000, Cisco experienced shortages of several key components. Customers had to wait for two and even three months for some of Cisco's most popular products. Some frustrated cus-

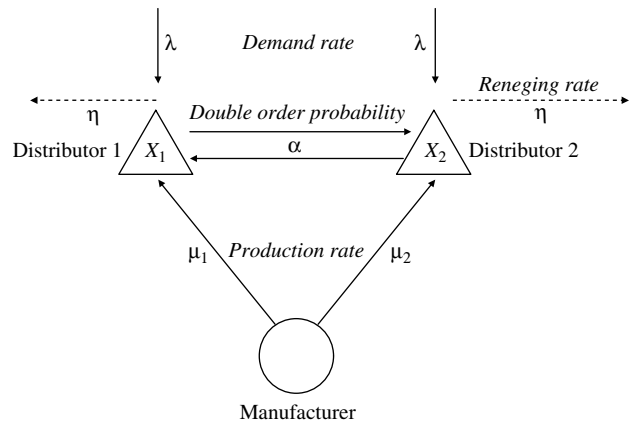
tomers chose to cancel their orders and buy equipment from Cisco's competitors (Juniper Networks, Nortel Networks). Customers and resellers also began to order from multiple distributors with the intention of cancelling duplicate orders as soon as one distributor shipped the product. Cisco failed to recognize the extent of the double orders and therefore, although the tech economy had already begun to slow down, Cisco maintained its ambitious sales forecasts. To avoid long lead times and lost sales, Cisco added workers and stockpiled components. Cisco also loaned \$600M without interest to contract manufacturers to buy even more parts. This expanded capacity did indeed serve to reduce production lead times throughout the fall of 2000. The order backlog disappeared as customers cancelled duplicate orders, and new orders anticipated by Cisco failed to materialize (Thurm 2001a). Cisco was saddled with excess capacity. Despite the write-off of \$2.2 billion in component inventory in April of 2001, Cisco carried \$1.68 billion in parts and unsold equipment on its books at the end of fiscal year 2001 (Thurm 2001b). According to Cisco Chief Strategy Officer Michelangelo Volpi, "We didn't know the magnitude [of duplication in the order backlog]. Without the misleading information we might have seen better and made better decisions." (Thurm 2001a).

Cisco's experiences raise several interesting questions: Will order duplication cause a manufacturer to overestimate the demand rate and the rate at which sales will be lost if customers are forced to wait (the renegeing rate)? If so, by how much? How will this affect capacity investment? How can the manufacturer ascertain the true demand rate, incidence of duplicate orders, and renegeing rate?

To address these questions, we analyze a stylized model of a manufacturer with two independent distributors (see Figure 1 for an illustration). At each distributor, customers arrive according to a Poisson process with rate λ . If a customer finds that his distributor is out of stock, then with probability α he will also place an order with the other distributor; as soon as one distributor supplies the product, the customer cancels his order with the other. Furthermore, the customer will renege,¹ cancelling all outstanding orders after a length of time that is exponentially distributed with rate η . The manufacturer has visibility of each distributor's inventory level or number of outstanding orders, but not customer identities. (The manufacturer can infer a distributor's inventory level from

¹ In most queueing models of service systems, a customer will not renege during his service, but will renege only while waiting for service to begin. In contrast, in our queueing model of manufacturing, the customer at the head of the line may renege, although he has claim to the product in process.

Figure 1 Illustration of the Model with a Single Manufacturer and Two Distributors



her own order queue if the distributor uses a base-stock policy. Since 2001, firms in the high-tech, automotive, chemical, and home appliance industries have implemented software to monitor distributors' inventory levels.) A detailed model formulation is provided in §2.

In §3, we derive maximum-likelihood estimators (MLEs) for λ and η in a system without double orders ($\alpha = 0$). Next, we assume that some customers will double order (unknown to the manufacturer) and we prove that by using the MLEs for the system with $\alpha = 0$, the manufacturer will *overestimate* the demand rate λ and the renegeing rate η . The basic problem is that double orders are counted as additional customer arrivals, and cancellations of double orders are counted as lost sales. Section 3 also contains sensitivity analysis of the *systematic error*, the difference between the parameter values estimated by the manufacturer (assuming that $\alpha = 0$) and the true parameter values.

The MLEs are valid for an arbitrary schedule of shipments from the manufacturer to the distributors. However, to investigate the impact of duplicate orders on capacity investment, production is modeled as a Poisson process with rate μ . We investigate two plausible allocation rules for finished goods. The base case is that each distributor i has a fixed portion of the capacity μ_i (with $\mu_1 + \mu_2 = \mu$). This is relevant when the distributors are located in different geographical regions and transportation costs are high, so the manufacturer serves them from different production facilities. (Cisco has regional production facilities, and some resellers, particularly those at an intermediate location, will duplicate-order from distributors in different regions.) The base case with $\mu_1 = \mu_2$ also approximates a "fair" division of the output between the distributors. The second case is that capacities are pooled, and distributors' orders are filled first in–first out (FIFO). (Cisco is concerned with fairness.)

Based on her estimation of demand, the manufacturer chooses μ to minimize the cost of capacity and lost sales. She assumes that each distributor uses a base-stock policy for inventory control with a fixed base stock level B . In reality, capacity is a strategic decision and inventory policy is a short-term, tactical decision that responds to capacity utilization (lead times). In Cisco's experience, distributors lower their inventory levels in response to an increase in the production capacity. Moreover, distributors learn about demand and adjust their inventory policies dynamically, and the optimal inventory policy for one distributor depends on the inventory policy of the other, as they compete for capacity and customers. Complete analysis of the strategic interaction between the three parties is beyond the scope of this paper. We simply characterize the manufacturer's best response μ to a given base stock level B used by both distributors. Opportunities for further research lie in integrating strategic capacity investment with the rich literature on inventory competition and Bayesian inventory management.

Section 4 demonstrates how overestimating the demand rate and the renege rate can cause the manufacturer to purchase too much capacity (like Cisco). On the contrary, when the cost of capacity is very high, a manufacturer that is unaware of duplicate orders will *underinvest* in capacity. For any fixed capacity level, duplicate ordering reduces the number of lost sales and thus increases the manufacturer's profit. Unfortunately, unrecognized duplicate ordering reduces the manufacturer's profit through errors in capacity planning. This serves as a warning to manufacturers: Watch out for double orders!

For the watchful manufacturer, §5 provides estimators for λ , η , and α , based on a general shipment schedule and continuous observation of inventory levels. Commonly, the manufacturer observes inventory levels infrequently if at all. Therefore, in §5.1 we adapt the estimators to handle discrete-time information about inventory levels.

1.1. Literature Review

The literature related to this research falls into four categories: (1) estimation of customer characteristics, (2) dynamic inventory control under demand uncertainty, (3) capacity investment under uncertainty, and (4) strategic interaction between a manufacturer and competitive retailers with customer substitution.

For queues with impatient customers, Mandelbaum and Zeltyn (1998) and Daley and Servi (2001) derive MLEs for the demand rate and the renege rate. Motivated by applications in networking and call centers, these authors assume that the queue length is not observable; they use only transaction data (the points in time that a customer begins or completes service).

Hence, their MLEs differ from the ones derived in this paper, where queue lengths (inventory levels) are observable. Anupindi et al. (1998) consider a retail store in which customers arrive according to a Poisson process; if the desired product is not in stock, a customer may substitute it with another item, or depart without making a purchase. Given discrete-time observations of the inventory in the store, they derive MLEs for the demand rate and substitution probability. This resembles estimating the demand rate and double-order probability (the probability that a customer will "substitute" an alternative distributor) in our model. Lee et al. (1997) observe that the *variance* of orders from a distributor to a manufacturer is larger than the variance of actual sales to end customers, the famous bullwhip effect. They conclude that for effective forecasting, the manufacturer needs sales data. Our results are complementary: Duplicate orders distort the *mean*. Therefore, the manufacturer would like to know the identity of end customers, not just the sales quantity, to correct for duplicate orders.

Scarf (1959) introduced the problem of Bayesian inventory management: How should a retailer dynamically control his inventory level while learning about the demand distribution as sales evolve over time? Many researchers have tackled this challenging problem. For a variety of plausible demand distributions, assuming linear ordering and holding costs and complete backordering of demand, Azoury (1985) establishes optimality of a base-stock policy—with the base stock level scaled by a sufficient statistic for observed demand. Lovejoy (1990, 1992) shows that the adaptive base-stock policy is optimal or near optimal under more general conditions, e.g., a constant (known) renege rate, Markov-modulated demand with cheap disposal of excess inventory. Lariviere and Porteus (1999) and Ding et al. (2002) assume unobservable lost sales: The optimal base stock level is increased to learn more about demand. Larson et al. (2001) incorporate a fixed cost of ordering, and derive an optimal adaptive (s, S) policy. Toktay and Wein (2001) consider a capacity-constrained production-inventory system. The demand distribution is known, but the forecast for actual demand in future periods evolves dynamically. A dynamic (forecast-adjusted) base-stock policy minimizes inventory holding and backorder costs.

