Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Managing an Available-to-Promise Assembly System with Dynamic Short-Term Pseudo-Order Forecast

Long Gao, Susan H. Xu, Michael O. Ball,

To cite this article:
Long Gao, Susan H. Xu, Michael O. Ball, (2012) Managing an Available-to-Promise Assembly System with Dynamic Short-Term Pseudo-Order Forecast. Management Science 58(4):770-790. http://dx.doi.org/10.1287/mnsc.1110.1442

MANAGEMENT SCIENCE
Vol. 58, No. 4, April 2012, pp. 770–790
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.1110.1442

Managing an Available-to-Promise Assembly System with Dynamic Short-Term Pseudo-Order Forecast

Long Gao
The A. Gary Anderson Graduate School of Management, University of California, Riverside, Riverside, California 92521, long.gao@ucr.edu

Susan H. Xu
Smeal College of Business, The Pennsylvania State University, University Park, Pennsylvania 16802, shx@psu.edu

Michael O. Ball
Robert H. Smith School of Business, University of Maryland, College Park, Maryland 20742, mball@rhsmith.umd.edu

We study an order promising problem in a multiclass, available-to-promise (ATP) assembly system in the presence of *pseudo orders*. A pseudo order refers to a tentative customer order whose attributes, such as the likelihood of an actual order, order quantity, and confirmation timing, can change dynamically over time. A unit demand from any class is assembled from one manufactured unit and one inventory unit, where the manufactured unit takes one unit of capacity and needs a single period to produce. An accepted order must be filled before a positive delivery lead time. The underlying order acceptance decisions involve trade-offs between committing resources (production capacity and component inventory) to low-reward firm orders and reserving resources for high-reward orders. We develop a Markov chain model that captures the key characteristics of pseudo orders, including demand lumpiness, nonstationarity, and volatility. We then formulate a stochastic dynamic program for the ATP assembly system that embeds the Markov chain model as a short-term forecast for pseudo orders. We show that the optimal order acceptance policy is characterized by *class prioritization*, *resource-imbalance-based rationing*, and *capacity-inventory-demand matching*. In particular, the rationing level for each class is determined by a critical value that depends on the resource imbalance level, defined as the net difference between the production capacity and component inventory levels. Extensive numerical tests underscore the importance of the key properties of the optimal policy and provide operational and managerial insights on the value of the short-term demand forecast and the robustness of the optimal policy.

Key words: available-to-promise; pseudo orders; Markov; stochastic dynamic programming; prioritization; resource and demand matching; resource-imbalance-based rationing; short-term and long-term forecasts; robustness

History: Received January 22, 2009; accepted June 22, 2011, by Martin Lariviere, operations management. Published online in *Articles in Advance* November 4, 2011.

1. Introduction

In pursuing sales of higher-value products, it is common for sales personnel to closely track and classify the status of "prospects." In many cases, the information maintained on potential orders is quite rich, including some measure of the likelihood of an actual order and estimates of the order quantity, confirmation timing, and revenue. In fact, there is a vibrant business in "sales force automation" software that maintains such information and provides management access to the basic data as well as derived information such as revenue and sales forecasts (e.g., see http://www.business-software.com/).

For manufactured products with significant lead times, this information can also be of value in production planning. We refer to these potential orders as *pseudo orders*. A pseudo order can either evolve into a firm order, whose characteristics may or may not be consistent with those originally forecasted, or be canceled. The uncertainty in this evolution can result in highly volatile demand estimates. It is also common for managers to reject or delay some low-reward orders on hand to reserve component inventory and production capacity in anticipation of high-reward orders considered to be likely. For example, Toshiba uses an electronic available-to-promise (ATP) system to process orders from different classes of business customers. It employs a make-to-order framework for many of its product lines. Critical components such as the LCD screen are shared among several models. Its

sales division keeps track of pseudo-order information and critical resources are frequently reserved for high-priority future orders. Similar ATP systems and business practices are also used by Dell and Maxtor for order promising and fulfillment (Ball et al. 2004). However, these activities typically are ad hoc at best and not well integrated into business planning and control systems. Academic research is scant on this topic and also on how demand signalling or, more generally, volatile real-time information, should be captured and used to improve cross-functional decision making in supply chains.

Compared with conventional ATP, whose function consists of a simple database lookup into the master production schedule, advanced ATP is a decision-making mechanism. Its purpose is to mitigate the discrepancy between forecast-driven push activities and order-driven pull activities across a supply chain. Advanced ATP specifies decisions relative to a set of orders. Each order is satisfied by producing the required demand quantity before a particular delivery date. Through a product's bill of material (BOM), ATP needs to match the committed demand quantity to available system resources, such as production capacity and component inventory, and delivers the finished goods before a quoted delivery date.

The literature on advanced ATP systems implicitly assumes that all orders are firm orders. Motivated by the wide-spread practice of advanced ATP systems in diverse industries, this paper develops models and tools to integrate pseudo-order information into advanced ATP systems. Decisions to favor pseudo orders over firm orders typically result when a higher profitability or priority is associated with pseudo orders. Thus, our model assumes multiple order classes. For example, Dell segments its customers into transaction, relationship, and public/international customer classes. By doing so, Dell is able to better meet the needs of each customer class. The principal goals of this paper are threefold: (1) developing a modeling framework that explicitly captures the stochastic and dynamic nature of pseudo orders; (2) integrating pseudo-order information within an available-to-promise assembly (ATP-A) system and developing structural properties of the optimal order acceptance policy; and (3) investigating numerically the behavior of the optimal policy and when and how to use pseudo-order information, possibly containing *systematic forecast errors*, under a wide range of ATP-A system settings. Our model and results can help companies to systematically quantify pseudo-order information and feed information into ATP decision making. Moreover, our dynamic programming model based on the dynamic pseudo-order forecast reveals new insights and management principles that can help companies improve their order promising and fulfillment processes.

We consider a multiperiod ATP-A system that serves demands of multiple classes (or multiple product types). At the beginning of each period, a fixed amount of the planned component inventory and production capacity become available. A unit demand, independent of its class, is assembled from two major components: one manufactured component that needs one unit of production capacity and a single period to produce; this manufactured component is then assembled with one unit of component inventory (called *inventory* hereafter) to become a finished product. The assembly time is assumed negligible. Future demand information is captured by a *dynamic short-term forecast*. In each period, the firm receives new confirmed orders (i.e., realized pseudo orders) from different classes and also updates the short-term forecast. A committed order must be delivered during a fixed time window (called the delivery lead time hereafter). The objective is to make order acceptance decisions to maximize the expected profit, while taking the inventory holding and capacity idleness costs into consideration.

Our main results pertaining to the three aforementioned goals can be summarized as follows.

1.1. Markov Chain Models for Pseudo Orders

We develop a Markov chain modeling framework that captures several key characteristics of pseudo orders, including demand *lumpiness*, *nonstationarity*, and *volatility*. Here, demand lumpiness refers to the possibility of null demand, due to the nonnegligible probability of order cancellation; nonstationarity is the result of frequent demand information updates (changes in prospect status received from a sales force automation software, e.g., Toshiba's point-of-sales terminal order processing control (Ball et al. 2004)); and volatility means that the attributes of a pseudo order can vary substantially during its lifetime. The volatile nature of demand signals has been recognized by many but explicitly modeled by few (Duenyas et al. 1997, Kempf 2004). In this paper, we first develop a Markov chain model for each *individual* pseudo order to capture the evolution of a stochastic attribute of the order, such as the order quantity or confirmation date (the date that a pseudo order becomes a firm order). We then aggregate individual Markov chains into a single Markov chain that captures collective information of overall demand of each class based on the confirmation date. Demand aggregation is essential in practice because an aggregate demand forecast is less volatile and more accurate than individual demand forecasts, and can be readily integrated with the ATP execution. We note that it is possible demand information of some low-priority classes is not dynamically forecasted (for example, demand information from individual consumers may not be continuously

updated); in this case we let demand for such a class form independently and identically distributed (iid) random variables, similar to the conventional stochastic demand assumption in the inventory management literature.

1.2. The Optimal Policy in the ATP-A System with Pseudo Orders

We formulate a dynamic programming (DP) model for the ATP-A system that embeds our pseudo-order model as a short-term forecast. The ATP-A system uses two types of resources, *perishable* capacity and *nonperishable* inventory, to produce and assemble the end products. We show that three key principles drive the ATP-A decisions: *class prioritization*, *capacity-inventory-demand* (*CID*) *matching*, and *resource-imbalance-based (RIB) rationing*. *Class prioritization* means that confirmed orders should be accepted in a decreasing order of their profitabilities. *CID matching* states that the firm needs to strike a balance between the availabilities of perishable and nonperishable resources and the aligned resource availability should be matched with the accepted demand as closely as possible. *RIB rationing* specifies an inventory rationing level for each class to be reserved for higher-valued orders, and this inventory rationing threshold uniquely determines the capacity rationing threshold. In the literature, resource rationing proves to be an effective control mechanism to improve system performance in different problem contexts (e.g., the expected marginal seat revenue heuristic in airline industries and rationing policies in inventory systems; see, e.g., Talluri and Van Ryzin 2005, Frank et al. 2003). However, RIB rationing in our problem context is significantly different from the rationing control policies reported in the literature in two important aspects. First, the existing literature considers rationing control for a *single resource type* (mostly *nonperishable* inventory), whereas we consider rationing control for *both nonperishable resource* (inventory) and *perishable resource* (capacity) types. We show that the two-resource rationing control can be achieved by inventory rationing control alone, and the inventory rationing threshold depends only on the resource imbalance level between the two resource types rather than their individual levels. Therefore, our RIB rationing captures two notions in a two-resource rationing control: *resource reservation* and *resource balancing*. Second, in the literature inventory rationing generally means reserving currently available inventory for future use. In contrast, our RIB rationing allows the optimal inventory rationing threshold to be either *positive* or *negative*. When it is positive, the currently available inventory will be reserved for future use; when it is negative, future inventory (i.e., available inventory during the delivery lead time) will be assigned to the current demand. Consequently, RIB rationing serves two purposes: preventing the loss of future, high-valued orders via current inventory reservation and preventing the loss of current, high-valued orders via advance booking of future inventory.

1.3. Experiments on ATP-A Systems

We conduct extensive numerical tests to investigate the behavior of the optimal policies in ATP-A systems and the value of pseudo-order information. We first illustrate the three properties of the optimal policy—class prioritization, CID matching and RIB rationing—using a three-class ATP-A system. We then consider a special case of a two-class ATP *inventory* (ATP-I) system. This simpler system allows us to systematically investigate, with manageable computation efforts, the value of pseudo-order information and when and how to use such information. To understand the volatile nature of pseudo orders, we compare trade-offs between using a short-term forecast (our pseudo-order model) and a long-term forecast based on historical data. Further we ask, how robust is the optimal policy when pseudo-order signals contain systematic forecast errors? To a large extent, the answers to these questions determine the applicability of pseudo-order information in practice. We find that using the short-term forecast is *always* better than using the long-term forecast, provided that both forecasts are accurate. We also investigate a range of forecast error scenarios and show that when the short-term forecast reaches a certain level of inaccuracy then it is better to ignore it and rely on the long-term forecast, which tends to be more reliable.

The remainder of this paper is organized as follows. In §2 we briefly review the relevant literature. In §3 we present a pseudo-order model and embed it in a DP formulation of the ATP-A system. In §4 we develop the structural results for the ATP-A system and discuss several special cases and extensions. In §5 we conduct numerical experiments. In §6 we summarize our major contributions, offer managerial implications, and discuss future research directions. The proofs can be found in the electronic companion (available at http://longgao.wordpress.com/).

2. Literature Review

Our work is related to several research streams, including demand signal models, quantitative ATP decision models, and rationing models for production-inventory control. We briefly review each of the research streams below.

In the operations management literature, downstream demand signals have been modeled in several ways to study their effects on system performance. One body of the literature applies a time series framework, using a Martingale model of forecast evolution

(MMFE) model in particular, to capture the evolution of the aggregate demand forecast (Heath and Jackson 1994, Aviv 2001, Gallego and Özer 2001, Özer 2003, Özer and Wei 2004, Milner and Kouvelis 2005). Another stream of the literature focuses on the impact of biased demand forecasts, or imperfect advanced demand information, on inventory management in different problem settings (see, e.g., DeCroix and Mookerjee 1997, Tan et al. 2007, Gayon et al. 2009, Li et al. 2011, among others). The third approach uses Markov chains to study aggregate demand forecasts in inventory control problems (Song and Zipkin 1993, Chen and Song 2001, Sethi and Cheng 1997, among others). Most of the above literature considers only demand quantity uncertainty at an aggregate level. In contrast, our work studies the uncertainty of demand quantity, confirmation date, and likelihood of confirmation, at both individual and aggregate levels.

Although the importance of advanced ATP systems has been widely acknowledged by practitioners, there are relatively few papers addressing quantitative models for ATP decision making. Most studies employ optimization techniques to study various aspects of ATP decisions, including due-date and lead time quotations (Taylor and Plenert 1999, Baker and Bertrand 1981, Hopp and Sturgis 2001), resource allocation (Ervolina and Dietrich 2001), production scheduling (Moses et al. 2004), requirements planning (Balakrishnan and Geunes 2000), and order promising (Kilger and Schneeweiss 2000, Robinson and Carlson 2007, among others). We refer the reader to a book chapter by Ball et al. (2004) for a detailed survey of this literature. Notably, most of the aforementioned ATP models treat future demand as deterministic and formulate optimization models accordingly. Only a few papers incorporate stochastic demand (de Kok 2000, Chen et al. 2008) in advanced ATP systems. De Kok (2000) assumes stochastic demand based on a relatively stable long-term forecast. In contrast, we capture the characteristics of more volatile demand signals, via a short-term pseudo-order forecast. Chen et al. (2008) consider a two-period, two-class, inventory model with a known demand distribution in each period. In contrast, our paper considers a multiperiod, multiclass, two-resource APT-A system with dynamic demand signals evolving in a Markov fashion.