These results support our assumption that the distributors follow a base-stock policy, but suggest that the base stock level will evolve dynamically, in contrast with our simplifying assumption that the base stock level is fixed. Indeed, capacity investment decisions typically occur on a quarterly or annual basis and are irreversible in the short term. Therefore, when the manufacturer chooses her capacity investment,

she should ideally account for future dynamics in distributors' base stock levels. Van Mieghem (2003) provides an extensive review of the literature on capacity investment under uncertainty, and observes that few of these papers consider capacity investment and demand estimation. A notable example, Ryan (2003) analyzes capacity expansion with an autocorrelated demand process, and discrete capacity increments with long lead times. Our contribution is to show how an erroneous belief about system structure (disregarding duplicate orders) produces an error in estimation, and how the error in estimation both influences and is influenced by the capacity investment decision. We show how this results in excess capacity in equilibrium. This is strikingly similar to the heuristic equilibrium with excess quantity in the newsvendor model by Cachon and Kok (2002), where the procurement quantity depends on the estimated salvage value, and the estimated salvage value depends on the quantity remaining at the end of the season.

In making a capacity investment, the manufacturer should ideally anticipate how each distributor will modify his base stock level in response to the production lead time and the base stock level of the other distributor. Researchers have used game theory to analyze strategic inventory management in settings with consumer substitution, assuming that the substitution probability and distribution of demand is known. Parlar (1988), Lippman and McCardle (1997), and Mahajan and van Ryzin (2001) characterize Nash equilibria in a single-period (newsvendor) game. Anupindi and Bassok (1999) and Netessine et al. (2001) consider the stocking decisions of two retailers in a multiperiod problem with stationary demand. In each period, if a customer finds that his retailer is out of stock, he may purchase from the other retailer. Assuming that the manufacturer has unlimited production capacity, they prove existence of myopic Nash equilibria, i.e., the multiperiod game reduces to a static (newsvendor) problem. Li (1992) analyzes a queueing model very similar to ours, in the special case $\alpha = 1$ and $\eta = 0$; customers arrive according to a Poisson process and attempt to buy from one of n competing firms. If the firm has a queue of customer orders, the customer places an order with every firm, buys from the one that delivers first, then cancels all other orders. Li characterizes the conditions under which all the firms will choose to make to order rather than carry inventory. The firms, acting selfishly, may choose to carry inventory even if expected profit would be greater if all firms chose to make to order. Indeed, a common conclusion in these papers is that, in competition for customers, firms will stock *more* than is optimal. Netessine and Rudi (2003) give a counterexample with asymmetric firms, and one stocking less than is globally optimal.

In contrast, Cachon (2001) shows that when retailers compete for supply from a common manufacturer but do not compete for customers, they may carry *less* inventory than would be optimal for the supply chain as a whole. Cachon and Lariviere (1999) show how retailers' order quantities depend on the allocation scheme chosen by the manufacturer, not just his capacity investment.

2. Model Formulation

Consider a manufacturer that sells a single product through two independent distributors. For brevity, we will at times use the pronoun "he" to refer to a distributor and "she" to refer to the manufacturer. At each distribution center, customers arrive according to a Poisson process with rate λ (which is independent of customer arrivals at the other distribution center), and each customer demands one unit of the product. Let $X_i(t)$ denote the inventory level for distributor i ($i = 1, 2$) at time t ; $[X_i(t)]^- = -\min\{X_i(t), 0\}$ indicates the number of outstanding orders from customers. If a customer arrives when his distributor is out of stock ($X_i(t) \leq 0$), then with probability α the customer buys immediately from the other distributor (if the other distributor has inventory) or orders the product from *both* distributors. As soon as one of the distributors delivers the product to him, the customer will cancel the duplicate order. With probability $1 - \alpha$, the customer orders from his original distributor only. The customer is impatient; after waiting for a time that is exponentially distributed with rate η , he will cancel all outstanding orders and leave the system without making a purchase. (One may interpret this waiting time before renege as the time for an alternative manufacturer to deliver. Note that the customer at the head of the line may renege although he has claim to the product in process. This contrasts with queueing models of service systems, in which a customer will not renege during his service, only while waiting for service to begin.) We will denote by $D(t)$ the number of duplicate orders that are outstanding at time t . Clearly, the number of customers waiting for the product at time t is given by $[X_1(t)]^- + [X_2(t)]^- - D(t)$, and if $X_i(t) \geq 0$ for either $i = 1$ or 2 , then $D(t) = 0$.

Each distributor follows a base-stock policy. In particular, each distributor orders one unit from the manufacturer every time a customer orders a unit from him, and cancels an order with the manufacturer every time a customer cancels an order with him. Let $Y_i(t)$ denote the number of outstanding orders from distributor i to the manufacturer. Then, $Y_i(t) = B - X_i(t)$, where B is the base stock level. The manufacturer does not hold inventory and has a total

production capacity of rate μ . When she has outstanding orders from distributor i ($Y_i(t) > 0$), the manufacturer delivers the product according to a Poisson process with rate μ_i , which is independent of the production process for the other distributor. The manufacturer knows the base-stock policy used by the distributors, and can therefore infer the inventory level and the number of customer orders outstanding for each distributor from her own order process ($Y_1(t), Y_2(t)$). Furthermore, the manufacturer knows when a downward transition in $Y_i(t)$ corresponds to an order cancellation and when it corresponds to an order fulfillment, and therefore effectively observes the orders and cancellations made by customers.

To completely describe the system dynamics, it remains to specify the sequence in which customer orders are filled. We will assume that each distributor knows which of his customers have placed a duplicate order, and gives priority to serving these customers (to avoid losing a sale to the other distributor). Hence, (X_1, X_2, D) is a continuous-time Markov chain. The assumption that distributors can identify double orders is plausible because any customer that double orders has an incentive to reveal this to the distributors to shorten his lead time. Furthermore, software for channel management enables distributors to share information in real time about customer identity and purchasing behavior. The most plausible alternative assumption is that distributors serve customers on a FIFO basis. For most of the propositions in this paper, we have an analogous result for the system with FIFO sequencing. Under FIFO sequencing (X_1, X_2, D) is not a continuous-time Markov chain; one must keep track of the precise position of double orders in the customer order queue to obtain a Markov process. We comment on how to extend each proof from the simple case with priority sequencing to the complicated case with FIFO sequencing.

For brevity, we focus on the above system in which the manufacturer dedicates a fraction of her capacity to each distributor. To demonstrate that our results are robust, we have also analyzed a system in which the manufacturer uses a FIFO sequencing policy (with simultaneous orders placed in front of each other in the queue with equal probabilities) and in which the distributors prioritize duplicate orders. Let $(D; C) = (d; c(1), c(2), \dots, c(n))$ be the state descriptor, with d the number of outstanding duplicate orders, n the total number of outstanding orders (counting double orders twice), and $c(k) \in \{1, 2\}$ the distributor that made the order which is currently in position k in the manufacturer's queue ($k = 1, \dots, n$). Then, under these two sequencing and prioritizing assumptions $(D; C)$ is a continuous-time Markov chain. All the results in this paper also hold for this FIFO system,

with one minor exception: We have proven Proposition 3 only for $\alpha = 1$. We comment briefly within the paper on adapting our proofs to this FIFO system; details are in the online appendix (available at <http://mansci.pubs.informs.org/ecompanion.html>).

Finally, to guarantee that the Markov process (X_1, X_2, D) is ergodic, we assume that the number of dedicated orders at each distributor and the number of duplicate orders are bounded by a very large number M ; that is, $X_i^- - D \leq M$ for $i = 1, 2$, and $D \leq M$. Throughout, we omit the time index t whenever we refer to the whole process, and write $t = \infty$ to refer to the process in steady state.

3. Maximum-Likelihood Estimation When $\alpha = 0$ (in the Manufacturer's Opinion)

We begin by analyzing the basic system in which each distributor has a dedicated stream of customers ($\alpha = 0$), and derive MLEs for the demand rate and renegeing rate from the manufacturer's point of view. Then, we evaluate the *systematic error* (difference between the limiting estimator and the true parameter value) when some customers double order ($\alpha > 0$), but the manufacturer is unaware of this and uses the estimator for the system with $\alpha = 0$. We prove that the manufacturer overestimates the demand rate and the renegeing rate. Finally, we investigate how the systematic error varies with the underlying system parameters. Business-press pundits (*Business Week* 2001, Thurm 2001a) attribute Cisco's forecast error to shortages in the summer of 2000. Therefore, we pay particular attention to how the systematic error varies with capacity μ . We prove that the systematic error in estimating the demand rate is decreasing in μ and converges to zero in the limit as $\mu \rightarrow \infty$ (as the production capacity becomes much larger than the demand rate). However, the systematic error in estimating the renegeing rate η is initially increasing in μ , and may be strictly positive in the limit $\mu \rightarrow \infty$. Even if capacity is much greater than demand, so that backordering rarely occurs, a manufacturer that is unaware of duplicate ordering will make a significant error in estimating the renegeing rate.

Our first proposition introduces estimators of the demand and the renegeing rates. *The MLEs are valid for an arbitrary production and shipment schedule.* They depend only on the number of customers that have ordered from distributor i , $N_i(T)$, and the number of these orders that have been cancelled, $Z_i(T)$.

PROPOSITION 1. *For the system with $\alpha = 0$, the MLEs for λ and η are given by*

$$\hat{\lambda}(T) = \frac{N_1(T) + N_2(T)}{2T}$$

and

$$\hat{\eta}(T) = \frac{Z_1(T) + Z_2(T)}{\int_0^T [Y_1(t) - B]^+ + [Y_2(t) - B]^+ dt}$$

$$= \frac{Z_1(T) + Z_2(T)}{\int_0^T X_1^-(t) + X_2^-(t) dt},$$

respectively.

PROOF. This problem can be viewed as estimating the transition rate parameters in a continuous-time Markov chain. In particular, $Y_1(\cdot)$ and $Y_2(\cdot)$ are two independent continuous-time Markov chains with generators (transition rate matrices) Q_1, Q_2 , which satisfy for $y \geq 0$: $Q_i(y, y + 1) = \lambda$, $Q_i(y, y - 1) = \mu_i 1_{\{y > 0\}} + \eta[y - B]^+$. For a given distributor i , we count the number of transitions out of state y during the time interval $[0, T]$, including $N_i(y; T)$ arrivals, $Z_i(y; T)$ order cancellations, and $E_i(y; T)$ service completions. In addition, let $\tau_i(y; T)$ denote the total amount of time during this interval that the queue length of outstanding orders from distributor i is equal to y . Thus, the likelihood function given the observation of $Y_i(\cdot)$ can be written as follows:

$$\mathcal{L}_i(\lambda, \eta) = \prod_{y \geq 0} \exp(-\{(\lambda + \mu_i 1_{\{y > 0\}} + \eta[y - B]^+) \tau_i(y; T)\})$$

$$\cdot \lambda^{N_i(y; T)} (\eta[y - B]^+)^{Z_i(y; T)} \mu_i^{E_i(y; T)}$$

$$= \exp\left(-\left\{\lambda T + \eta \sum_{y > B} (y - B) \tau_i(y; T)\right\}\right)$$

$$\cdot \lambda^{N_i(T)} \eta^{Z_i(T)} C_i,$$

where C_i stands for a constant that involves only terms that are not a function of λ or η . Let $\mathcal{L}(\lambda, \eta) = \mathcal{L}_1(\lambda, \eta) \cdot \mathcal{L}_2(\lambda, \eta)$. Then, the values of λ and η that maximize $\mathcal{L}(\lambda, \eta)$ are $\hat{\lambda}(T)$ and $\hat{\eta}(T)$ as given in the statement of the proposition. \square

3.1. Systematic Error in Maximum-Likelihood Estimation

Now suppose that some customers double order ($\alpha > 0$) unbeknown to the manufacturer, who uses the MLE for the system with $\alpha = 0$. To compute the resulting systematic error in estimation, some additional notation is needed. The superscript “0” will indicate that the distributor is out of stock, and the superscript “1” will indicate that the corresponding distributor has items in stock. The first superscript will refer to Distributor 1, and the second to Distributor 2. For example, $N_i^{01}(T)$ denotes the number of orders placed with distributor i up to time T , when Distributor 1 is out of stock, and Distributor 2 has some items in inventory immediately prior to the arrival. Similarly, $N_i^{00}(T)$ denotes the number of orders placed with distributor i while both distributors are out of stock (immediately before the customer

arrives). $N_i^{10}(T)$ and $N_i^{11}(T)$ are defined in an analogous fashion. Also, let $\tau^{00}(T)$ be the total time the system spends in states where both distributors are out of stock during the corresponding time interval, and let $P^{00} = P(X_1(\infty) \leq 0, X_2(\infty) \leq 0)$ be the steady-state probability that both distributors are out of stock. Finally, recall that $D(\infty)$ is the steady-state (random) number of duplicate orders in the system and EX_i^- is the expected backlog level at distributor i in steady state.

The next proposition characterizes the systematic error, establishing that the manufacturer will overestimate the demand rate and the renegeing rate.

PROPOSITION 2. Suppose that customers double order with positive probability ($\alpha > 0$), but the manufacturer uses the MLE for the system with $\alpha = 0$ given in Proposition 1. Then, the systematic error in estimating the demand rate is given by

$$\hat{\lambda} - \lambda = \lambda \alpha P^{00} > 0,$$

and the systematic error in estimating the renegeing rate is given by

$$\hat{\eta} - \eta = \frac{\mu P(D(\infty) > 0)}{EX_1^- + EX_2^-} > 0,$$

where $\hat{\lambda} = \lim_{T \rightarrow \infty} \hat{\lambda}(T)$ and $\hat{\eta} = \lim_{T \rightarrow \infty} \hat{\eta}(T)$.

PROOF. Given the notation introduced above, the systematic error in estimating the demand rate is

$$\hat{\lambda} - \lambda$$

$$= \lim_{T \rightarrow \infty} \hat{\lambda}(T) - \lambda$$

$$= \lim_{T \rightarrow \infty} \frac{N_1(T) + N_2(T)}{2T} - \lambda$$

$$= \lim_{T \rightarrow \infty} \frac{N_1^{00}(T) + N_2^{00}(T)}{2\tau^{00}(T)} \frac{\tau^{00}(T)}{T}$$

$$+ \frac{[N_1^{01}(T) + N_2^{01}(T)] + [N_1^{10}(T) + N_2^{10}(T)] + [N_1^{11}(T) + N_2^{11}(T)]}{2(T - \tau^{00}(T))}$$

$$\cdot \frac{T - \tau^{00}(T)}{T} - \lambda$$

$$= (\lambda + \lambda \alpha) P^{00} + \lambda(1 - P^{00}) - \lambda = \lambda \alpha P^{00},$$

where the last equality follows from the strong law of large numbers (SLLN) for renewal processes. The above equalities indicate that the systematic error is strictly positive because double orders are counted as true customer arrivals. In addition, they imply that as the production capacity μ increases to ∞ , the systematic error goes to 0.

Calculation of the systematic error in the estimator for η is more involved. Because distributors prioritize double orders, whenever $D(t) > 0$ each service completion is coupled with an order cancellation. Hence, one cancellation is seen whenever a nonduplicate order is cancelled or a service completion occurs for

a duplicate order. The resulting rate of one cancellation at time t is $\eta(X_1^-(t) + X_2^-(t) - 2D(t)) + \mu 1_{\{D(t) > 0\}}$. Two simultaneous cancellations will be observed with rate $\eta D(t)$. Let $Z_i(x_1, x_2, d; T)$ be the number of order cancellations during the time interval $[0, T]$, when the state immediately prior to the cancellation is $(X_1(t), X_2(t), D(t)) = (x_1, x_2, d)$. We deal first with the numerator of the expression for $\hat{\eta}(T)$:

$$\begin{aligned} & \lim_{T \rightarrow \infty} \frac{Z_1(T) + Z_2(T)}{T} \\ &= \lim_{T \rightarrow \infty} \frac{\sum_{x_1 \leq B} \sum_{x_2 \leq B} \sum_{d \geq 0} \frac{Z_1(x_1, x_2, d; T) + Z_2(x_1, x_2, d; T)}{\tau(x_1, x_2, d; T)}}{\frac{\tau(x_1, x_2, d; T)}{T}} \\ &= \sum_{x_1 \leq B} \sum_{x_2 \leq B} \sum_{d \geq 0} (\eta(x_1^- + x_2^- - 2d) + \mu 1_{\{d > 0\}} + 2\eta d) \\ & \quad \cdot P(X_i(\infty) = x_i, i = 1, 2, D(\infty) = d) \\ &= \eta(EX_1^- + EX_2^-) + \mu P(D(\infty) > 0), \end{aligned}$$

where the second equality follows from the SLLN for renewal processes. The denominator of the expression for $\hat{\eta}(T)$ is simpler to analyze. Specifically,

$$\lim_{T \rightarrow \infty} \frac{\int_0^T X_1^-(t) + X_2^-(t) dt}{T} = EX_1^- + EX_2^-,$$

from ergodicity. Hence,

$$\begin{aligned} \hat{\eta} - \eta &= \lim_{T \rightarrow \infty} \hat{\eta}(T) - \eta \\ &= \lim_{T \rightarrow \infty} \frac{(Z_1(T) + Z_2(T))/T}{(\int_0^T X_1^-(t) + X_2^-(t) dt)/T} - \eta \\ &= \frac{\mu P(D(\infty) > 0)}{EX_1^- + EX_2^-}. \quad \square \end{aligned}$$

The systematic error expressions in Proposition 2 are exactly the same for the system in which the manufacturer uses a FIFO policy and distributors prioritize double orders. When distributors serve customers on a FIFO basis, Proposition 2 is true except that $P(D(\infty) > 0)$ in the numerator of the error expression for η is replaced by the probability that a job at the head of the line corresponds to a double order.