Most rationing models are of the continuous-review type (see, e.g., Ha 1997, Zhao et al. 2005, Benjaafar and ElHafsi 2006, and references therein). Only a few papers deal with rationing decisions in the periodic-review setting. Among them, Frank et al. (2003) propose a simple ordering and rationing policy for an inventory system with two classes. Duran et al. (2007) prove that a modified order-up-to policy is optimal for an inventory problem with two demand classes and tactical inventory. Gupta and Wang (2007) consider an allocation problem where a capacitated manufacturer faces both contractual and transactional demands.

To further differentiate our model with the above literature, we also point out that ATP and inventory management serve different business functions in supply chain operations (Ball et al. 2004). First, ATP is operated in a *make-to-order* or an *assemble-to-order* environment, and requires using several system resources, including perishable resources, to *produce* and *assemble* an accepted order during a delivery lead time, whereas an inventory system is usually a *make-to-stock* system and uses nonperishable resources, i.e., inventories, to meet demand. Second, dictated by their respective business functionalities, the principal decisions in a multiclass ATP are *order acceptance* and *resource allocation*, to meet promised orders during a lead time; in contrast, the major decisions in a multiclass inventory system are *inventory replenishment and inventory rationing*, and demand, either lost or backlogged, typically does not have a delivery lead time. Third, because of the short decision horizon in ATP and the relatively long resource replenishment lead time, most manufacturing resources are planned in advance and considered fixed, whereas an inventory system has replenishment opportunities (e.g., Li and Gao 2008). Finally, because ATP typically operates in a business-to-business environment, it is common for sales personnel to closely track and classify the status of prospective orders, making it possible to use demand signals from business customers to assist ATP decision making; in contrast, because an inventory system often operates in a business-to-consumer environment and customers' demand signals cannot be systematically collected, its inventory replenishment decisions often depend on long-term demand forecasts.

3. Model Formulation

We consider an ATP assembly system with the following modeling constructs: (a) at the beginning of each period, the ATP-A system receives the confirmed orders (realized pseudo orders) of multiple classes, and receives class-dependent revenue for accepted orders; (b) the key attributes of a future pseudo order, including the order quantity and confirmation date, are forecasted dynamically over time; (c) the system makes the order acceptance decisions for the confirmed orders from different classes, subject to the constraint that all accepted orders must be fulfilled during a given delivery lead time. The rejected orders are lost. The order acceptance decisions involve the trade-offs between committing resources to current, lower-reward confirmed orders and reserving

resources for future, higher-valued pseudo orders; (d) the system utilizes two types of resources, *production capacity* and *component inventory*, to satisfy demand. Each resource type is shared across all the demand classes. The resource level of each type is planned in advance and considered exogenously given in each period; (e) a unit demand from any class requires one manufactured unit and one inventory unit. The manufactured unit takes one unit of capacity and needs a single period to produce. The manufactured unit and the inventory unit are then assembled into a finished product to meet demand. The assembly time is negligible.

In the remaining section, we first describe Markov chain models that capture dynamic forecast of pseudo orders at both the individual order and aggregate demand levels (§3.1). We then formulate the order acceptance problem in ATP-A as a discrete-time, finite horizon dynamic program (§3.2).

3.1. Pseudo Orders

The pronounced traits of pseudo orders include (1) *lumpiness*: a pseudo order has a nonnegligible probability of being cancelled, resulting in a null demand; (2) *nonstationarity*: demands in different periods are often not identically distributed; and (3) *volatility*: the attributes of a pseudo order change dynamically over time before the order either becomes a firm order or is cancelled.

Each pseudo order is characterized by several stochastic attributes. In this paper, we focus on two such attributes: *demand distribution* and *confirmation date*. We assume that the two attributes are independent of each other and are also independent of the attributes of other orders. This assumption allows us to develop a pseudo-order model for each individual class. Suppose the ATP system has a finite execution period T, which is traced backward with $t = T$ as the beginning of the planning horizon and $t = 1$ as the end. For expositional simplicity, we assume that all pseudo orders are available at time T, i.e., no new orders arrive from the outside world during the ATP execution period. Later, we discuss how to extend our model to incorporate new arrivals.

We first quantify the attributes of individual orders, and then aggregate individual orders by the confirmation date and obtain the aggregate demand in each period. Suppose we are at the beginning of period t, $1 \le t \le T$, and the orders with confirmation date t have just been confirmed. Let M_t be the index set of all future orders at the beginning of period t. A future order k, $k \in M_t$, is characterized by two attributes:

1. *Demand Distribution.* Demand of order k may follow one of several distributions. Let e_k be the indicator of distribution F_{e_k} and, hereafter, we call e_k a *distribution state* of order k. Let $\mathscr{E}$ be the set of distribution states of all orders. The distribution state of order k evolves according to a stationary Markov chain with the transition probability $q(e'_k \mid e_k)$, $e_k, e'_k \in \mathscr{E}$. For example, if $e_k = L$ and $e'_k = H$, then $q(H \mid L)$ quantifies the likelihood that the demand distribution forecast for order k will be revised from F_L in the current period to F_H in the next period.

2. *Confirmation Date.* The confirmation date of an order is the date at which the order becomes firm. The confirmation date of each order evolves according to a Markov chain with transition probability $h_t(s' \mid s)$, where $s, s' < t$ are the possible confirmation dates. For example, if $s = 3$ and $s' = 2$, then $h_t(2 \mid 3)$ represents the likelihood that an order, which in period t is forecasted to be released to ATP in period 3, will postpone its confirmation date to period 2 in the next period, $t - 1$.

Our Markov chain model aims at capturing the *volatility, nonstationarity*, and *dynamic information availability* of individual orders, the key characteristics of short-term forecasts. The transition probability matrices $\{q(\cdot \mid \cdot)\}$ and $\{h_t(\cdot \mid \cdot)\}$ are known in advance. In each period t, the state pairs of all orders, $\{(e_k, s_k): k \in M_t,\ e_k \in \mathscr{E}, s_k < t\}$, along with transition probability matrices $\{q(\cdot \mid \cdot)\}$ and $\{h_t(\cdot \mid \cdot)\}$, completely describe stochastic attributes of future orders in period t and the probability laws that govern their evolution. Note that although all the orders are governed by the same transition matrices, the demand process is in the transient phase and hence is nonstationary. To illustrate, suppose that only the distribution states of orders evolve over time and their confirmation dates are certain. Let the demand of each order, if realized, follow a Poisson distribution with mean λ, denoted by PP(λ). The order, however, is lumpy: Given it is currently "alive," the likelihood that it will be cancelled in the next period is π. Then the set of distribution states is $\mathscr{E} = \{0, 1\}$, where state 0 represents the cancellation state and state 1 corresponds to PP(λ). The transition matrix of distribution states is given by

$$\{q(\cdot \mid \cdot)\} = \begin{pmatrix} 1 & 0 \\ \pi & 1-\pi \end{pmatrix}.$$

Next, we use this example to illustrate how information availability can help the firm to manage volatility of pseudo orders. Observe that if order cancellation information is not available, then demand uncertainty comes from two sources: order cancellation and demand variability induced from PP(λ). Therefore, the demand distribution is the mixture of a point mass at zero with probability π and PP(λ) with probability $(1 - \pi)$. This distribution is known as the *zero-inflated Poisson* (ZIP) distribution (Thas and Rayner 2005), which assigns a greater probability to point 0 as compared to that in PP(λ). A ZIP distribution is more variable than the Poisson distribution with the same

mean, and can be used to depict demand lumpiness. When an order has a significant probability of being cancelled, as is often the case of a pseudo order, information updating through our Markov chain model can remove one source of demand uncertainty, and thus reduce the volatile nature of pseudo orders.

Next, we aggregate individual orders by their confirmation dates. For this purpose, let $\mathbf{e}_{t,s} = \{e_k\colon k \in M_t,\ s_k = s\}$ be the distribution states of the orders with confirmation date s, $1 \le s < t$. Then $E_t = (\mathbf{e}_{t,1}, \mathbf{e}_{t,2}, \ldots, \mathbf{e}_{t,t-1})$ represents the distribution states of all future pseudo orders in period t, aggregated by confirmation dates s, $s \le t$. The aggregate distribution state forecast in the next period, $E_{t-1} = (\mathbf{e}_{t-1,1}, \mathbf{e}_{t-1,2}, \ldots, \mathbf{e}_{t-1,t-1})$, is an update of E_t through the transition probability

$$\begin{aligned} &P(E_{t-1} \mid E_t) \\ &\quad = P((\mathbf{e}_{t-1,1}, \mathbf{e}_{t-1,2}, \cdots, \mathbf{e}_{t-1,t-1}) \mid (\mathbf{e}_{t,1}, \mathbf{e}_{t,2}, \ldots, \mathbf{e}_{t,t-1})) \\ &\quad = \prod_{k \in M_t} h_t(s_k' \mid s_k) \cdot q(e_k' \mid e_k), \end{aligned} \tag{1}$$

where $\mathbf{e}_{t-1,s} = \{e_k'\colon k \in M_{t-1},\ s_k' = s\}$, $s \le t-1$, represents the updated distribution states of the orders with confirmation date s, forecasted at the beginning of period $t-1$. The product form of the above transition probability uses the fact that aggregate demands across different confirmation dates are independent and each distribution state evolves independently. In particular, the transition probability is independent of the realized demand in period t.

Let us consider the distribution of the aggregate demand at a fixed confirmation date. Given distribution states $\mathbf{e}_{t,s} = \{e_k\colon k \in M_t,\ s_k = s\}$ in period t, denote $Y^k(e_k)$ as the random demand of order k in state e_k. Then the aggregate demand with confirmation date s in period t is given by

$$X_{t,s}(\mathbf{e}_{t,s}) = \sum_{k\colon s_k = s} Y^k(e_k), \quad s = 1, \ldots, t-1.$$

Therefore, the distribution of $X_{t,s}(\mathbf{e}_{t,s})$, denoted by $F_{\mathbf{e}_{t,s}}$, is the convolution of distributions F_{e_k} with $s_k = s$. Clearly, $F_{\mathbf{e}_{t,s}}$ is uniquely determined by states $\mathbf{e}_{t,s}$. The pseudo-order confirmation and information update in each period occur in the following sequence: At the beginning of period $t-1$, demand $X_{t-1,t-1}(\mathbf{e}_{t-1,t-1}) = \sum_{k\colon s_k' = t-1} Y^k(e_k')$ is realized according to distribution $F_{\mathbf{e}_{t-1,t-1}}$; the order index set in period t, M_t, is updated to $M_{t-1} = M_t \setminus \{k \in M_t\colon s_k' = t-1\}$; and the distribution state E_{t-1} is updated by the transition probability (1).

It is often the case that the distribution states of aggregate demands, $(\mathbf{e}_{t,1}, \ldots, \mathbf{e}_{t,t-1})$, have simpler representations. This can further simplify the computation of $F_{\mathbf{e}_{t,s}}$ and the transition matrix of the distribution states. To illustrate, suppose each order has two states, 0 and 1, with state 1 corresponding to PP(λ) and state 0 order cancellation, which occurs with probability π. Then the aggregate demand with confirmation date s can be represented by the total number of uncancelled orders with confirmation date s, that is, state $(\mathbf{e}_{t,1}, \ldots, \mathbf{e}_{t,t-1})$ can be simplified to state $(n_{t,1}, \ldots, n_{t,t-1})$, where $n_{t,s} = |\mathbf{e}_{t,s}|$ is the cardinality of $\mathbf{e}_{t,s}$, $s \le t-1$. The aggregate demand in period s, forecasted in period t, follows PP($n_{t,s}\lambda$). Denote $n(0 \mid s)$ as the number of cancelled orders with confirmation date s and $n(s' \mid s)$ as the number of uncancelled orders that switch their confirmation date from s to s'', $1 \le s, s' \le t-1$. Then we can write

$$\begin{aligned} n_{t,s} &= \sum_{s'=0}^{t-1} n(s' \mid s), \quad s = 1, \ldots, t-1, \quad \text{and} \\ n_{t-1,s'} &= \sum_{s=1}^{t-1} n(s' \mid s), \quad s' = 1, \ldots, t-1. \end{aligned} \tag{2}$$

Denote $\mathbf{n}_t = \{n(s' \mid s)\colon 0 \le s' \le t-1,\ 1 \le s \le t-1\}$. Note that each $\mathbf{n}_t$ that satisfies (2) specifies a transition from $(n_{t,1}, n_{t,2}, \ldots, n_{t,t-1})$ to $(n_{t-1,1}, n_{t-1,2}, \ldots, n_{t-1,t-1})$. Then (1) can be written as

$$\begin{aligned} &P((n_{t-1,1}, n_{t-1,2}, \ldots, n_{t-1,t-1}) \mid (n_{t,1}, n_{t,2}, \ldots, n_{t,t-1})) \\ &\quad = \sum_{\mathbf{n}_t} \left[\prod_{s=1}^{t-1} \frac{n_{t,s}!}{n(0 \mid s)!(n_{t,s} - n(0 \mid s))!} \pi^{n(0 \mid s)} (1-\pi)^{n_{t,s} - n(0 \mid s)} \right. \\ &\qquad \left. \times \prod_{s=1}^{t-1} \left(\frac{(n_{t,s} - n(0 \mid s))!}{n(1 \mid s)! \cdots n(t-1 \mid s)!} \prod_{s'=1}^{t-1} (h_t(s' \mid s))^{n(s' \mid s)} \right) \right]. \end{aligned}$$

Here, the summation is over all possible $\mathbf{n}_t$ satisfying (2). The first product form corresponds to the probability of cancellations across different confirmation dates, and the second product form corresponds to the probability of confirmation date changes of uncancelled orders. Another example of our demand aggregation and updating scheme is given in §EC.2 in the electronic companion.

We conclude this subsection with a remark on how to incorporate other features of pseudo orders into our Markov chain model. One possible relaxation is to allow new orders to arrive during the ATP execution period. Let an arriving order carry with it a distribution state and a confirmation date, which can be observed upon its arrival. An attribute of a new order evolves independently according to a Markov chain, similar to that of an existing order. Our demand aggregation scheme in each period will be based on the combined distribution states of new and existing orders. Another generalization is to allow the transition matrix for a stochastic attribute to be time and order dependent. This approach may be desirable if the number of pseudo orders is small but the demand

quantity is large, and each pseudo order can be associated with a major client, whose stochastic behavior can be extracted from historical data. Then our Markov chain model still applies.

3.2. Dynamic Programming Formulation

Consider an ATP-A system that produces several products during the execution period T, with time traced backward. This ATP-A adopts the make-to-order (MTO) strategy, that is, it will not produce in advance in anticipation of future orders. Let $\mathcal{J} = \{j: j = 1, \ldots, J\}$ be the index set of product (or customer class) types. In period t, S_t units of component inventory and K_t units of production capacity become available. Resources $S = \{S_t\}$ and $K = \{K_t\}$ are planned in advance of the ATP execution period and are treated as exogenous input parameters. A unit demand, independent of the product type, is assembled from two major components: one component needs one unit of production capacity (called *capacity* hereafter) and a single period to produce; the manufactured component then needs to be assembled with one unit of component inventory (called *inventory* hereafter) to become a finished product. We assume that the assembly time is negligible. This kind of production-assembly system is common in practice, in which the final product is assembled from the major components produced in-house or supplied by other venders. A confirmed order, which may consist of several units of a given product type, can be partially accepted. Denote the profit margin of product j by r^j, $j \in \mathcal{J}$. We rank product types in an increasing order of the profit margins so that $r^1 > r^2 > \cdots > r^J$. The leftover inventory not committed to the accepted demand incurs an increasing convex holding cost $h(\cdot)$ in each period. To smooth production and improve capacity utilization, we charge a unit penalty cost p for capacity idleness. Note that inventory is a *nonperishable* resource that can be carried forward to the next period, whereas production capacity is a *perishable* resource that can only be utilized in the current period.

Denote I_t and Q_t as the *net inventory* (i.e., inventory on hand minus *inventory backlogs*, where inventory backlogs are the accepted demands whose inventory requirement has not been met) and the *net production capacity* (i.e., zero minus *capacity backlogs*, where capacity backlogs are the previously accepted demands whose capacity requirement has not been met), respectively, at the beginning of period t, before the planned resources in period t, i.e., S_t and K_t, are received. If $I_t < 0$, then $-I_t$ is the inventory backlog of the previously accepted orders; if $I_t \geq 0$, then all the inventory requirements for these orders have been satisfied and I_t can be used to meet the new commitment. On the other hand, because production capacity is perishable and cannot be carried forward, the net capacity Q_t cannot be positive. If $Q_t < 0$, then $-Q_t$ is the capacity needed to produce the manufactured units for the previously accepted orders; if $Q_t = 0$, then all these orders have their manufactured units produced already by time t.

In each period t, we keep track of information of two types of orders, namely, newly confirmed orders and future pseudo orders with different confirmation dates. Because partial acceptance of an order is allowed, we can aggregate each class of orders according to the confirmation date s and product type $j \in \mathcal{J}$, as follows.

Pseudo orders (future demand). Our pseudo-order model for the single-product case is described in §3.1. For the multiproduct case, we append superscript j to each relevant notation to indicate its dependence on product j. For example, $E_t^j = (\mathbf{e}_{t,1}^j, \mathbf{e}_{t,2}^j, \ldots, \mathbf{e}_{t,t-1}^j)$ represents the distribution states of pseudo orders for product (class) j forecasted in period t, which is updated via transition probability $P(E_{t-1}^j \mid E_t^j)$ given in (1). With a slight abuse of notation, we redefine $E_t = \{E_t^j = (\mathbf{e}_{t,1}^j, \mathbf{e}_{t,2}^j, \ldots, \mathbf{e}_{t,t-1}^j): j \in \mathcal{J}\}$ as the collection of the distribution states of all classes in period t. Because the attributes of different classes are independent, the transition probability of the distribution states across different classes has the product form

$$P(E_{t-1} \mid E_t) = \prod_{j \in \mathcal{J}} P(E_{t-1}^j \mid E_t^j). \tag{3}$$

Confirmed orders (current demand). These are the realizations of pseudo orders with confirmation date t. Let N_t^j be a realization of X_t^j, the class j demand with the confirmation date t, and $N_t = (N_t^1, \ldots, N_t^J)$. Because of the independence of $X_t = (X_t^1, \ldots, X_t^J)$, the joint probability mass function of X_t is given by

$$p(N_t \mid E_t) \equiv P(X_t = N_t \mid E_t) = \prod_{j \in \mathcal{J}} P(X_t^j = N_t^j). \tag{4}$$

Given N_t, the firm can accept, partially accept, or reject the entire demand of a given class. We assume that the accepted demand must be delivered within L periods of its acceptance and call L the delivery lead time. Because the BOM and the lead time for different classes are the same, the accepted demand will be produced on a first-in, first-out (FIFO) basis (although the BOMs are identical, the production capacity is sufficiently flexible to customize products for different demand classes). Our *hard* lead time delivery commitment assumption is consistent with the business function of ATP systems, whose purpose is to support order promising and fulfillment decisions by matching customer orders with available resources and deliver accepted orders before the promised delivery date.

Collectively, the state of the system in period t can be represented by (I_t, Q_t, N_t, E_t), which represent, respectively, the net inventory, the net production capacity, the newly confirmed orders and the short-term forecast of future demands in period t. The sequence of events in each period unfolds as follows. At the beginning of period t, the pseudo orders with the confirmation date t are realized as N_t and the future demand forecast is updated to E_t. The firm observes net inventory I_t and net capacity Q_t and also receives the planned resources S_t and K_t. The firm then makes the order acceptance decision $\mathbf{x}_t = \{x_t^j : j \in \mathcal{J}\}$, where $x_t^j \le N_t^j$ is the demand accepted for class j in period t, subject to the constraints that sufficient inventory and capacity resources are available to meet the overall commitment $\mathbf{x}_t$ within L periods. After the acceptance decision, the firm produces and assembles as much as possible of the committed demand, on the FIFO basis, using available resources in period t. If available resources are insufficient to satisfy all the accepted demand, unmet capacity or inventory requirements for the accepted demand are backlogged and carried forward to the next period.

Denote $|\mathbf{x}_t| = \sum_{j\in\mathcal{J}} x_t^j$ as the total accepted demand in period t. For any scalars a and b, let $a^+ = \max\{a, 0\}$, $a \wedge b = \min\{a, b\}$. For any vector B, denote $[B]_s^t = \sum_{\tau=s}^t B_\tau$, $s < t$, with $[B]_s^t = 0$ if $s < 1$. For example, $[S]_{t-L}^t = \sum_{s=t-L}^t S_s$ represents the cumulative inventory replenishment for the next $L+1$ periods. Let $V_t(I_t, Q_t, N_t, E_t)$ be the maximum expected profit from period t onward given the current state (I_t, Q_t, N_t, E_t). Then $V_t(I_t, Q_t, N_t, E_t)$ satisfies the optimality equations

$$\begin{aligned} &V_t(I_t, Q_t, N_t, E_t) \\ &= \max_{\mathbf{x}_t\in\mathcal{A}_t}\Bigg\{\sum_{j=1}^J r^j x_t^j - h(I_t + S_t - |\mathbf{x}_t|) - p\times(Q_t + K_t - |\mathbf{x}_t|)^+ \\ &\qquad + \sum_{E_{t-1}}\sum_{N_{t-1}} p(N_{t-1}\mid E_t)P(E_{t-1}\mid E_t) \\ &\qquad \cdot V_{t-1}(I_{t-1}, Q_{t-1}, N_{t-1}, E_{t-1})\Bigg\}, \end{aligned} \tag{5}$$

where $V_0(I_0, Q_0, N_0, E_0) \equiv 0$ is the boundary condition, and $p(N_{t-1} \mid E_t)$ and $P(E_{t-1} \mid E_t)$ satisfy (4) and (3), respectively. The action space $\mathcal{A}_t$ is defined by

$$\begin{aligned} \mathcal{A}_t = \big\{&\mathbf{x}_t : 0 \le \mathbf{x}_t \le N_t, \\ &|\mathbf{x}_t| \le I_t + [S]_{t-L}^t, |\mathbf{x}_t| \le Q_t + [K]_{t-L}^t\big\}, \end{aligned} \tag{6}$$

and the net inventory and the net capacity in the next period are updated, respectively, as

$$I_{t-1} = I_t + S_t - |\mathbf{x}_t|, \tag{7}$$

$$Q_{t-1} = [Q_t + K_t - |\mathbf{x}_t|] \wedge 0. \tag{8}$$

We now provide a brief interpretation of the above dynamic program formulation. The first term in the objective function (5) is the profit function, the second and third terms are the inventory holding and capacity idleness penalty costs, respectively, and the last term is the expected future profit. Action space (6) states that the newly accepted demand for each class cannot be greater than the newly realized demand for that class, and the system must have a sufficient resource of each type, after first satisfying the resource requirement for the backlogged demand, if any, to deliver the new commitment $|\mathbf{x}_t|$ within L periods. Equation (7) updates the net inventory by removing the inventory committed to the newly accepted demand $|\mathbf{x}_t|$ from $I_t + S_t$. Finally, (8) means that the system in period t will use the available capacity K_t to meet its production commitment $-Q_t + |\mathbf{x}_t|$. The leftover capacity, if any, will expire at the end of the period.

4. Structural Properties

In this section, we first establish the structural results of the optimal policy for the ATP-A system in §4.1. We then discuss two special cases of the ATP-A system in §4.2.

4.1. Structural Properties of the Optimal Order Acceptance Policy

We shall show that in our ATP-A system, the optimal acceptance policy has three key drivers: *class prioritization*, *capacity-inventory-demand matching*, and *resource-imbalance-based rationing*. These properties will be discussed in the following three subsections, respectively.

4.1.1. Class Prioritization. The following lemma establishes *class prioritization* as the first driver of the optimal policy. The proof uses the standard interchange argument and we omit details.

Lemma 1 (The Lowest-Indexed Class First (LICF) Rule). *In any period t, the optimal policy always accepts the demand from a lower-indexed class before the demand from a higher-indexed class. That is, if $\mathbf{x}_t^* = (x_t^{1*}, \ldots, x_t^{J*})$ are the optimal acceptance quantities, then $0 \le x_t^{j*} < N_t^j$ implies $x_t^{i*} = 0$ for all $i > j$.*

Under the LICF rule, we can replace the *accepted demand vector* $\mathbf{x}_t$ by the *totally accepted demand* $x = |\mathbf{x}_t|$. Conversely, given x and N_t, the accepted class j demand satisfies $x_t^j = (x - [N]_1^{j-1})^+ \wedge N_t^j$ for $j \in \mathcal{J}$. The profit $r(N_t, x)$ in period t, given the realized demand N_t and the totally accepted demand x, can be written as

$$r(N_t, x) = \sum_{j\in\mathcal{J}} r^j \cdot ((x - [N]_1^{j-1})^+ \wedge N_t^j). \tag{9}$$

Because r^j is decreasing in j, the profit function $r(N_t, x)$ is a *piecewise linear, increasing concave* function of (N_t, x). To ease notation, define the profit-to-go function from period t to the end by

$$V_{t-1}(I_{t-1}, Q_{t-1} \mid E_t) = \sum_{N_{t-1}} \sum_{E_{t-1}} p(N_{t-1} \mid E_t) P(E_{t-1} \mid E_t) \cdot V_{t-1}(I_{t-1}, Q_{t-1}, N_{t-1}, E_{t-1}). \quad (10)$$

We may understand $V_{t-1}(I_{t-1}, Q_{t-1} \mid E_t)$ as the future profit before updating the demand forecast E_{t-1} and observing the pseudo-order demand in the next period. Then, under the LICF rule, we can rewrite the optimality equations as

$$\begin{aligned} &V_t(I_t, Q_t, N_t, E_t) \\ &\quad = \max_{x \in \mathscr{A}_t} \{ r(N_t, x) - h(I_t + S_t - x) - p \times (Q_t + K_t - x)^+ \\ &\qquad + V_{t-1}(I_t + S_t - x, (Q_t + K_t - x) \wedge 0 \mid E_t) \} \quad (11) \\ &\quad = \max_{x \in \mathscr{A}_t} \{ G_t(I_t, Q_t, N_t, E_t, x) \}, \quad (12) \end{aligned}$$

where

$$\begin{aligned} &G_t(I_t, Q_t, N_t, E_t, x) \\ &\quad = r(N_t, x) - h(I_t + S_t - x) - p \times (Q_t + K_t - x)^+ \\ &\qquad + V_{t-1}(I_t + S_t - x, (Q_t + K_t - x) \wedge 0 \mid E_t) \quad (13) \end{aligned}$$

is the maximum expected profit from period t to the end, given state (I_t, Q_t, N_t, E_t) and the total acceptance x, with the boundary condition $V_0(I_t, Q_0, N_0, E_0) \equiv 0$. With a slight abuse of notation, we redefine the action set $\mathscr{A}_t$ in state (I_t, Q_t, N_t, E_t) in (6) as

$$\mathscr{A}_t = \{ x: 0 \le x \le [N]_1^J,\ x \le I_t + [S]_{t-L}^t,\ x \le Q_t + [K]_{t-L}^t \}. \quad (14)$$