3.2. Sensitivity Analysis of Systematic Errors

Overestimation of λ and η occurs because the manufacturer fails to recognize the potential for duplicate orders. One might therefore expect that as the production capacity increases, the systematic error will decrease because there is less opportunity for double ordering to occur. In this section, we prove that, as expected, the systematic error in the estimator for λ decreases with μ . However, the error in the estimator for η is not so well behaved. In fact, we observe situations in which $\hat{\eta} - \eta$ first increases with μ , and

only then starts to decrease. We explain this behavior by teasing apart the various drivers of systematic error in the reneging rate estimator. Finally, we present numerical results to illustrate that the systematic error increases with α .

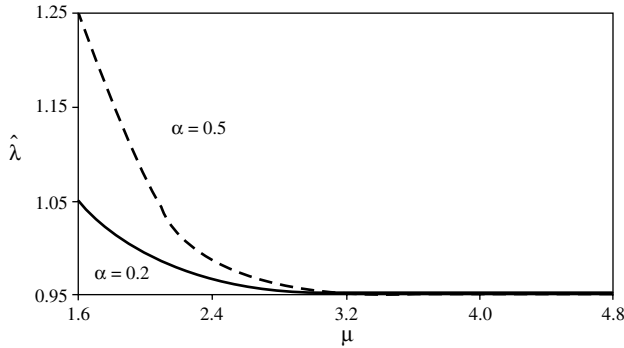
The steady-state probability distribution (and hence the systematic error) can be expressed in closed form only in the special cases $\alpha = 0$ and $\alpha = 1$. Therefore, our method of proof involves sample-path arguments and coupling. By this method, one can prove statements that are stronger than what we need. Specifically, our sensitivity analysis is concerned with comparisons of certain quantities in the *limit* as time goes to infinity. Instead, the sample-path arguments establish stochastic ordering of the relevant quantities for every time t . This approach works when varying the capacity μ , but not the double-order probability α (the relevant quantities are not ordered in the prelimit).

PROPOSITION 3. *The systematic error in estimating the demand rate λ is decreasing in μ for any fixed allocation $\mu_1 = p\mu$ and $\mu_2 = (1-p)\mu$, where $0 < p < 1$.*

PROOF. From Proposition 2, the systematic error in the demand rate is $\hat{\lambda} - \lambda = \lambda \alpha P^{00} = \lambda \alpha P(X_1(\infty) \leq 0, X_2(\infty) \leq 0)$. In three steps, we will prove that $P(X_1(\infty) \leq 0, X_2(\infty) \leq 0)$, the steady-state probability that both distributors are out of stock, decreases with μ . First, consider the uniformized discrete-time Markov chain with one-step transition probabilities equal to the corresponding transition rates of the continuous-time Markov chain, divided by $v \geq 2(\lambda + \lambda\alpha) + \mu + 3M\eta$. Transitions from a state to itself are allowed to ensure that the transition probabilities sum up to 1. The steady-state distribution of the uniformized discrete-time chain is identical to that of the continuous-time Markov chain.

Second, let $\mu_L < \mu_H$, and denote by $(-X_1^L, -X_2^L, D^L)$ the state of the system when $\mu = \mu_L$. Similarly, denote by $(-X_1^H, -X_2^H, D^H)$ the corresponding state descriptor when $\mu = \mu_H$. We use *sample-path coupling arguments* to show that $(-X_1^L, -X_2^L, D^L) \geq^{st} (-X_1^H, -X_2^H, D^H)$, where \geq^{st} denotes stochastic ordering. This will imply, in particular, that $P(X_1^L(\infty) \leq 0, X_2^L(\infty) \leq 0) \geq P(X_1^H(\infty) \leq 0, X_2^H(\infty) \leq 0)$. The sample-path coupling argument works as follows: Let $Z^L = (-X_1^L + D^L, -X_2^L + D^L, D^L)$ and $Z^H = (-X_1^H + D^H, -X_2^H + D^H, D^H)$. Note that $Z^L(\cdot) \leq M$ and $Z^H(\cdot) \leq M$ (where the inequalities hold component-wise). We can construct versions of $Z^L(\cdot)$ and $Z^H(\cdot)$ (which for notational simplicity are denoted the same as the original processes), such that $Z^L(\cdot) \geq Z^H(\cdot)$ with probability 1. Specifically, we let $v = 2(\lambda + \lambda\alpha) + \mu^H + 3M\eta$ and assume that $Z^L(0) \geq Z^H(0)$. Then, by coupling the transitions of both chains and using induction on n , it follows that $Z^L(n) \geq Z^H(n)$.

Figure 2 Maximum-Likelihood Estimator for the Demand Rate as a Function of the Production Capacity for the System with $\lambda = 0.95$, $\eta = 0.1$, $B = 5$, and $\mu_1 = \mu_2 = \mu/2$



Third, the relationship $Z^L(\cdot) \geq^{st} Z^H(\cdot)$ implies that $(-X_1^L, -X_2^L, D^L) \geq^{st} (-X_1^H, -X_2^H, D^H)$, because $(-X_1, -X_2, D)$ can be expressed as an increasing function of $Z = (-X_1 + D), -(X_2 + D), D$. \square

The second step in this proof breaks down if the manufacturer fills distributors' orders FIFO. The online appendix gives an alternative proof for FIFO assuming $\alpha = 1$.

Figure 2 illustrates how the systematic error in estimating the demand rate λ decreases with the capacity μ .

In contrast, the systematic error in estimating the renegeing rate η initially increases with μ .

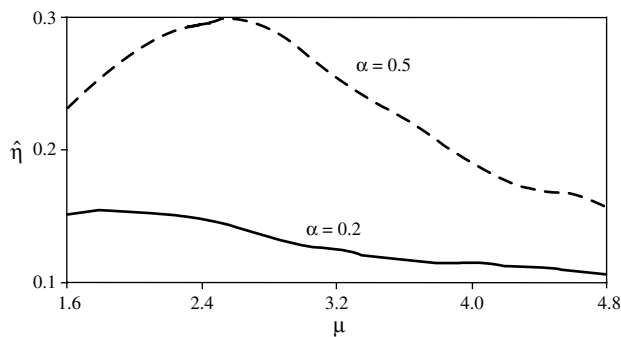
PROPOSITION 4. Let $\hat{\eta} = \lim_{T \rightarrow \infty} \hat{\eta}(T)$ be the limit of the renegeing rate MLE as the time horizon grows to infinity. Then, $\hat{\eta} - \eta$ is increasing in μ at $\mu = 0$.

PROOF. It is easy to see that when $\mu = 0$, $\hat{\eta} - \eta = 0$. However, $\hat{\eta} - \eta > 0$ for all $\mu > 0$. \square

However, the systematic error in estimating the renegeing rate η may subsequently decrease with μ as shown in Figure 3.

To understand this nonmonotonicity, one must closely examine the source of the error. Orders are cancelled for one of the following reasons: (1) a customer reneges and consequently cancels all outstanding orders (two orders are cancelled if he has placed

Figure 3 Maximum-Likelihood Estimator for the Renegeing Rate as a Function of the Production Capacity for the System with $\lambda = 0.95$, $\eta = 0.1$, $B = 5$, and $\mu_1 = \mu_2 = \mu/2$



an order with both distributors), and (2) a customer receives the product from one distributor and cancels a duplicate order from the other distributor (one order is cancelled). Order cancellations of Type 1 do not contribute to the systematic error because customers that double order are counted twice as part of the backlog and twice as an order cancellation. Order cancellations of Type 2 are the ones that cause the systematic error. That is, the systematic error occurs because the cancellation of a duplicate order is counted as a renegeing customer, but no customer is actually renegeing.

The effect of μ on order cancellations of Type 2 is twofold. Having greater production capacity reduces the proportion of time that the system spends in the backordered states, so fewer duplicate orders occur. This tends to reduce order cancellations of Type 2. On the other hand, as the capacity μ increases, a larger fraction of all duplicate orders are cancelled due to service completion (Type 2) rather than because of renegeing (Type 1). This tends to increase the number of order cancellations of Type 2. This second effect is the one that tends to increase the error. Note that $\hat{\eta} - \eta$ can be written as $\mu / (EX_1^- + EX_2^-) \cdot P(D(\infty) > 0)$, where the first term in the product is increasing in μ , while the second term is decreasing. This explains why $\hat{\eta}$ is not monotone in μ .

We have observed numerically that the systematic error is increasing with the double-order probability as shown in Figure 4. Furthermore, the systematic error in the demand rate is only slightly higher if the distributors use FIFO sequencing, rather than giving priority to the customers that double order. However, giving priority to the customers that double order increases the error in estimating the renegeing rate. That is because giving priority to customers that double order increases the frequency of order cancellations of Type 2 (in which a customer receives the product and cancels a duplicate order).