Next, we discuss the solution procedure for the dynamic program (12)–(14). We first define the difference function of V_{t-1} with respect to capacity as

$$\begin{aligned} &\Delta_Q V_{t-1}(I, Q \wedge 0 \mid E_t) \\ &\quad = V_{t-1}(I, Q \wedge 0 \mid E_t) - V_{t-1}(I, (Q-1) \wedge 0 \mid E_t). \end{aligned}$$

Here, Q can be understood as the ending capacity in the previous period before the unused capacity, if any, expires, and $Q \wedge 0$ (≤ 0) is the beginning capacity in the current period after the unused capacity in the previous period expires but before the planned capacity is installed. Note that $\Delta_Q V_{t-1}(I, Q \wedge 0 \mid E_t)$ represents the marginal value of the extra capacity $Q \wedge 0 - (Q-1) \wedge 0$, if any, at the resource level $(I, Q \wedge 0)$ at the beginning of period $t-1$. Similarly, the difference function of V_{t-1} with respect to inventory at the capacity level $Q \wedge 0$ is given by

$$\begin{aligned} &\Delta_I V_{t-1}(I, Q \wedge 0 \mid E_t) \\ &\quad = V_{t-1}(I, Q \wedge 0 \mid E_t) - V_{t-1}(I-1, Q \wedge 0 \mid E_t), \quad (15) \end{aligned}$$

which is the marginal value of a unit of inventory at the resource level $(I, Q \wedge 0)$. Because the maximum benefit of a *unit pair* of resources is r^1 (when they are used to meet the class 1 demand), we have, for any resource levels $(I, Q \wedge 0)$,

$$\begin{aligned} r^1 &\ge V_{t-1}(I, Q \wedge 0 \mid E_t) - V_{t-1}(I-1, (Q-1) \wedge 0 \mid E_t) \\ &= \Delta_I V_{t-1}(I, Q \wedge 0 \mid E_t) + \Delta_Q V_{t-1}(I-1, Q \wedge 0 \mid E_t). \quad (16) \end{aligned}$$

Because the maximum cost of a unit of capacity is p (when it is left idle in a period), we have

$$\Delta_Q V_{t-1}(I-1, Q \wedge 0 \mid E_t) \ge -p. \quad (17)$$

The following theorem establishes the concavity properties of the objective function G_t and the value function V_t.

THEOREM 1. (i) *$G_t(I_t, Q_t, N_t, E_t, x)$ is a concave function of (I_t, Q_t, x) for fixed (N_t, E_t).*

(ii) *$V_t(I_t, Q_t, N_t, E_t)$ is a concave function of (I_t, Q_t) for fixed (N_t, E_t).*

(iii) *$V_t(I_t, Q_t, N_t, E_t)$ is a nonincreasing concave function of N_t for fixed (I_t, Q_t, E_t).*

Theorem 1 confirms the common understanding that a firm's profit depends on the proper balance between supply and demand. In the context of ATP-A, it means that, first, during the planning stage, the firm needs to effectively allocate resources (S, K) over the planning horizon, using pseudo-order information available at time T and the short-term forecast model developed in §3.1; second, in the execution stage, the firm needs to coordinate demand and resource management activities effectively to ensure both high service level and high resource utilization.

The joint concavity of $G_t(I_t, Q_t, N_t, E_t, x)$ in (I_t, Q_t, x) implies the concavity of G_t in x. Let $\hat{x}_t(I_t, Q_t, N_t, E_t)$ be the largest nonnegative maximizer of $G_t(I_t, Q_t, N_t, E_t, x)$, without imposing other constraints in action set $\mathscr{A}_t$ defined in (14):

$$\begin{aligned} &\hat{x}_t(I_t, Q_t, N_t, E_t) \\ &= \max\{ x \ge 0:\ G_t(I_t, Q_t, N_t, E_t, x) \\ &\qquad - G_t(I_t, Q_t, N_t, E_t, x-1) \ge 0 \} \\ &= \max\{ x \ge 0:\ \Delta_x r(N_t, x) \\ &\qquad \ge -\Delta h(I_t + S_t - x + 1) - \Delta p(Q_t + K_t - x + 1)^+ \\ &\qquad + \Delta_I V_{t-1}(I_t + S_t - x + 1, (Q_t + K_t - x + 1) \wedge 0 \mid E_t) \\ &\qquad + \Delta_Q V_{t-1}(I_t + S_t - x, (Q_t + K_t - x + 1) \wedge 0 \mid E_t) \}, \end{aligned} \quad (18)$$

where $\hat{x}_t(I_t, Q_t, N_t, E_t) \equiv 0$ if the conditions in (18) do not hold for any $x > 0$. Here, $\Delta_x r(N_t, x) = r(N_t, x) - r(N_t, x-1)$ is the marginal revenue, and

$\Delta h(I_t + S_t - x + 1)$ and $\Delta p(Q_t + K_t - x + 1)^+$ are the marginal inventory holding and capacity idleness costs, respectively. We call $\hat{x}_t(I_t, Q_t, N_t, E_t)$ the state-dependent, *optimal base demand acceptance level*. After imposing constraints in $\mathcal{A}_t$, the optimal solution of the dynamic program (12)–(14) is given by

$$x_t^*(I_t, Q_t, N_t, E_t) = \min\{[N_t]_1^J, I_t + [S]_{t-L}^t, Q_t + [K]_{t-L}^t, \hat{x}_t(I_t, Q_t, N_t, E_t)\}. \quad (19)$$

Equation (19) states that the optimal demand acceptance policy is of an *accept-up-to* type: we accept demand, in decreasing order of class indices, until either the total demand $[N_t]_1^J$ is accepted, or one of the lead time resources is exhausted, or the optimal base demand acceptance level $\hat{x}_t(I_t, Q_t, N_t, E_t)$ is reached, whichever occurs first.

4.1.2. Capacity-Inventory-Demand Matching. In the next theorem, we summarize the *modularity* properties of the value functions G_t and V_t. These properties reveal the joint effect of the resource levels (I_t, Q_t) on the acceptance decision x. We call a function $f(x, y)$ on $\mathbb{Z}^2$ *supermodular* in (x, y) if $\Delta_x f(x, y) \le \Delta_x f(x, y+1)$, and *submodular* in (x, y) if the inequality is reversed. Supermodularity (submodularity) of $f(x, y)$ means that the marginal value of one variable increases (decreases) as another variable increases. We call a function $f(x, y)$ on $\mathbb{Z}^2$ *diagonally dominant* in x (or y) if $\Delta_x f(x, y) \le \Delta_x f(x+1, y+1)$ (or $\Delta_y f(x, y) \le \Delta_y f(x+1, y+1)$), and *diagonally dominated* in x (or y) if the inequality is reversed. The diagonally dominant (dominated) functions are also known in the literature as *generalized supermodular (submodular)* functions, respectively (Glasserman and Yao 1994). Generalized supermodularity (submoduarity) of $f(x, y)$ means that the marginal value of one variable increases (decreases) as *both* variables increase. Modular and generalized modular functions have been exploited in the literature to study the structure of the optimal control policy (see, e.g., Topkis 1998, Yang and Qin 2007, Zhao et al. 2008).

Theorem 2. (i) $G_t(I_t, Q_t, N_t, E_t, x)$ *is a supermodular function of* (I_t, x) *for fixed* (Q_t, N_t, E_t) *and a supermodular function of* (Q_t, x) *for fixed* (I_t, N_t, E_t).

(ii) $\hat{x}_t(I_t, Q_t, N_t, E_t)$, *defined in* (18), *and* $x_t^*(I_t, Q_t, N_t, E_t)$, *defined in* (19), *are increasing functions of net inventory* I_t *and net capacity* Q_t *for fixed* (N_t, E_t). *In addition,*

$$\hat{x}_t(I_t, Q_t, N_t, E_t) \le \hat{x}_t(I_t+1, Q_t+1, N_t, E_t) \le \hat{x}_t(I_t, Q_t, N_t, E_t) + 1, \quad (20)$$

$$x_t^*(I_t, Q_t, N_t, E_t) \le x_t^*(I_t+1, Q_t+1, N_t, E_t) \le x_t^*(I_t, Q_t, N_t, E_t) + 1. \quad (21)$$

(iii) $V_t(I_t, Q_t, N_t, E_t)$ *is diagonally dominated in* I_t *and* Q_t, *i.e., for* $Q < 0$,

$$\Delta_{I_t} V_t(I_t+1, Q_t+1, N_t, E_t) \le \Delta_{I_t} V_t(I_t, Q_t, N_t, E_t) \quad \text{for fixed } (Q_t, N_t, E_t),$$

$$\Delta_{Q_t} V_t(I_t+1, Q_t+1, N_t, E_t) \le \Delta_{Q_t} V_t(I_t, Q_t, N_t, E_t) \quad \text{for fixed } (I_t, N_t, E_t).$$

Theorem 2(i) and (ii) state that $\hat{x}_t(I_t, Q_t, N_t, E_t)$ and $x_t^*(I_t, Q_t, N_t, E_t)$, the optimal base demand acceptance levels before and after imposing the resource availability and demand constraints, respectively, increase when either resource level increases. Therefore, they conform with the common belief that more demand should be accepted when system resources become abundant. Theorem 2(ii) further states that when the resource levels increase by a *unit pair*, the optimal demand acceptance level increases by at most one unit. Theorem 2(iii) means that the marginal value of the I_tth unit of inventory at the capacity level Q_t is higher than the marginal value of the (I_t+1)th unit of inventory at the net capacity level (Q_t+1); similarly, the marginal value of the Q_tth unit of capacity at the inventory level I_t is higher than the marginal value of the (Q_t+1)th unit of capacity at the inventory level (I_t+1). In other words, the marginal value of each resource diminishes as both resource levels increase.

Remark 1. Notably, the value function $V_t(I_t, Q_t, N_t, E_t)$ is *neither* supermodular *nor* submodular in (I_t, Q_t). In other words, the marginal value of one resource need not be monotone as another resource level increases. For example, the marginal value of the Q_tth unit may be low when inventory is scarce (i.e., when I_t is substantially lower than Q_t), because this capacity unit cannot be utilized to meet demand due to lack of inventory. On the other hand, the marginal value of the Q_tth unit may also be low when inventory is abundant (i.e., when I_t is substantially higher than Q_t), because this unit is not needed due to lack of demand. However, the Q_tth unit of capacity will have a high marginal value when it can be properly paired with the I_tth unit of inventory to meet demand. In other words, to simultaneously achieve high resource utilization and high profitability, the inventory, capacity, and demand trio must be aligned in close proximity. We shall refer to this property as *capacity-inventory-demand (CID) matching.*

4.1.3. Resource-Imbalance-Based Rationing. Although the concavity and modularity properties of the value functions G_t and V_t help identify the structure of the optimal policy, it reveals little insight on how the optimal solution $\hat{x}_t(I_t, Q_t, N_t, E_t)$

to $\max_{x\ge 0}\{G_t(I_t, Q_t, N_t, E_t, x)\}$ and the optimal solution $x_t^*(I_t, Q_t, N_t, E_t)$ to $\max_{x\in\mathscr{A}_t}\{G_t(I_t, Q_t, N_t, E_t, x)\}$ depend on each component of state (I_t, Q_t, N_t, E_t). In the remainder of this section, we aim at characterizing the properties of the optimal policy parameters. These properties not only ease computational complexity in determining the optimal policy parameters in applications, but also identify *resource-imbalance-based (RIB) rationing* as the third principle in effective management of ATP-A systems.

For this purpose, let $D_t = Q_t - I_t$ be the *net resource imbalance level* between the net capacity and net inventory levels (i.e., the net capacity overage to inventory) in period t, before receiving resource replenishment. We shall decompose the dynamic program (12)–(14) into a sequence of subproblems, each of which determines the optimal *capacity rationing* level for a given class that depends only on the net resource imbalance level D_t and the forecast E_t. Because under the LICF rule with the total acceptance level x, we accept demand in an increasing order of class indices until x units of demand are accepted, we can define

$$\begin{aligned} &G_t(I_t, Q_t, N_t, E_t, x)\\ &\quad = G_t^j(I_t, Q_t, N_t, E_t, x)\\ &\quad = \sum_{i=1}^{j-1} r^i N_t^i + r^j\bigl(x - [N_t]_1^{j-1}\bigr)\\ &\qquad - h(I_t + S_t - x) - p\times(Q_t + K_t - x)^+\\ &\qquad + V_{t-1}\bigl(I_t + S_t - x, (Q_t + K_t - x)\wedge 0 \mid E_t\bigr)\\ &\qquad\qquad \text{if } [N_t]_1^{j-1} \le x \le [N_t]_1^j,\ j\in\mathscr{J}, \quad (22)\end{aligned}$$

where $[N_t]_1^0 \equiv 0$. Because $G_t(I_t, Q_t, N_t, E_t, x)$ is a concave function of x, each of its constituting pieces $G_t^j(I_t, Q_t, N_t, E_t, x)$ is a piecewise concave function of x. As such, to determine $x_t^*(I_t, Q_t, N_t, E_t)\in\mathscr{A}_t$ that maximizes $G_t(I_t, Q_t, N_t, E_t, x)$, it is equivalent to solving the following subproblems $j\in\mathscr{J}$:

$$\max_{x\in\mathscr{A}_t^j}\{G_t^j(I_t, Q_t, N_t, E_t, x)\}, \quad (23)$$

with the action set for the subproblem j defined as

$$\begin{aligned}\mathscr{A}_t^j = \{x:\ &[N_t]_1^{j-1} \le x \le [N_t]_1^j,\ x \le I_t + [S]_{t-L}^t,\\ &x \le Q_t + [K]_{t-L}^t\}, \quad j\in\mathscr{J}. \quad (24)\end{aligned}$$

We can solve the above subproblems sequentially starting from $j = 1$, and, once the optimal solution for the jth problem is strictly less than $[N_t]_1^j$ for some j, we stop and declare that this optimal solution is the global optimal solution $x^*(I_t, Q_t, N_t, E_t)$ to the dynamic program (12)–(14).

Toward this end, we first relax $x\in\mathscr{A}_t^j$ to allow $x\in\mathbb{Z}$. We define the difference function of G_t^j in (22) with respect to x by

$$\begin{aligned}&\Delta_x G_t^j(I_t, Q_t, N_t, E_t, x)\\ &\quad = r^j - \Delta_x h(I_t + S_t - x) - \Delta_x p\cdot(Q_t + K_t - x)^+\\ &\qquad + \Delta_x V_{t-1}(I_t + S_t - x, (Q_t + K_t - x)\wedge 0 \mid E_t). \quad (25)\end{aligned}$$

Note that $\Delta_x G_t^j$ is independent of N_t after relaxing $x\in\mathscr{A}_t^j$. The *unconstrained* optimizer of G_t^j, without imposing $x\in\mathscr{A}_t^j$, is given by

$$\begin{aligned}&\hat{x}_t^j(I_t, Q_t, E_t)\\ &= \max\{x:\ \Delta_x G_t^j(I_t, Q_t, N_t, E_t, x)\ge 0\}\\ &= \max\bigl\{x:\ r^j \ge \Delta_x h(I_t + S_t - x) + p\Delta_x(Q_t + K_t - x)^+\\ &\qquad - \Delta_x V_{t-1}(I_t + S_t - x, (Q_t + K_t - x)\wedge 0 | E_t)\bigr\}, \quad (26)\end{aligned}$$

which is the optimal *base demand acceptance level* of the first j classes, $j\in\mathscr{J}$. Now we perform the following two linear transformations to (26): $y = I_t + S_t - x$ and $D_t = Q_t - I_t$. Here, y is the *net ending inventory* after accepting the first j class demand and D_t is the *net resource imbalance level*. Using these two linear transformations, we can express the ending capacity level as

$$Q_t + K_t - x = Q_t + K_t - I_t - S_t + y = y + D_t + K_t - S_t.$$

We then obtain from (26) that the optimal *ending inventory rationing level* as

$$\begin{aligned}&\hat{y}_t^j(D_t, E_t)\\ &\quad = (I_t + S_t) - \hat{x}_t^j(I_t, Q_t, E_t)\\ &\quad = \min\bigl\{y:\ r^j \ge -\Delta_y h(y) - p\Delta_y(y + D_t + K_t - S_t)^+\\ &\qquad + \Delta_y V_{t-1}\bigl(y, (y + D_t + K_t - S_t)\wedge 0 \mid E_t\bigr)\bigr\}, \quad (27)\end{aligned}$$

where $\hat{y}_t^j(D_t, E_t) = -\infty$ if the constraint in (27) holds for all y. The optimal ending *capacity* level for class j is given by $\hat{y}_t^j(D_t, E_t) + D_t + S_t - K_t$, $j\in\mathscr{J}$. Here, the pair $(\hat{y}_t^j(D_t, E_t), \hat{y}_t^j(D_t, E_t) + D_t + K_t - S_t)$ can be understood as the desirable resource rationing levels reserved for the higher valued, future demand than the class j demand. In economic terms, at the rationing levels $(\hat{y}_t^j(D_t, E_t), \hat{y}_t^j(D_t, E_t) + D_t + K_t - S_t)$, the marginal value of a *unit pair* of resources equals the marginal profit of a class j demand. Because the optimal inventory rationing level $\hat{y}_t^j(D_t, E_t)$ is independent of N_t and depends on (I_t, Q_t) only via their difference D_t, we have established *RIB rationing* as a key driver of the optimal acceptance decision in the ATP-A system.

Remark 2. RIB rationing studied in this paper is significantly different from the rationing control

policies reported in the literature. First, the literature mostly considers rationing control of a *single, nonperishable* resource type, whereas we consider rationing control for both the *perishable* and *nonperishable* resource types. Our RIB rationing captures not only the notion of *resource reservation*, but also the notion of *resource balancing*. Additionally, resource rationing in the literature, with the exception of Duran et al. (2007, 2008), generally means to reserve currently available, scarce resources for future use. In our case, the optimal inventory *rationing* level $\hat{y}_t^j(D_t, E_t)$ can be either *positive* or *negative*. If $\hat{y}_t^j(D_t, E_t) \geq 0$, the optimal base demand acceptance level for the first j classes is less than the available inventory, so $\hat{y}_t^j(D_t, E_t)$ units of inventory should be *set aside* for future, higher-reward demand than class j. If $\hat{y}_t^j(D_t, E_t) < 0$, the optimal base demand acceptance level for the first j classes is more than the currently available inventory, so $\hat{x}_t^j(I_t, Q_t) - I_t - S_t = -\hat{y}_t^j(D_t, E_t)$ units of the future inventory should be *booked* to meet the inventory requirement of the first j class demand. As such, RIB rationing can reserve a resource for both the current and future high-valued orders.

Because $r^1 \geq r^2 \geq \cdots \geq r^J$ and $r^1 \geq \Delta_y V_{t-1}(y, (y + D_t + K_t - S_t) \wedge 0 \mid E_t)$ for any (D_t, E_t, y), class j's ending inventory rationing level is increasing in j, i.e.,

$$\hat{y}_t^1(D_t, E_t) \leq \hat{y}_t^2(D_t, E_t) \leq \cdots \leq \hat{y}_t^J(D_t, E_t). \quad (28)$$

By restricting x to $\mathcal{A}_t^j$ defined in (24), and using the identity $\hat{x}_t^j(I_t, Q_t, E_t) = (I_t + S_t) - \hat{y}_t^j(D_t, E_t)$, the *optimal acceptance level for class* j is then obtained as, for $j \in \mathcal{J}$,

$$\begin{aligned} x_t^{j*}(I_t, Q_t, N_t, E_t) &= \min\{N_t^j, [I_t + [S]_{t-L}^t - [N_t]_1^{j-1}]^+, \\ &\qquad [Q_t + [K]_{t-L}^t - [N_t]_1^{j-1}]^+, \\ &\qquad [I_t + S_t - \hat{y}_t^j(D_t, E_t) - [N_t]_1^{j-1}]^+\}, \end{aligned} \quad (29)$$

where in the above expression, the first term is the class j demand in period t, the second and third terms are the remaining inventory and the remaining capacity to meet the class j demand after fulfilling the first $j-1$ class demand, respectively, and the last term is the optimal base demand acceptance level for class j demand.

We note that (27) and (29) provide the explicit functional forms of the optimal policy (OPT). Specifically, for class j, OPT makes the order acceptance decision sequentially starting from class 1; it will reject the entire class j demand if class $j-1$ demand is either rejected or partially accepted; otherwise, it accepts class j demand until either all N_t^j is accepted, or one of the lead time resources to meet the class j demand, i.e., $[I_t + [S]_{t-L}^t - [N_t]_1^{j-1}]^+$ or $[Q_t + [K]_{t-L}^t - [N_t]_1^{j-1}]^+$, is exhausted, or the optimal base demand acceptance level for class j, $[I_t + S_t - \hat{y}_t^j(D_t, E_t) - [N_t]_1^{j-1}]^+$, is reached, whichever occurs first. Furthermore, (29) also shows the simplicity of the policy structure. The transformation between $x_t^{j*}(I_t, Q_t, N_t, E_t)$ and $\hat{y}_t^j(D_t, E_t)$ reveals, in an explicit analytical form, how the maximum order acceptance level for class j depends on each component of the state (N_t, I_t, Q_t, E_t). The computational efficiency of the optimal policy values $\{x_t^{j*}(I_t, Q_t, N_t, E_t)\}$ is a direct consequence of the fact that $\hat{y}_t^j(D_t, E_t)$ only depends on (D_t, E_t), rather than the individual components of the state (I_t, Q_t, N_t, E_t).

The next theorem summarizes the comparative statics of the policy parameters $\hat{x}_t^j$ and $\hat{y}_t^j$.

THEOREM 3. (i) *The optimal base demand acceptance level* $\hat{x}^j(I_t, Q_t, E_t)$ *of the first* j *classes, defined in* (26), *satisfies*

$$\hat{x}_t^j(I_t + 1, Q_t + 1, E_t) = \hat{x}_t^j(I_t, Q_t, E_t) + 1, \quad j \in \mathcal{J}. \quad (30)$$

(ii) *The inventory rationing level* $\hat{y}_t^j(D_t, E_t)$ *of class* j, *defined in* (27), *is nonincreasing in the resource imbalance level* D_t, *i.e.*,

$$\hat{y}_t^j(D_t, E_t) - 1 \leq \hat{y}_t^j(D_t + 1, E_t) \leq \hat{y}_t^j(D_t, E_t), \quad j \in \mathcal{J}. \quad (31)$$

Theorem 3(i) states that a unit pair increase in the resource levels will result in a unit increase in the optimal acceptance level $\hat{x}_t^j(I_t, Q_t, E_t)$. This is evident from the linear transformation $\hat{y}_t^j(D_t, E_t) = Q_t + K_t - \hat{x}_t^j(I_t, Q_t, E_t)$: because the rationing level is invariant to I_t and Q_t as long as their difference remains unchanged, a pairwise increase in the resource levels must result in a unit increase in $\hat{x}_t^j(I_t, Q_t, E_t)$. Theorem 3(ii) states that one unit increase of the resource imbalance level will lower the optimal ending inventory rationing level, but by at most one unit. The intuition is that a higher net capacity "overage" dampens the incentive to reserve inventory for future use, thus preventing the likelihood of further resource imbalance caused by capacity expiration. Therefore, *RIB rationing* is a critical principle for ATP-A systems: the system favors a balanced supply of the perishable and nonperishable resources, and aims at reducing their mismatch via demand management.

4.2. Special Cases

In this section, we discuss two special cases of the ATP-A model, including the single-resource ATP system and the ATP production system.

4.2.1. Single-Resource ATP Models. When only a unit of inventory or a unit of capacity is required to produce an end product, our system reduces to a single-resource ATP system, with either the perishable or nonperishable resource. It can be shown that for a singe-resource ATP, the key drivers of the order acceptance policy are class prioritization and resource rationing. First, the optimal acceptance policy must be of a LICF type; second, because resource imbalance is no longer an issue, the rationing level $\hat{y}_t^j(E_t)$ depends only on the forecast E_t and is independent of (Q_t, N_t) or (I_t, N_t). Consequently, the computational efficiency of the policy parameters can be further improved. For the single-resource ATP, we can also allow the class-specific resource requirement. For example, in the ATP *inventory* system, let a^j be the inventory requirement to fill a unit of class j demand. We index the classes such that $r^1/a^1 > r^2/a^2 > \cdots > r^J/a^J$, where r^j/a^j is the profit margin per unit inventory of class j (a similar approach can be found in the market selection literature; see, e.g., Romeijn et al. 2007, Taaffe et al. 2008). We can show that class prioritization and resource rationing remain the key drivers of the optimal acceptance policy. Most ATP and rationing models are of the single, nonperishable resource type (see, e.g., Ha 1997, Frank et al. 2003, Duran et al. 2007, Chen et al. 2008). In contrast, our single-resource ATP models extend the existing literature by capturing new modeling features including the dynamic short-term demand forecasts, multiple demand classes, multiple periods, positive lead time and perishable resources. In addition, as we pointed out in Remark 2, RIB rationing is more general than the one used in the above literature, as our rationing control threshold permits either reserving the current resource for future orders or booking future resources for current orders.

4.2.2. An ATP Production System Model with Zero Lead Time. Our ATP-A model does not require simultaneous availability of both resource types to produce an end product. For this assembly system, it only requires, for each accepted demand, the availability of its constituting components during the lead time. It fits into many ATP environments where one major component is produced internally and another is supplied externally, and the two components are then assembled into the end product to meet customer demand. In contrast, in an *ATP production* (*ATP-P*) *system*, production can take place only when a unit of capacity and a unit of inventory (interpreted as raw material) are available *simultaneously*. The analysis for the APT-P system is more complex than for the ATP-A system, because in ATP-P we cannot simply remove the committed resources from the available resources. However, when the lead time is zero, i.e., when an accepted demand must be fulfilled in the same period, the system must guarantee simultaneous availability of both resource types for the accepted demand. Therefore, the ATP-A system with zero lead time is a special case of the ATP-P system with zero lead time. As such, our study serves as the first step to understand the basic properties of the ATP-P system. The study of the ATP-P system with a positive lead time poses an interesting future research topic.

We close this section by noting that variable production costs can be easily incorporated into our model. For example, let c^j be the unit production cost of class j. Then $\sum_{j\in\mathcal{J}} c^j x_t^j$ is the total production cost in period t. If we modify the profit margin r^j to $\tilde{r}^j = (r^j - c^j)$ and rank order the classes such that $\tilde{r}^j$ is decreasing in j, then all our results still hold. Likewise, our results extend to the case when there is a class-specific penalty cost for not accepting a demand from that class.

5. Numerical Results

This section conducts extensive numerical tests to investigate (1) the behavior of the optimal policy and (2) the value of dynamic pseudo-order information. To achieve the first objective, in §5.1, we illustrate the three properties of the optimal policy—class prioritization, RIB rationing, and CID matching, as discussed in §4.1, using a three-class ATP-A system. Because our focus here is on the three properties of the optimal policy, which hold irrespective of the dynamic forecast E_t, to isolate the policy effect as well as to suppress the "curse of dimensionality" in our numerical tests, we use the long-term pseudo-order forecast, i.e., E_t is not updated. To achieve the second objective, in §5.2, we investigate a two-class ATP *inventory* (ATP-I) system with only the nonperishable resource type, focusing exclusively on the value of dynamic pseudo-order information updating and the robustness of the optimal policy in the presence of systematic short-term forecasting errors. The motivation to investigate this simpler system is to isolate the impact of this key factor with manageable computational effort. We believe that the lessons learned on the value of pseudo-order information from the study of this simpler system are also applicable to more complex systems.