4. The Manufacturer's "Optimal" Capacity Investment

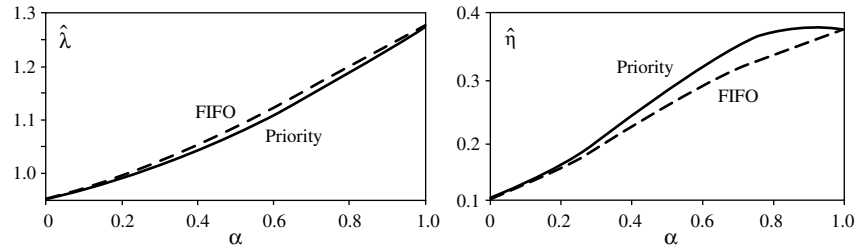
The conventional wisdom following Cisco's notorious inventory write-off is that duplicate ordering by customers will lead a manufacturer to overinvest in capacity. This is not necessarily true. In this section, we show that, in fact, a manufacturer that mistakes double orders for true customers' orders may buy too little capacity.

Suppose that the manufacturer chooses capacity according to²

$$\min_{\mu} [c\eta E(X_1^- + X_2^- - D) + k\mu], \tag{1}$$

² More generally, one may consider choosing capacities μ_1, μ_2 to minimize the cost function $[c\eta E(X_1^- + X_2^- - D) + k(\mu_1 + \mu_2)]$. However, we have observed in all numerical experiments that the

Figure 4 Maximum-Likelihood Estimator for the Demand Rate and Reneging Rate as a Function of the Double-Order Probability α for the System with $\lambda = 0.95$, $\eta = 0.1$, and $B = 5$



where the decision variable μ is the total capacity devoted to both distributors; c is the manufacturer’s contribution per unit sold, so the first term in the objective function is the expected cost of lost sales; the second term is the cost of capacity. (Without loss of generality, we will assume that $c = 1$.) In practice, a distributor’s inventory policy depends on the delivery lead time and hence upon the manufacturer’s capacity. However, in solving for the optimal capacity in (1) we disregard strategic interaction, and assume that the base stock level B per distributor is fixed. The solution to (1) can be interpreted as the manufacturer’s best response to the distributors’ inventory policies.

Conventional wisdom is that consumers are increasingly impatient, and this increases capacity requirements. However, the optimal total capacity μ is *not* monotone increasing in the reneging rate η . Intuitively, when customers become extremely impatient (as $\eta \rightarrow \infty$), the optimal capacity may drop down to zero. This is because the capacity required to capture a certain amount of sales from customers with an increasing reneging rate may become too high, and hence prohibitively expensive.³

Therefore, whereas overestimating λ always leads the manufacturer to buy more capacity, overestimating η may lead the manufacturer to buy less capac-

ity.⁴ Furthermore, for fixed λ and η , the optimal level of capacity investment may be increasing in α . For these two reasons, a manufacturer that is unaware of duplicate ordering may purchase *too little* capacity. We demonstrate this result through numerical examples.

Consider a system in which $\alpha = 1$ (every customer that must wait for the product will place a duplicate order), but the manufacturer believes that $\alpha = 0$. For this system, we have the steady-state probability distribution in closed form (see Appendix A), and can therefore compute the limiting MLEs and the cost function exactly. The expected rate of lost sales is strictly lower in the system with $\alpha = 1$ than in the system with $\alpha = 0$ because, in choosing to double order, each customer increases the likelihood that he will obtain the product before reneging. Effectively, inventory and capacity are pooled in the system with $\alpha = 1$. Furthermore, as illustrated by Figure 5, when the capacity μ is very small, the marginal value of capacity is greater in the system with $\alpha = 1$ than in the system with $\alpha = 0$ (i.e., increasing μ does more to reduce lost sales when $\alpha = 1$ than when $\alpha = 0$). However, if the capacity μ is sufficiently large, additional capacity is more beneficial when $\alpha = 0$ than when $\alpha = 1$. Therefore, if the manufacturer knows the true demand rate and reneging rate, but incorrectly assumes that $\alpha = 0$, she will underinvest when the cost of capacity k is large and overinvest when the cost of capacity k is small. When the manufacturer’s capacity is pooled and distributors’ orders are filled FIFO, duplicate orders are still beneficial in pooling the distributors’ inventories. The pooling effects of duplicate orders are less pronounced with pooled capacity than with dedicated capacities, but remain significant and qualitatively the same as the effects illustrated in Figure 5.

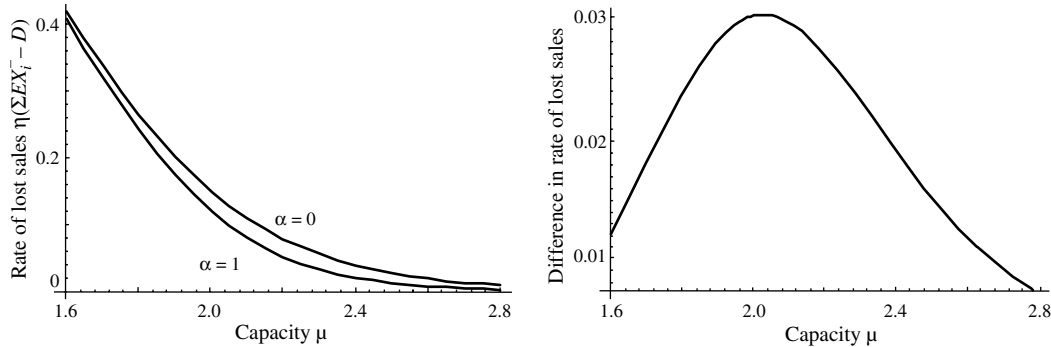
Suppose that the manufacturer has been operating the system at some fixed initial level of capacity μ , and uses the MLEs $\hat{\lambda}$ and $\hat{\eta}$ (which depend on the level of capacity μ) to compute his “optimal” capacity investment. Figure 6 shows that when the initial capacity level is larger than the demand rate and

solution to this optimization problem is symmetric: $\mu_1 = \mu_2 = \mu/2$. Intuitively, if the manufacturer dedicates greater capacity to one distributor, that distributor tends to have greater inventory. When his inventory reaches the base stock level the greater capacity is idled, while the other distributor typically has outstanding orders. Hence, asymmetric capacity is inefficient. Only in the degenerate case that all customers duplicate order ($\alpha = 1$) and distributors do not carry inventory ($B = 0$) does asymmetric capacity perform as well as symmetric capacity. In this case, the expected rate of lost sales is constant for all μ_1 and μ_2 such that $\mu_1 + \mu_2 = \mu$. Asymmetric capacity might, however, yield greater system profit if the distributors choose their inventory levels in Nash equilibrium after observing capacity.

³ This result is proven in Armony et al. (2005) for the special case that $B = 0$ and $\alpha = 0$, a make-to-order system without double orders. We have observed nonmonotonicity in numerical examples with $B > 0$ and $\alpha \in (0, 1)$, and conjecture that the result is true in general.

⁴ In fact, it is plausible that under any setting in which one overestimates the reneging rate, underinvestment in capacity may occur (even in the absence of duplicate orders).

Figure 5 The Difference in the Expected Rate of Lost Sales in the Case $\alpha = 0$ and the Case $\alpha = 1$ for a System with $\lambda = 1$, $\eta = 0.2$, $B = 10$, and $\mu_1 = \mu_2 = \mu/2$



the cost of capacity is relatively large, this “optimal” capacity investment will be strictly smaller than the true optimal capacity. That is, the manufacturer will underinvest in capacity.

One might suspect that this underinvestment phenomenon occurs because the manufacturer devotes a fraction of her production capacity exclusively to each distributor. However, underinvestment also occurs with resource pooling. Figures 5 and 6 can be essentially reproduced for a system in which the manufacturer uses a FIFO sequencing policy and the distributors prioritize their double orders. Here, the steady-state probabilities are easily calculated for $\alpha = 1$, but when $\alpha = 0$ things are more intricate; however, one can obtain fairly simple expressions for those steady-state probabilities in the pure loss model (i.e., $\eta = \infty$; see Appendix B).

Now, let us suppose that the manufacturer repeatedly runs the system for long enough to compute the estimators $\hat{\lambda}$ and $\hat{\eta}$, and then adjusts capacity to the “optimal” level. In all of our numerical experiments, the capacity converges to an equilibrium that appears to be “optimal” if the manufacturer assumes that $\hat{\lambda}$ and $\hat{\eta}$ (evaluated at the current capacity level) are the true demand rate and renegeing rate. Figure 7 shows

that in equilibrium the manufacturer overinvests in capacity.

5. Maximum-Likelihood Estimation When $\alpha > 0$

If the manufacturer is aware of the potential for double orders and observes the system continuously, she can recognize a double order whenever both distributors order simultaneously, or cancel an order simultaneously. One may argue that in reality no two events will occur at exactly the same time. As a practical alternative, if the manufacturer has visibility of end-customers’ identities, she can pair two orders made by the same customer at approximately the same time, and recognize a double order. In this section, we spell out the MLEs of λ , η , and α in the case of full information (continuous observation or visibility of customers’ identities). These estimators are valid for a general shipment schedule from the manufacturer to the distributors.