5.1. Three-Class ATP Assembly Systems

We first provide a numerical example for a three-class ATP-A system. Let F_1, F_2, and F_3 be the distributions of the uniform random variable $\mathcal{U}[1, 5]$, $\mathcal{U}[6, 10]$, and $\mathcal{U}[11, 15]$, respectively. We assume that in each period the demand of each class follows the stationary distribution

$$F_L(N) = P(X_L^j \le N) = 0.2 + 0.3F_1(N) + 0.3F_2(N) + 0.2F_3(N), \quad N \ge 0, \qquad (32)$$

Table 1 **Inventory Rationing Level $\hat{y}_t^j(D_t)$:** $r=(6,3,1)$, $h=p=0.5$, $(S,K)\equiv(10,15)$, $L\in\{0,2,4\}$

Lead time	Demand class	Remaining period	Resource imbalance level D_t											
			−4	−3	−2	−1	0	1	2	3	4	5	6	7
$L=0$	$j=2$	$t=2$	0	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
		3	0	0	0	0	0	0	0	0	0	0	0	0
		4	0	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
		5	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
	$j=3$	$t=2$	4	4	4	4	4	4	4	4	4	4	4	4
		3	1	1	1	1	1	1	1	1	1	1	1	1
		4	1	1	1	1	1	1	1	1	1	1	1	1
		5	0	0	0	0	0	0	0	0	0	0	0	0
$L=2$	$j=2$	$t=2$	0	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
		3	−4	−4	−4	−4	−4	−5	−5	−5	−5	−5	−5	−5
		4	−5	−5	−5	−5	−5	−5	−5	−5	−5	−5	−5	−5
		5	−4	−4	−4	−4	−4	−5	−5	−5	−5	−5	−5	−5
	$j=3$	$t=2$	4	4	4	4	4	4	4	4	4	4	4	4
		3	0	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
		4	0	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
		5	0	0	0	0	0	0	0	0	0	0	0	0
$L=4$	$j=2$	$t=2$	0	−1	−2	−2	−2	−2	−2	−2	−2	−2	−2	−2
		3	−4	−4	−4	−4	−4	−5	−5	−5	−5	−5	−5	−5
		4	−5	−5	−5	−5	−5	−5	−6	−6	−6	−6	−6	−6
		5	−5	−5	−5	−5	−5	−5	−5	−5	−5	−5	−5	−5
	$j=3$	$t=2$	4	4	4	4	4	4	4	4	4	4	4	4
		3	0	0	0	0	0	0	0	0	0	0	0	0
		4	0	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
		5	−1	−1	−2	−2	−2	−2	−2	−2	−2	−2	−2	−2

where (0.2, 0.3, 0.3, 0.2) in the above expression represents, respectively, the stationary distribution of a Markov chain in state 0 (order cancellation) and the distribution state i with the distribution F_i, $i=1,2,3$ (see §5.2.1 for an explanation of (32) as a long-term forecast of pseudo orders). The expected stationary demand for each class in each period works out as $\mathbb{E}[X_t^i]=0.3\times 3+0.3\times 8+0.2\times 13=5.9$, and the total demand across all three classes is $\mathbb{E}[\sum_i X_t^i]=17.7$. The profit margins are set as $r^1=6$, $r^2=3$, $r^3=1$. The unit holding cost is $h=0.5$, the unit capacity idleness cost is $p=0.5$, and the ATP-A execution period is $T=5$. Also let the planned resources in each period be $S_t=10$ and $K_t=15$ for each t. We set the delivery lead time as $L\in\{0,2,4\}$. Our parameter setting is consistent with the inventory literature (see, e.g., Yang et al. 2005).

Table 1 reports the inventory rationing levels $\hat{y}_t^j(D_t)$ of classes 2 and 3 as a function of the remaining period t, the class index j, the resource imbalance level D_t, and the delivery lead time L. Because in the last period there is no rationing control we omit $t=1$. We have the following observations.

1. As expected, for each period and each class, our computation shows that the optimal inventory rationing level $\hat{y}_t^j(D_t)$ only depends on the resource imbalance level $D_t=Q_t-I_t$, reflecting RIB rationing as a key property of the optimal policy. Table 1 also shows that $\hat{y}_t^j(D_t)$ can be either positive or negative, as we noted in Remark 2. For example, when the lead time $L=4$ and the resource imbalance level $D_2=0$, we have $y_2^3(0)=4$, meaning that we reserve four units of the available inventory in period 2 for the future, classes 1 and 2 demand. In doing so, we keep $9=15-(10-4)$ units of capacity in period 2 idle to reduce the future resource imbalance level. As another example, for $L=2$ and $D_5=0$, $y_5^2(0)=-4<0$, thus we book four units of the future inventory for class 2 and allow $1=15-[10-(-4)]$ unit of capacity to be idle in period 5.

2. In addition, Table 1 verifies the comparative statics of the inventory rationing level $\hat{y}_t^j(D_t)$ discussed in (28) and Theorem 3(ii), i.e., $\hat{y}_t^j(D_t)$ increases in the class index j and decreases in the resource imbalance level D_t at the rate of no more than one. As seen from Table 1, when D_t is sufficiently large (i.e., significant capacity overage), a unit increase in the imbalance level tends not to affect the optimal inventory rationing level. This is intuitive because when inventory is the bottleneck, the firm needs to save scarce inventory for the future demand, and can afford to let some unneeded capacity idle (in fact, Table 1 shows that all $\hat{y}_t^j(D_t)$ do not change for $D_t\geq 1$). On the other hand, when $D_t<0$ and is sufficiently small (i.e., significant inventory overage), capacity becomes the bottleneck and the firm needs to prevent capacity idleness by reserving less inventory for the future; consequently, a unit increase in the imbalance level

tends to reduce the optimal inventory rationing level by one (e.g., Table 1 shows that for $L = 4$, $\hat{y}_2^2(-4) = 0 \geq -1 = \hat{y}_2^2(-3)$).

3. Finally, Table 1 suggests that the rationing level $\hat{y}_t^j(D_t)$ decreases in the lead time L. This is because, by (27), $\hat{y}_t^j(D_t)$ is the inventory level at which the unit revenue of class j equals the marginal value of a *unit pair* of resources at the resource levels $(\hat{y}_t^j(D_t), \hat{y}_t^j(D_t) + D_t + K_t - S_t)$. The increase in lead time L gives more order acceptance flexibility, and therefore decreases the marginal value of a unit resource pair. As the consequence of a longer lead time L, the increased lead time resource availability and decreased rationing level jointly raise the *actual acceptance of* class j demand $x_t^{j*}(I_t, Q_t, N_t)$, defined in (29), which translates into a higher profit for the firm.

Next, we examine the impact of resource planning (S_t, K_t) on system performance. With slight abuse of notation, we denote $V_T(S, K)$ as the expected total profit with the zero initial net resources $(I_T, Q_T) = (0, 0)$ and the resource replenishment plan $(S_t, K_t) = (S, K)$ for all t. Figures 1 and 2 report $V_T(S, K)$ under different resource replenishment plans (S, K), with $S \in (0, 25]$ and $K \in (0, 25]$. We observe that for each given planned capacity level $K \in \{10, 15, 20, 25\}$, the planned inventory S that is slightly below K achieves the maximum profit. This is intuitive because the left-over inventory can be accumulated over time and the idle capacity expires after each period. Therefore, to match the two resource levels, the planned capacity should be above the planned inventory, according to the CID matching principle. In addition, the maximum profit of each value function V_T in Figure 2 first increases as the matched resource levels rise to the expected demand (undersupply) and then decreases as the matched resource levels significantly exceed the expected demand (oversupply). That is, for a given

Figure 1 $V_T(S, K)$, $r = (6, 3, 1)$, $h = p = 0.5$, $\mathbb{E}[\sum_i X_t^i] = 17.7$

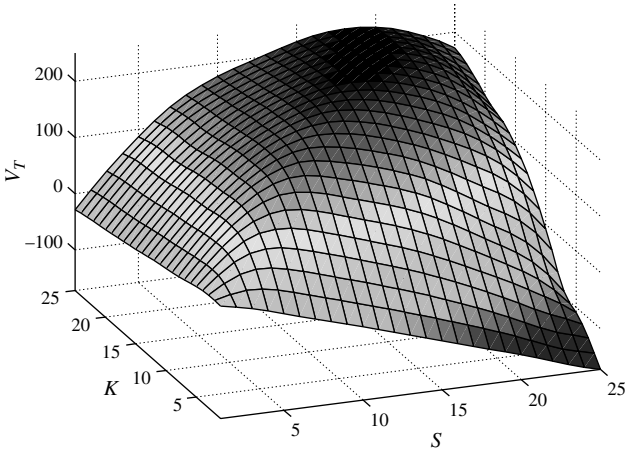
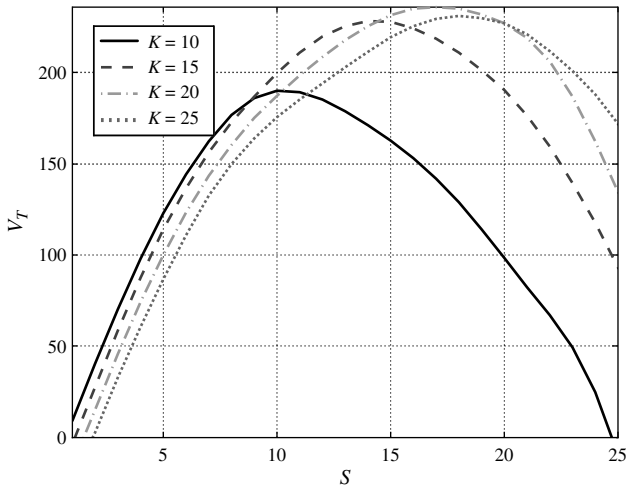


Figure 2 $V_T(S, K)$, $K \in \{10, 15, 20, 25\}$, $r = (6, 3, 1)$, $h = p = 0.5$, $\mathbb{E}[\sum_i X_t^i] = 17.7$



demand forecast, the resource replenishment plans with their levels slightly above the 45-degree line on the (S, K)-plane tend to achieve a higher profit than other imbalanced resource replenishment plans, provided that the matched resources are properly balanced with the total mean demand. In particular, the balanced replenishment plan, $(S, K) = (17, 19)$, best matches the total expected demand 17.7 and yields the maximum profit 238.2 in our setting. Therefore, our CID matching principle provides both tactical and strategic support for the ATP planning and execution.

Using Figures 3 and 4, we demonstrate that the value function $V_T(S, K)$ is *diagonally dominated* in (S, K) (i.e., the marginal value of one resource decreases as both resource levels increase, as shown in Theorem 2(iii)), but is not necessarily supermodular (i.e., the marginal value of one resource may

Figure 3 $\Delta_K V_T(S + k, K + k)$, $r = (6, 3, 1)$, $h = p = 0.5$, $(S, K) = \{(0, 0), (0, 2), (2, 0)\}$, $\mathbb{E}[\sum_i X_t^i] = 17.7$

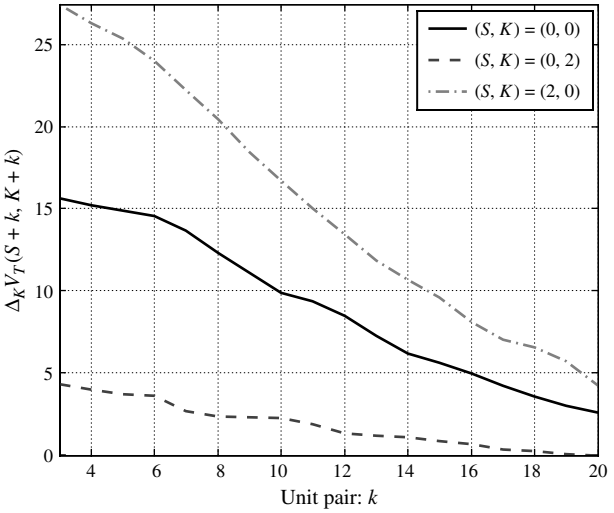
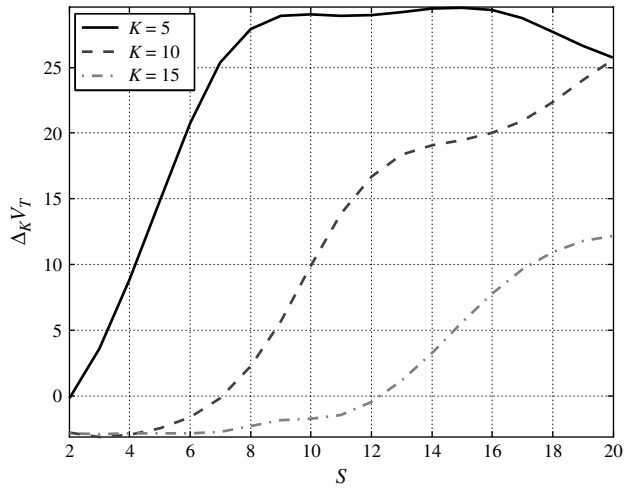


Figure 4 $\Delta_K V_T(S,K)$, $r = (6,3,1)$, $h = p = 0.5$, $K \in \{5,10,15\}$, $\mathbb{E}[\sum_i X_t^i] = 17.7$



not increase as another resource level increases, see Remark 1). Figure 3 reports the marginal value of the planned capacity $\Delta_K V_t(S+k, K+k) = V_t(S+k, K+k) - V_t(S+k, K+k-1)$ at resource level $(S+k, K+k)$ along the 45-degree line originated from (S,K) on the (S,K)-plane, and Figure 4 reports the marginal value of the planned capacity $\Delta_K V_T(S,K) = V_T(S,K) - V_T(S,K-1)$ as a function of the planned inventory. In Figure 3, we observe that the marginal value of capacity $\Delta_K V_T(S+k, K+k)$ decreases as both resource levels increase pairwise from the different initial levels $(S,K) = \{(0,0),(0,2),(2,0)\}$. In Figure 4, we observe that for $K = 5$ the marginal value of capacity $\Delta_K V_T(S,5) = V_T(S,5) - V_T(S,4)$ is nonmonotone in inventory. This is because the marginal value of capacity is low when inventory is either scarce (because additional capacity cannot be utilized due to lack of inventory) or excess (because additional capacity is not needed due to lack of demand). This result again signifies the importance of CID matching in ATP-A management.