To write down the MLEs for these three parameters, we need to introduce some additional notation. Let $N(T)$ denote the total number of orders made by both distributors in the period $[0, T]$ (accounting only once for those orders that are immediately switched

Figure 6 True Optimal Capacity and the “Optimal” Capacity Investment for a Manufacturer Who Assumes that $\alpha = 0$, $\lambda = \hat{\lambda}$, and $\eta = \hat{\eta}$ for the System with $\alpha = 1$, $\eta = 0.2$, $\lambda = 1$, $B = 10$, and Initial Capacities $\mu_1 = \mu_2 = 1.8$

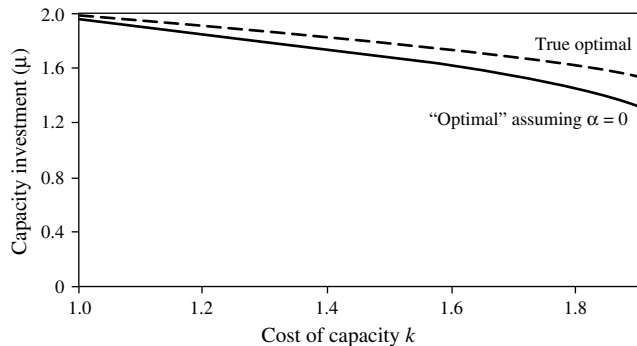
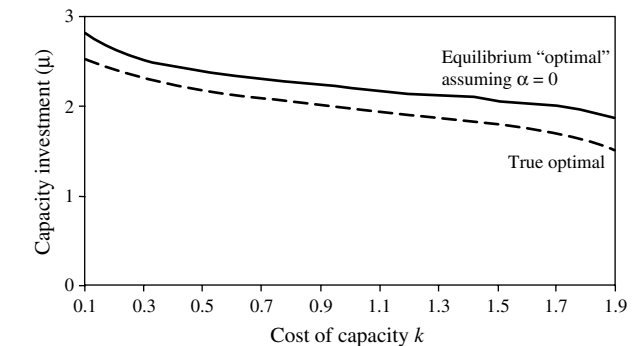


Figure 7 True Optimal Capacity and Equilibrium “Optimal” Capacity Investment for a Manufacturer Who Assumes that $\alpha = 0$, $\lambda = \hat{\lambda}$, and $\eta = \hat{\eta}$, for the System with $\alpha = 1$, $\eta = 0.2$, $\lambda = 1$, and $B = 10$



from an out-of-stock distributor to a distributor with the item in inventory, but twice for double orders). Also, let $Z(T)$ correspond to the total number of order cancellations from both distributors in the same time interval. Let $D_{in}(T)$ be the total number of duplicate orders made between time 0 and time T (counting only those duplicate orders that occur while both distributors are out of stock), and let $D_{out}(T)$ be the total number of double orders (counted in $D_{in}(T)$) that have both been cancelled by time T . Finally, let $Sw(T)$ be the number of customers who switch from an out-of-stock distributor to one with positive inventory in the time interval $[0, T]$, and denote by $N_-(T)$ the total number of arriving customers who find the first distributor they turn to being out of stock. As in the case with $\alpha = 0$ analyzed in §3, maximizing the likelihood function given continuous-time transition information yields the MLEs described in the following proposition. The proof of Proposition 5 is very similar to that of Proposition 1, and hence is omitted.

PROPOSITION 5. *The MLEs of λ , η , and α are given by*

$$\begin{aligned}\tilde{\lambda}(T) &= \frac{N(T) - D_{in}(T)}{2T}, \\ \tilde{\eta}(T) &= \frac{Z(T) - D_{out}(T)}{\int_0^T [X_1^-(t) + X_2^-(t) - D(t)] dt}, \\ \tilde{\alpha}(T) &= \frac{D_{in}(T) + Sw(T)}{N_-(T)}.\end{aligned}$$

These estimators are consistent: $(\tilde{\lambda}(T), \tilde{\eta}(T), \tilde{\alpha}(T)) \rightarrow (\lambda, \eta, \alpha)$ as $T \rightarrow \infty$.

REMARK. To ensure consistency of the estimators for an arbitrary shipment schedule, all that needs to be verified is that as $T \rightarrow \infty$, the system spends enough time in both the backordered and the positive inventory states. This condition is guaranteed to hold under our Poisson production assumption with dedicated capacities or with pooled capacity and FIFO fulfillment of distributors' orders.

Before the manufacturer can compute the estimators for $\tilde{\eta}(T)$ and $\tilde{\alpha}(T)$, some customers must experience a stockout. By reducing the production rate μ and risking some lost sales, the manufacturer can obtain an improved estimate of her customer's tolerance for delay and willingness to switch to an alternative distributor of her product.

5.1. Periodic Observation of Inventory Levels

In practice, production occurs in batches, distributors order by the truckload, and the manufacturer observes distributors' inventory levels infrequently, if at all. Specifically, prior to the spring of 2001, Cisco did not have information systems in place to track distributors' inventory levels. Management maintained a record of all shipments, and could, by placing a phone call to a distributor, check on the

aggregate dollar value of Cisco products in inventory. Such check-ups occurred on an infrequent, ad hoc basis (Kothari 2001). Recently, Cisco and many other high-tech manufacturers, including Sony, HP, Toshiba, and Sun, have installed software that enables them to review distributors' inventory levels at the SKU level on a weekly, and in some cases daily, basis (Chua 2003).

Let us assume that at discrete times t_k with $0 < t_1 < t_2 < \dots < t_K$, the manufacturer has some interaction with distributor i_k ($i_k \in \{1, 2\}$); either the manufacturer delivers $S^k > 0$ units of the product to distributor i_k or the manufacturer observes the inventory level $X_{i_k}(t_k)$ (in the latter case $S_k \doteq 0$). Recall that X_i may take negative values: X_i^- is the number of backorders for distributor i , including double orders. We assume that the manufacturer cannot identify double orders (D is hidden from the manufacturer). On the contrary, each distributor knows which of his customers have placed a duplicate order, and gives priority to serving these customers (to avoid losing a sale to the other distributor). Therefore,

$$\begin{aligned}X_{i_k}(t_k) &= X_{i_k}(t_k^-) + S_k \quad \text{and} \\ X_j(t_k) &= X_j(t_k^-) + S_k \wedge D(t_k^-) \quad \text{for } j \neq i_k, \\ D(t_k) &= [D(t_k^-) - S^k]^+.\end{aligned}$$

We will derive the manufacturer's MLE for η , α , and λ , given the discrete-time observations of inventory levels and the schedule of deliveries. For each $k = 1, \dots, K$ and $t \in [t_{k-1}, t_k)$, the stochastic process $(X_1(t), X_2(t), D(t))$ is a continuous-time Markov chain with generator matrix A given by

$$\begin{aligned}A_{(X_1, X_2, D), (X_1-1, X_2, D)} &= A_{(X_1, X_2, D), (X_1, X_2-1, D)} = \lambda \\ &\quad \text{if } X_1 > 0 \text{ and } X_2 > 0, \\ A_{(X_1, X_2, D), (X_1-1, X_2, D)} &= \lambda(1 + \alpha) \quad \text{if } X_1 > 0 \text{ and } X_2 \leq 0, \\ A_{(X_1, X_2, D), (X_1, X_2-1, D)} &= \lambda(1 + \alpha) \quad \text{if } X_1 \leq 0 \text{ and } X_2 > 0, \\ A_{(X_1, X_2, D), (X_1-1, X_2, D)} &= A_{(X_1, X_2, D), (X_1, X_2-1, D)} = \lambda(1 - \alpha) \\ &\quad \text{if } X_1 \leq 0 \text{ and } X_2 \leq 0, \\ A_{(X_1, X_2, D), (X_1-1, X_2-1, D+1)} &= 2\lambda\alpha \quad \text{if } X_1 \leq 0 \text{ and } X_2 \leq 0, \\ A_{(X_1, X_2, D), (X_1+1, X_2, D)} &= \eta[X_1 - D]^+ \quad \text{if } X_1 < 0, \\ A_{(X_1, X_2, D), (X_1, X_2+1, D)} &= \eta[X_2 - D]^+ \quad \text{if } X_2 < 0, \\ A_{(X_1, X_2, D), (X_1+1, X_2+1, D-1)} &= \eta D \quad \text{if } D > 0, \\ A_{(X_1, X_2, D), (X_1', X_2', D')} &= 0 \quad \text{otherwise,}\end{aligned}$$

with the obvious modification to reflect our state space truncation: $D \leq M$ and $X_i^- - D \leq M$. The continuous-time Markov chain is completely characterized by the

initial distribution and the generator A . In particular, the transition matrix $P(t)$ is given by

$$P(t) = \exp(At) = \sum_{n=0}^{\infty} \frac{(At)^n}{n!},$$

that is, $P_{(X_1, X_2, D), (X'_1, X'_2, D')}(t)$ is the conditional probability that $(X_1(s+t), X_2(s+t), D(s+t)) = (X'_1, X'_2, D')$ given that $(X_1(s), X_2(s), D(s)) = (X_1, X_2, D)$ and $t_{k-1} \leq s < s+t < t_k$. Let q^0 denote the initial distribution of $(X_1(0), X_2(0), D(0))$. For example, if the distribution channel is empty at time zero, then

$$q^0_{(X_1, X_2, D)} = 1 \quad \text{if } X_1 = X_2 = D = 0,$$

$$q^0_{(X_1, X_2, D)} = 0 \quad \text{otherwise.}$$

Because the deliveries $(S_1^k, S_2^k)_{k=1..K}$ are known, the likelihood function can be computed recursively, as follows. For $k = 1, 2, \dots$ we will compute π^k , the unnormalized conditional distribution for $(X_1(t_k^-), X_2(t_k^-), D(t_k^-))$ given the manufacturer's observations and deliveries up to time t_{k-1} , and then compute q^k , the unnormalized conditional distribution for $(X_1(t_k), X_2(t_k), D(t_k))$ given the manufacturer's observations and deliveries up to time t_k . Define $t_0 = 0$ and

$$\pi^k \triangleq q^{k-1}P(t_k - t_{k-1}).$$

In the case that a delivery is made at time t_k ($S^k > 0$), then

$$q^k_{(X_1, X_2, 0)} \triangleq \begin{cases} \sum_{d=0}^{S_k} \pi^k_{(X_1-S_k, X_2-d, d)} & \text{if } i_k = 1, \\ \sum_{d=0}^{S_k} \pi^k_{(X_1-d, X_2-S_k, d)} & \text{if } i_k = 2, \end{cases}$$

$$q^k_{(X_1, X_2, D)} \triangleq \pi^k_{(X_1-S_k, X_2-S_k, S_k)} \quad \text{for } D > 0.$$

In the case that an observation is made at time t_k ($S^k = 0$),

$$q^k_{(X_1, X_2, D)} \triangleq \begin{cases} \pi^k_{(X_1, X_2, D)} & \text{if } X_{i_k} = X_{i_k}(t_k), \\ 0 & \text{otherwise.} \end{cases}$$

Then, the likelihood function $\mathcal{L}(\lambda, \eta, \alpha)$ is given by

$$\mathcal{L}(\lambda, \eta, \alpha) = |q^K|.$$

$\mathcal{L}(\lambda, \eta, \alpha)$ is the probability that under the delivery schedule $\{S_k, i_k, t_k\}_{k=1, \dots, K}$, the distributor i_k has inventory level $X_{i_k}(t_k)$ at time t_k for which $S_k = 0$. This establishes our main result:

PROPOSITION 6. *Suppose that the manufacturer cannot identify double orders, and observes each distributor's inventory level at discrete points in time. In this case, the MLEs of λ, η , and α are given by*

$$(\tilde{\lambda}(t_K), \tilde{\eta}(t_K), \tilde{\alpha}(t_K)) = \arg \max_{(\lambda, \eta, \alpha) \in \mathbb{R}_+^3} [\mathcal{L}(\lambda, \eta, \alpha)].$$

In extensive simulation experiments, we have found that for systems with a small base stock level $B \leq 10$

and $M = 10$, the estimator is consistent. Unfortunately, the time required to compute the matrix exponential $\exp(At)$ grows exponentially with B and M , and we have been unable to directly compute the MLE for systems with $B \geq 20$ or $M \geq 20$. Clearly, a more efficient estimator will be needed in practice. The expectation maximization (EM) algorithm (Elliott et al. 1995) can be used to compute a series of parameter values that converges to $(\tilde{\lambda}(t_K), \tilde{\eta}(t_K), \tilde{\alpha}(t_K))$ in a manner that avoids direct computation of the matrix exponential. An alternative, efficient method of moments estimators based on periodic sampling of a continuous-time Markov process has recently been developed in the finance literature (Hansen and Scheinkman 1995, Duffie and Glynn 2004).

6. Concluding Remarks

Our results suggest that Cisco's write-off was caused by estimation errors and cannot be blamed entirely on the economic downturn. Any manufacturer that fails to account for duplicate orders will overestimate the demand rate and the renegeing rate, and therefore err in capacity planning. Business-press pundits have attributed Cisco's multibillion-dollar overinvestment in capacity to a severe component shortage followed by a drop in demand, and failure by Cisco to recognize this downturn because of duplication in the order backlog. However, our analysis shows that excess capacity can be an insidious, chronic problem even under stable demand conditions. An acute drop in demand just exacerbates the problem. When the cost of capacity is relatively low, so that the production rate is greater than the demand rate, the error in estimating the demand rate is small, but the overestimate in the renegeing rate is large. Because customers appear to be very sensitive to delay, the manufacturer does not realize that she has too much capacity.

Surprisingly, we have also observed that when the cost of capacity is very high and the manufacturer is unaware of duplicate ordering, she may invest *too little* in capacity. At a low level of capacity, customers' tendency to switch to an alternative distributor reduces the number of lost sales and increases the marginal value of capacity.

For the manufacturer that *is* monitoring double orders, we give MLEs for the demand rate, the renegeing rate, and the probability that a customer will double order when forced to wait. These are valid for any production, transportation, and inventory policy. The basic MLE assumes that the manufacturer has real-time visibility of distributors' inventory levels, either through sophisticated software *or* because the distributor follows a base-stock policy (so the manufacturer can infer the inventory level from her own order queue). More commonly, distributors order in batches and the manufacturer observes distributors'

inventory levels infrequently if at all; we also provide the MLE for this setting.

An important insight is that the manufacturer must experience backorders and lose some customers before she can estimate her customers' tolerance for delay. In models with lost sales, other researchers have shown that carrying more inventory results in a better estimate of the demand rate. Our model is distinctive in that customers will wait, although only for a limited time, before the sale is lost. Carrying less inventory and/or reducing the production capacity results in a better estimate of the renegeing rate and the double-order probability, without affecting the estimate of the demand rate.

We have assumed a fixed probability that a customer duplicates orders, given that his distributor is out of stock. In reality, duplicate ordering is history dependent. Confronted with a long lead time, some customers will make the effort to seek out an alternative distributor. Having learned about their alternatives, these customers are more likely to duplicate-order in the future (Kothari 2001). This can be modeled as a Markov process with the double-order probability as a hidden state variable. Then, maximum-likelihood estimation requires repeated numerical evaluation of an exponential function of the generator matrix (a computationally intensive procedure). This is impractical for industrial-sized problems. Ongoing research will develop efficient method-of-moments estimators based on periodic sampling of a continuous-time (partially observed) Markov process, drawing on methods from the finance literature (Duffie and Glynn 2004).

Since 2001, Cisco has increased visibility and tightened control of its distribution channels. New information systems provide visibility of distributors' inventory levels and the ability to control the purchase price for resellers (www.comergent.com). Resellers that share demand information and provide a high service level to customers are rewarded with a reduced price for Cisco hardware. With survey information from resellers and corporate customers, Cisco has improved its demand forecasting (*Business Week* 2002a). In another new program, Cisco owns the hardware in the distribution channel, sets the price for the end customer, and pays channel partners. Resellers charge the end customer for the service of configuration, rather than selling hardware. Adoption of this new program has been slow, perhaps because channel partners are unwilling to cede information and control.

High-tech manufacturers including Sony, HP, Toshiba, and Sun have recently implemented iGINE software for monitoring channel inventory. According to Chua I-Pin, Vice-President of iGINE, manufacturers may need to provide significant financial incentives for distributors to share inventory data. Validation

is challenging, as the data sets are complex. For example, in the Asia-Pacific region, Hewlett-Packard monitors more than 100,000 SKUs at 100 Tier-1 wholesalers, and 300 Tier-2 resellers. Distributors typically will not share the identity of backordered customers or price information. In shortage conditions, distributors typically increase their prices, so demand estimation is complicated by the hidden variable of price as well as duplicate orders. With visibility of inventory levels, manufacturers are moving beyond simple FIFO order fulfillment or proportional allocation, and beginning to replenish inventory based on the overstock or understock conditions at various distributors. Will distributors reduce their inventory to the detriment of the manufacturer? Further research is needed to address dynamic strategic interaction with information asymmetry and estimation.

An online appendix to this paper is available at <http://mansci.pubs.informs.org/ecompanion.html>.

Acknowledgments

The authors thank Tushar Kothari, Vice-President of Distribution at Cisco; and Chua I-Pin, Vice-President of Operations at iGINE for discussions on duplicate ordering and inventory visibility in channel management. They thank Li Chen for numerical analysis, Halina Frydman for advice on statistical estimation, and Sridhar Seshadri for guidance in the sensitivity analysis. Finally, they thank Bill Lovejoy and the anonymous associate editor and referee for suggestions on model formulation, presentation, and literature.