5.2. Two-Class ATP Inventory (ATP-I) Systems

In this section, we conduct a numerical study of two-class ATP inventory (ATP-I) systems. We investigate the value of incorporating the short-term demand forecast, as opposed to the long-term demand forecast, in ATP-I decisions (§5.2.1) and the robustness of the optimal policies in the presence of *systematic forecast errors* (§5.2.2).

5.2.1. Short-Term vs. Long-Term Forecasts: Value of Real-Time Information. As discussed in §1, one distinguishing characteristic of pseudo orders is demand volatility. For class j, we can capture dynamic demand signals via the Markov chain $\{E_t^j, t = 0,\dots,T\}$. We shall assume, for the remainder of the section, that $E_t^j = (e_t^j,\dots,e_{t-L}^j)$ is the distribution states of class j pseudo orders during lead time L, forecasted in period t. A fundamental question is then, What is the value of the dynamic short-term demand forecast, as opposed to the long-term forecast based on historical data, on the performance of ATP-I? And how do different types of demand forecasts affect the behavior of the optimal policy? To answer these questions, we conduct an experiment to compare performance gaps and policy differences under the two types of forecast, for a wide range of parameter settings. Our results provide the cost justification for tracing pseudo-order information over time. We design our experiment as follows.

Pseudo-Order Forecast. We assume that for any given t, class 1 demand with confirmation date s, X_s^1, $t-L-1 \le s \le t-1$, is determined by its distribution state $e_s \in \{0,1,2,3\}$, where state 0 corresponds to null demand, and states 1, 2, and 3 correspond to the uniform random variables $\mathcal{U}[1,5]$, $\mathcal{U}[6,10]$, and $\mathcal{U}[11,15]$, with distributions F_1, F_2, and F_3, respectively. The short-term forecast is captured by transition matrix

$$\mathbf{P} = \{P(e_s' \mid e_s)\} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 2/3 & 0 & 1/3 & 0 \\ 0 & 1/3 & 0 & 2/3 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad \text{for } t-1-L \le s \le t-1.$$

Such Markov matrices have been used by other researchers to model stochastic monotone capacity evolutions (e.g., Yang et al. 2005) and queueing systems (e.g., Xu et al. 2007). For the long-term forecast, we assume that the firm ignores class 1 demand signals, and forecasts its future distribution state in any period by the stationary distribution of the Markov chain, $\pi = (0.2, 0.3, 0.3, 0.2)$. Then, X_L^1, the long-term class 1 demand forecast in any period, follows the stationary distribution

$$F_L(N) = P(X_L^1 \le N) = 0.2 + 0.3F_1(N) + 0.3F_2(N) + 0.2F_3(N), \quad N \ge 0. \tag{33}$$

For simplicity, we assume that class 2 demand follows the stationary distribution $F_L(N)$ of X_L^1. With either forecast, we assume that, in the true world, the distribution state of class 1 evolves according to the short-term forecast situation, and class 2 demands are independent and identically distributed following (33).

Resource Availability ρ. We characterize resource tightness via a *resource availability index*, defined as the ratio of the planned inventory S $(=S_t)$ and the expected total demand in each period: $\rho := S/\mathbb{E}(X^1 + X^2)$. This measure captures the imbalance between supply and demand. We let $\rho \in (0, 1.2)$,

where the three ranges, $0 < \rho \le 0.4$, $0.4 < \rho \le 0.8$ and $0.8 < \rho \le 1.2$, respectively, represent the systems with scarce, moderate, and ample inventory.

Profit Ratio γ. We characterize customer heterogeneity, or profit disparity, by the profit ratio $\gamma := r^1/r^2$, with $r^1 + r^2 = 10$ and $\gamma \in (1, 3)$. Customer heterogeneity increases when γ increases.

We set other parameter values for our experiment as $L = 2$ and $T = 5$. Let OPT_S and OPT_L be the optimal policies under the short-term and long-term forecasts, respectively. Because the true world evolves according to the Markov chain with transition matrix $\mathbf{P}$ and the short-term forecast captures this dynamic information, we expect that OPT_S outperforms OPT_L. We define the performance gap under the two forecasts by $\Delta_S V_L^* := ((V_L^{\text{OPT}_S} - V_L^{\text{OPT}_L})/V_L^{\text{OPT}_L}) \times 100\%$, where $V_L^{\text{OPT}_L}$ and $V_L^{\text{OPT}_S}$ are the expected profits under OPT_S and OPT_L, with lead time L, respectively.

Figure 5 reports performance gap $\Delta_S V_2^*$ over the tested region (ρ, γ). As seen, OPT_S outperforms OPT_L over the *entire* (ρ, γ) region, meaning the performance of ATP-I under the short-term forecast *always* dominates its counterpart under the long-term forecast. The maximum improvement can be as high as 7%, which occurs when the planned inventory availability is moderate ($\rho \approx 0.50$) and customers are highly heterogeneous ($\gamma \approx 3$). This parameter range is also where inventory rationing is most effective. When resource availability is low ($\rho \le 0.2$), the scarce inventory will be used almost exclusively for class 1, and the maximum profit is achieved by using class prioritization alone. On the other hand, when inventory is abundant ($\rho \ge 0.8$) or the profit ratio γ is low ($\gamma \le 1.2$), both demand classes will be accepted and there is no need to reserve inventory for future class 1 demand or reserve future inventory for the current class 1 demand. As such, when resource availability is sufficiently low or high ($\rho < 0.2$ or $\rho > 0.8$), the long-term

Figure 5 **Value of Short-Term Forecast**

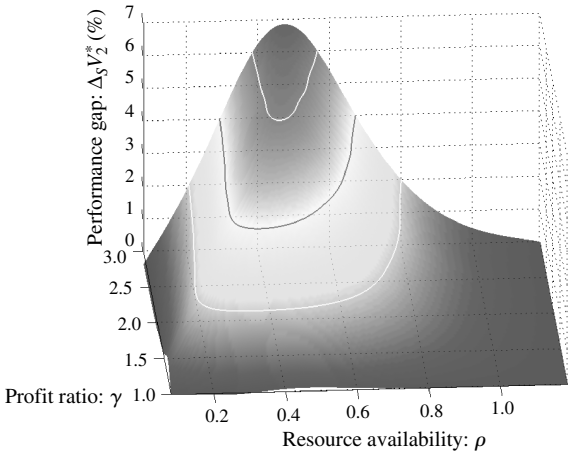
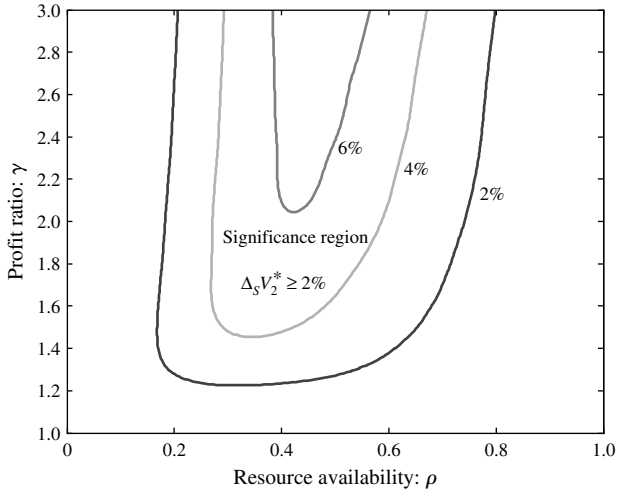


Figure 6 **Partition of Significance Regions**



forecast is sufficient to achieve good ATP performance ($\Delta_S V_2^* < 2\%$). It is in the moderate capacity range with highly heterogeneous customers that both class prioritization and rationing must be implemented, and real-time demand information provided by the short-term forecast proves to be most valuable. This result also conforms with the similar findings in other multiple-class rationing models (see, e.g., Ha 1997, Akçay et al. 2010, Gao et al. 2011).

In Figure 6, we further identify several significance (ρ, γ) regions under which the performance improvements of OPT_S reach 2%, 4%, and 6%, respectively. It is worth noting that the 2% significance region in Figure 4 covers a wide range of parameter values, $0.2 \le \rho \le 0.8$ and $\gamma \ge 1.2$, indicating that for most practical ATP systems, it pays to incorporate the short-term demand forecast in ATP decision making.

We now examine the impact of the short-term and long-term forecasts on the behavior of the optimal policy. Figure 7 shows the sensitivity of the optimal rationing levels $\{\hat{y}_t^2(E_t)\}$ with respect to the short-term forecast states E_t, when resource availability is tight ($\rho = 0.42$) and the profit ratio is high ($\gamma = 3$).

When the short-term forecast E_t predicts higher class 1 demand in the next period ($E_t = 2$ or 3), the system responds by *raising* the rationing threshold $y_t^2(E_t)$ for class 2. Conversely, when E_t predicts low class 1 demand ($E_t = 0, 1$) in the next period, the system dynamically adjusts $y_t^2(E_t)$ by reserving less inventory for the next period. Comparing these dynamic rationing levels $\{\hat{y}_t^2(E_t)\}$ with the static rationing level $\hat{y}_t^2$ when real-time demand information is absent, we find that $\hat{y}_t^2$ is roughly the *weighted average* of $\{\hat{y}_t^2(E_t)\}$, where the weights $\{0.2, 0.3, 0.3, 0.2\}$ are the stationary probabilities of the Markov chain in different distribution states. Observe that $y_t^2(0) < y_t^2(1) < y_t^2(2) < y_t^2(3)$ for any t, and the gaps among them are significant, suggesting that the rationing level is

Figure 7 **Impact of Short-Term Forecast E_t on Inventory Rationing Levels $\hat{y}_t^2$ and $\hat{y}_t^2(E_t)$: $E_t = \{0, 1, 2, 3\}$, $\gamma = 3$, $\rho = 0.42$**

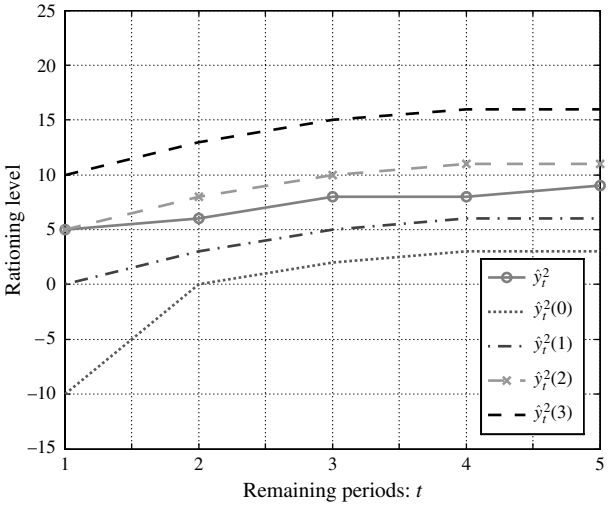
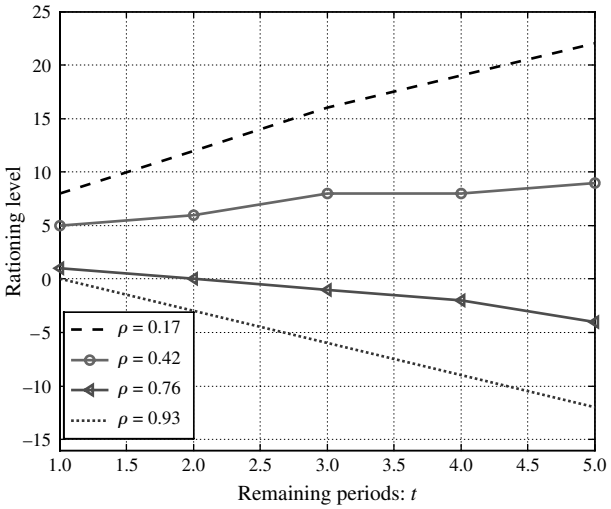


Figure 8 **Impact of Capacity Availability ρ on Inventory Rationing Level $\hat{y}_t^2$ Under Long-Term Forecast: $\rho \in \{0.17, 0.42, 0.76, 0.93\}$, $\gamma = 3$**



very sensitive to real-time demand information. This explains the significant performance gap under the two forecasts for the system with a tight resource availability and highly heterogeneous customers, as shown in Figure 5.

Figure 8 depicts the impact of the resource availability ρ on the behavior of the rationing level $\hat{y}_t^2$ for class 2 under the long-term forecast. It reveals that, when supply is scarce ($\rho = 0.17$ or 0.42), the ATP-I operates as an *inventory reservation system* by setting *a positive* inventory rationing level for class 2, and this positive rationing level $\hat{y}_t^2$ is decreasing as time t approaches the end of the planning horizon. In contrast, when supply is ample ($\rho = 0.93$), the ATP-I runs as a *demand backlogging system* by setting *a negative* inventory rationing level (i.e., using future inventory for the current demand), and this negative rationing level $\hat{y}_t^2$ is increasing as time t approaches the end of the planning period. When supply is moderate ($\rho = 0.76$), the system operates as a demand backlogging system when t is large and as an inventory reservation system when t is small.