Appendix A. The Steady-State Probabilities

When $\alpha = 1$

We describe how to derive the steady-state probabilities for the system with $\alpha = 1$, and the exclusive production capacity devoted to each distributor. Note that the states (X_1, X_2) where $X_1 < 0$ and $X_2 > 0$ (and vice versa) cannot be accessed. Therefore, one can partition the complete state space into the following three mutually exclusive sets: (1) $S^- := \{(-n, -n); n > 0\}$, (2) $S^{0+} := \{0\} \times \{0, 1, \dots, B\} \cup \{0, 1, \dots, B\} \times \{0\}$, and (3) $S^+ := \{1, \dots, B\} \times \{1, \dots, B\}$. It is straightforward to calculate the steady-state probabilities for the Markov chains that are restricted to each of these three sets, and hence to calculate the steady-state probabilities for the whole chain, using the transition rates between these sets. The resulting expressions are cumbersome. For example, the steady-state probability that both distributors have outstanding orders is

$$\pi(S^-) = \frac{\pi_{(0,0)}^+ 2\lambda}{\pi_{(0,0)}^+ 2\lambda + \pi_{(-1,-1)}^-(\mu + \eta)},$$

where

$$\pi_{(-1,-1)}^- = \left[\sum_{n=1}^{\infty} \frac{(2\lambda)^{n-1}}{\prod_{j=2}^n (\mu + j\eta)} \right]^{-1} \quad \text{and} \quad \pi_{(0,0)}^+ = \frac{q_1}{q_1 + q_2},$$

with

$$q_1 = \lambda \left(\frac{1 - \rho_1}{1 - \rho_1^B} \rho_1^{B-1} + \frac{1 - \rho_2}{1 - \rho_2^B} \rho_2^{B-1} \right),$$

$$q_2 = c \left(\mu_1 \frac{(2\rho_1)^B (1 - 1/(2\rho_2)^B)}{2\rho_2(1 - 1/(2\rho_2))} + \mu_2 \frac{1 - (2\rho_1)^B}{1 - 2\rho_1} \right),$$

$$c = \left[\frac{1 - (2\rho_1)^{B+1}}{1 - 2\rho_1} + \frac{(2\rho_1)^B (1 - 1/(2\rho_2)^B)}{2\rho_2(1 - 1/(2\rho_2))} \right]^{-1}, \quad \rho_i = \frac{\lambda}{\mu_i},$$

$\rho_i \neq 1/2, 1, i = 1, 2$. Similar expressions apply when $\rho_i = 1/2$ or 1 for $i = 1$ or 2.

Appendix B. The Steady-State Probabilities with FIFO

In this appendix, we describe how to calculate the steady-state probabilities when the manufacturer uses a FIFO policy, the distributors prioritize their double orders, orders made when distributors are out of stock are lost (that is, $\eta = \infty$), and there are no double orders ($\alpha = 0$). Under these assumptions, the state descriptor $C = (c(1), c(2), \dots, c(n))$ (with n the total number of outstanding orders, and $c(k) \in \{1, 2\}$ is the distributor who made the order that is currently in position k in line ($k = 1, \dots, n$)) is a continuous-time Markov chain. Let $Y_i(C) = \sum_{k=1}^n 1_{\{c(k)=i\}}$, $i = 1, 2$, be the number of orders from distributor i currently in queue; then $Y_i(C)$ is constrained to the set $\{0, \dots, B\}$. Without this constraint, the steady state of the system has a product form, and hence it has this form on the constrained state space as well. In summary, the steady-state distribution is as follows:

$$\pi(C) = \frac{b}{A} \left(\frac{\rho}{2} \right)^n,$$

where $\rho = 2\lambda/\mu$, $b = (1 - \rho)/(1 - \rho^{2B+1})$ if $\rho \neq 1$, and $b = 1/(2B + 1)$ when $\rho = 1$. Also,

$$A = 1 - b \sum_{n=B+1}^{2B} M_n \left(\frac{\rho}{2} \right)^n \quad \text{with} \quad M_n = 2 \sum_{k=B+1}^n \binom{n}{k}.$$

The loss rate in this system is

$$\begin{aligned} \lambda P(Y_1(C) = B) + \lambda P(Y_2(C) = B) &= 2\lambda P(Y_1(C) = B) \\ &= 2\lambda \sum_{n=B}^{2B} \binom{n}{B} \frac{b}{A} \left(\frac{\rho}{2} \right)^n. \end{aligned}$$

References

- Anupindi, R., Y. Bassok. 1999. Centralization of stocks: Retailers vs. manufacturer. *Management Sci.* **45** 178–191.
- Anupindi, R., M. Dada, S. Gupta. 1998. Estimation of consumer demand with stock-out based substitution: An application to vending machine products. *Marketing Sci.* **17** 406–423.
- Armony, M., E. Plambeck, S. Seshadri. 2005. Convexity properties and comparative statics for M/M/S queues with balking and renegeing. Working paper, New York University, New York.
- Azoury, K. S. 1985. Bayes solution to dynamic inventory models under unknown demand distribution. *Management Sci.* **31** 1150–1160.
- Business Week*. 2001. Management lessons from the bust. (August 27) 104–110.
- Business Week*. 2002a. Cisco: Behind the hype. (January 22).
- Business Week*. 2002b. Without chips, the economy is harder to read. (May 6).
- Cachon, G. P. 2001. Stock wars: Inventory competition in a two-echelon supply chain with multiple retailers. *Oper. Res.* **49** 658–674.
- Cachon, G. P., A. G. Kok. 2002. How to (and how not to) estimate the salvage value in the newsvendor model. Working paper, University of Pennsylvania, Philadelphia, PA.
- Cachon, G. P., M. A. Lariviere. 1999. Capacity choice and allocation: Strategic behavior and supply chain performance. *Management Sci.* **45** 1091–1108.
- Chua, I.-P. 2003. Conversation with iGINE Vice-President of Operations. (May 21).
- Daley, D. J., L. D. Servi. 2001. Estimating customer loss rates from transactional data. Working paper, GTE Laboratories, Waltham, MA.
- Ding, X., M. C. Puterman, A. Bisi. 2002. The censored newsvendor and the optimal acquisition of information. *Oper. Res.* **50** 517–527.
- Duffie, D., P. Glynn. 2004. Estimation of continuous-time Markov processes sampled at random times. *Econometrica* **72** 1773–1808.
- Elliott, R., L. Aggoun, J. Moore. 1995. *Hidden Markov Models: Estimation and Control*. Springer, New York, 35–76.
- Hansen, L., J. Scheinkman. 1995. Back to the future: Generating moment implications for continuous time Markov processes. *Econometrica* **63** 767–804.
- Kothari, T. 2001. Conversation with Cisco vice president for distribution channels. (June 7).
- Lariviere, M. A., E. L. Porteus. 1999. Stalking information: Bayesian inventory management with unobserved lost sales. *Management Sci.* **45** 346–363.
- Larson, C. E., L. J. Olson, S. Sharma. 2001. Optimal inventory policies when the demand distribution is not known. *J. Econom. Theory* **101** 281–300.
- Lee, H. L., V. Padmanabhan, S. Whang. 1997. Information distortion in a supply chain: The bullwhip effect. *Management Sci.* **43** 546–558.
- Li, L. 1992. The role of inventory in delivery time competition. *Management Sci.* **38** 182–197.
- Lippman, S. A., K. F. McCardle. 1997. The competitive newsboy. *Oper. Res.* **45** 54–65.
- Lovejoy, W. S. 1990. Myopic policies for some inventory models with uncertain demand distribution. *Management Sci.* **36** 724–738.
- Lovejoy, W. S. 1992. Stopped myopic policies in some inventory models with generalized demand processes. *Management Sci.* **38** 688–707.
- Mahajan, S., G. van Ryzin. 2001. Inventory competition under dynamic consumer choice. *Oper. Res.* **49** 646–657.
- Mandelbaum, A., S. Zeltyn. 1998. Estimating characteristics of queueing networks. *Queueing Systems* **29** 75–197.
- Netessine, S., N. Rudi. 2003. Centralized and competitive inventory models with demand substitution. *Oper. Res.* **51** 329–335.
- Netessine, S., N. Rudi, Y. Wang. 2001. Dynamic inventory competition and customer retention. Working paper, University of Pennsylvania, Philadelphia, PA.
- Parlar, M. 1988. Game theoretic analysis of the substitutable product inventory problem with random demand. *Naval Res. Logist.* **35** 397–409.
- Ryan, S. M. 2003. Capacity expansion with lead times and autocorrelated random demand. *Naval Res. Logist.* **50** 167–183.
- Scarf, H. E. 1959. Bayes solution of the statistical inventory problem. *Ann. Math. Statist.* **30** 346–363.
- Thurm, S. 2001a. Missed signals: Behind Cisco's woes are some wounds of its own making. *Wall Street Journal* (April 1).
- Thurm, S. 2001b. Cisco expects tough times for rest of year. *Wall Street Journal* (August 8).
- Toktay, L. B., L. M. Wein. 2001. Analysis of a forecasting-production-inventory system with stationary demand. *Management Sci.* **47** 1268–1281.
- Van Mieghem, J. A. 2003. Capacity management, investment, and hedging: Review and recent developments. *Manufacturing Service Oper. Management* **5** 269–302.