5.2.2. Robustness of the Optimal Policy. As shown in §5.2.1, OPT_S always outperforms OPT_L, as long as both forecasts are accurate. Unfortunately, the short-term forecast of pseudo orders is based on sales data solicited from sales personnel, who report and update periodically the attributes of pseudo orders. Without a sound sales force monitoring system in place, salespersons may *systematically* overclaim or underclaim future sales, i.e., commit *systematic forecast errors*. Furthermore, volatility of demand signals makes *random forecast errors* inevitable. Although developing accurate statistical models for pseudo-order forecasts is beyond the scope of this paper, it is important to recognize that a short-term forecast can be more error prone than a long-term forecast, even if sound data collection and statistical mechanisms are deployed to filter out potential biases in salespersons' reporting. This naturally leads to the following questions: How is the effectiveness of the optimal policy affected by systematic short-term forecast errors? When does OPT_L under an accurate long-term forecast (which ignores current demand signals and relies on historical data) outperform OPT_S under an inaccurate short-term forecast (which relies on real-time demand signals and ignores historical data)? To answer those questions, we compare the performance of the "optimal policy" OPT_S under a biased short-term forecast against that of OPT_L under the accurate long-term forecast. Because the impact of the forecast errors is more prominent for high-valued orders, for expositional simplicity, we assume that only class 1 demands are subject to forecast errors. The detailed experimental results and our detailed discussion are relegated to §EC.6 in the electronic companion. Here, we only outline our experiment and summarize the major findings.

To test the robustness of the optimal policy under a biased short term forecast, we carry out our experiment as follows: First, in the true world, pseudo orders evolve according to the example given in §5.2.1. Second, the accurate long-term forecast for each class is given in (33). Third, the short-term forecast errors occur to class 1 demand in states e_2 and e_3, according to $\hat{\mathcal{U}}_2 := \mathcal{U}[6, 10] \pm \varepsilon$ and $\hat{\mathcal{U}}_3 := \mathcal{U}[11, 15] \pm \varepsilon$, where $\varepsilon \in \{1, 2, 3\}$ represents the magnitude of the error. Let the accurate long-term forecast, given by (33), be used for class 2 demand. We classify the forecast errors for class 1 demand into four types, $(+\varepsilon, +\varepsilon)$, $(-\varepsilon, +\varepsilon)$, $(+\varepsilon, -\varepsilon)$, and

$(-\varepsilon, -\varepsilon)$, and call them Types I, II, III, and IV errors, respectively. By letting ε vary between 1 to 3 for each of the four types, we have totally $3 \times 4 = 12$ forecast error scenarios, reflecting different magnitudes and types of forecast bias. For each scenario, we examine $3 \times 5 \times 14 = 210$ combinations of system settings: 3 levels of lead time $L \in \{0, 2, 4\}$, 5 levels of profit ratio $\gamma \in \{1, 1.5, 2, 2.5, 3\}$, and 14 levels of resource availability $\rho \in [0.085, 1.186]$ (corresponding to $S = 1, 2, \ldots, 14$). Altogether, we test $12 \times 210 = 2{,}520$ problem instances, which include most practical system configurations and reasonable magnitudes and types of forecast errors (see Table EC.2 in §EC.6 in the electronic companion). In each scenario, we report, in Table EC.2, the percentage of problem instances that the optimal policy OPT_S outperforms OPT_L, and the worst performance gap of OPT_S against OPT_L. Our major findings are as follows:

(1) OPT_S statistically outperforms OPT_L in 9 of 12 scenarios, and 8 of these 9 scenarios correspond to small to moderate forecast errors. For the other 3 scenarios, where the mean forecast errors reach $\pm 25.42\%$, the percentage of problem instances that OPT_L dominates OPT_S is slightly over 50%. These results indicate that OPT_S is robust and outperforms OPT_L for small to moderate forecast errors; however, OPT_L should be adopted when the short-term forecast is highly biased.

(2) Among the four types of errors, OPT_S appears most vulnerable, in terms of the worst performance, to the error type that significantly overestimates the mean and underestimates the variability. To understand this, we note, in general, overestimating the mean results in overrationing, and underestimating the variability also leads to overrationing. Therefore, the compound effect of overrationing under this type of errors causes ATP to *overreject* low-profit orders, leading to a significant profit loss. Our results also suggest that the consequence of overrationing is more detrimental than that of underrationing, because the former results in overrejection of low-profit orders, hence *unutilized* system resources (which would be used for low-profit orders under an accurate forecast), and the latter means overacceptance of low-profit orders, hence *underutilized* system resources (which would be used for high-profit orders under an accurate forecast rather than low-profit ones). Therefore, OPT_S is more robust against the errors that cause underrationing than the errors that cause overrationing.

(3) Because the effectiveness of OPT_S depends on the quality of pseudo-order information, to support ATP execution, firms need effective sales force monitoring systems to ensure high-quality sales information reporting and updating, and statistical mechanisms to filter out systematic forecast errors from salespersons' reporting. In particular, those decision support systems should provide the safeguard against salespersons' tendencies to overestimate the likelihoods or quantities of potential orders, because, as our results show, overestimating potential sales causes more harm than underestimating future sales to firms.

6. Concluding Remarks

In this paper, we study an order promising problem in an ATP-A system with pseudo orders and two types of resources. Our contributions to the academic literature and ATP management are summarized as follows.

From an academic research perspective, we develop a Markov chain modeling framework to capture stochastic attributes of ever-changing real-time demand information, which appears to be a new research endeavor. We integrate the dynamic short-term forecast with the optimal order acceptance policy and derive strong analytical results. We show that *class prioritization*, *CID matching*, and *RIB rationing* are the three key principles driving the optimal policy in the ATP-A system. Our study enriches the ATP models and methods, where few analytical results exist even for the models with a single-resource type and a stable demand forecast. To a larger extent, our modeling paradigm demonstrates how real-time, dynamic data sources can be used to support execution-level decision making, although the bulk of research into supply chain decision models has been based on stable demand forecasts.

From a management perspective, we provide a simple policy structure that appeals to managers' intuition and requires reduced computational efforts. Our comprehensive numerical tests suggest several rules of thumb on *when and how the dynamic short-term forecast should be incorporated into the ATP execution*: (1) the optimal policy under an accurate short-term forecast significantly improves the ATP performance over the optimal policy under an accurate long-term forecast, when the planned system resource is *moderate* and customer heterogeneity is *high*; (2) the optimal policy under a short-term forecast with small to moderate forecast errors is robust, and outperforms the optimal policy under an accurate long-term forecast. However, the optimal policy under an accurate long-term forecast should be used when short-term forecast errors are significant; and (3) to take full advantage of pseudo-order information, firms need effective sales force monitoring systems and forecast mechanisms. These decision support systems should provide a safeguard against salespersons' tendencies to overestimate potential orders. Our results suggest that, as opposed to underestimating, overestimating

potential sales, as is often the case for a sales force, causes more harm to firms. Firms with sound decision support systems can gain a significant competitive advantage by using pseudo-order information to increase profits, improve customer service, and shorten delivery lead times.

There are several possible future research directions. First, although it appears that the optimal policy does not possess a simple form when an end product has a general BOM structure, in some special cases, such as when a demand class requires a *class-specific* inventory resource and a *common* capacity resource, or when a demand class has a *class-specific* lead time, we believe our modeling framework and solution approach are applicable to derive analytical properties of the optimal policy. We can utilize the insights gained from these simpler models to develop efficient heuristics for the system with more complex BOM requirements. Second, statistical mechanisms need to be developed to capture and model sales data, which serve as the inputs to our pseudo-order model. Third, it is of interest to understand how pseudo-order information can be used within a model that makes integrated resource management, order promising, and production control decisions. Finally, the ATP production system with a positive lead time, as discussed in §4.2, poses another interesting research topic.

Acknowledgments

The authors thank department editor Martin Lariviere, the associate editor, and the three referees for helpful comments and constructive suggestions that have resulted in a significantly improved paper. The research of Susan H. Xu was supported in part by the National Science Foundation [Grants 0825928, 101691].

References

Akçay, Y., A. Balakrishnan, S. H. Xu. 2010. Dynamic assignment of flexible service resources. *Production Oper. Management* **19**(3) 279–304.

Aviv, Y. 2001. The effect of collaborative forecasting on supply chain performance. *Management Sci.* **47**(10) 1326–1343.

Baker, K. R., J. W. M. Bertrand. 1981. An investigation of due-date assignment rules with constrained tightness. *J. Oper. Management* **1**(3) 109–120.

Balakrishnan, A., J. Geunes. 2000. Requirements planning with substitutions: Exploiting bill-of-materials flexibility in production planning. *Manufacturing Service Oper. Management* **2**(2) 166–185.

Ball, M. O., C.-Y. Chen, Z.-Y. Zhao. 2004. Available to promise. D. Simchi-Levi, S. D. Wu, Z.-J. Shen, eds. *Handbook of Quantitative Supply Chain Analysis: Modeling in the E-Business Era*, Chap. 11. Kluwer Academic, Boston.

Benjaafar, S., M. ElHafsi. 2006. Production and inventory control of a single product assemble-to-order system with multiple customer classes. *Management Sci.* **52**(12) 1896–1912.

Chen, C. Y., Z. Y. Zhao, M. O. Ball. 2008. An available-to-promise model for periodical order promising of a non-perishable resource. Working paper, George Mason University, Fairfax, VA.

Chen, F., J.-S. Song. 2001. Optimal policies for multiechelon inventory problems with Markov-modulated demand. *Oper. Res.* **49**(2) 226–234.

DeCroix, G. A., V. S. Mookerjee. 1997. Purchasing demand information in a stochastic-demand inventory system. *Eur. J. Oper. Res.* **102**(1) 36–57.

de Kok, T. G. 2000. Capacity allocation and outsourcing in a process industry. *Internat. J. Production Econom.* **68**(3) 229–239.

Duenyas, I., W. J. Hopp, Y. Bassok. 1997. Production quotas as bounds on interplant JIT contracts. *Management Sci.* **43**(10) 1372–1386.

Duran, S., T. Liu, D. Simchi-Levi, J. Swann. 2007. Optimal production and inventory policies of priority and price-differentiated customers. *IIE Trans.* **39**(9) 845–861.

Duran, S., T. Liu, D. Simchi-Levi, J. L. Swann. 2008. Policies utilizing tactical inventory for service-differentiated customers. *Oper. Res. Lett.* **36**(2) 259–264.

Ervolina, T., B. Dietrich. 2001. Moving toward dynamic available to promise. Working paper. IBM Research Division, T. J. Watson Research Center, Yorktown Heights, NY.

Frank, K. C., R. Q. Zhang, I. Duenyas. 2003. Optimal policies for inventory systems with priority demand classes. *Oper. Res.* **51**(6) 993–1002.

Gallego, G., Ö. Özer. 2001. Integrating replenishment decisions with advance demand information. *Management Sci.* **47**(10) 1344–1360.

Gao, L., M. Gorman, Z. Li. 2011. Optimal load acceptance polices for intermodal freight transportation. Working paper. University of California, Riverside, Riverside.

Gayon, J. P., S. Benjaafar, F. de V'ericourt. 2009. Using imperfect advance demand information in production-inventory systems with multiple customer classes. *Manufacturing Service Oper. Management* **11**(1) 128–143.

Glasserman, P., D. D. Yao. 1994. *Monotone Structure in Discrete-Event Systems*. John Wiley & Sons, New York.

Gupta, D., L. Wang. 2007. Capacity management for contract manufacturing. *Oper. Res.* **55**(2) 367–377.

Ha, A. Y. 1997. Inventory rationing in a make-to-stock production system with several demand classes and lost sales. *Management Sci.* **43**(8) 1093–1103.

Heath, D., P. Jackson. 1994. Modeling the evolution of demand forecast with application to safety stock analysis in production/distribution systems. *IIE Trans.* **26**(3) 17–30.

Hopp, W. J., M. R. Sturgis. 2001. A simple, robust leadtime-quoting policy. *Manufacturing Service Oper. Management* **3**(4) 321–336.

Kempf, K. G. 2004. Control-oriented approaches to supply chain management in semiconductor manufacturing. *Proc. 2004 Amer. Control Conf., Boston*.

Kilger, C., L. Schneeweiss. 2000. Demand fulfillment and ATP. H. Stadtler, C. Kilger, eds. *Supply Chain Management and Advanced Planning: Concepts, Models, Software and Case Studies*. Springer, Berlin, 179–196.

Li, K. J., D. K. H. Fong, S. H. Xu. 2011. Managing trade-in programs based on product characteristics and customer heterogeneity in business-to-business markets. *Manufacturing Service Oper. Management* **13**(1) 108–123.

Li, Z., L. Gao. 2008. The effects of sharing upstream information on product rollover. *Production Oper. Management* **17**(5) 522–531.

Milner, J. M., P. Kouvelis. 2005. Order quantity and timing flexibility in supply chains: The role of demand characteristics. *Management Sci.* **51**(6) 970–985.

Moses, S., H. Grant, L. Gruenwald, S. Pulat. 2004. Real-time due-date promising by build-to-order environments. *Internat. J. Production Res.* **42**(20) 4353–4375.

Özer, Ö. 2003. Replenishment strategies for distribution systems under advance demand information. *Management Sci.* **49**(3) 255–272.

Özer, Ö., W. Wei. 2004. Inventory control with limited capacity and advance demand information. *Oper. Res.* **52**(6) 988–1000.

Robinson, A. G., R. C. Carlson. 2007. Dynamic order promising: real-time ATP. *Internat. J. Integrated Supply Management* **3**(3) 283–301.

Romeijn, H. E., J. Geunes, K. Taaffe. 2007. On a nonseparable convex maximization problem with continuous knapsack constraints. *Oper. Res. Lett.* **35**(2) 172–180.

Sethi, S. P., F. Cheng. 1997. Optimality of (s, S) policies in inventory models with Markovian demand. *Oper. Res.* **45**(6) 931–939.

Song, J.-S., P. Zipkin. 1993. Inventory control in a fluctuating demand environment. *Oper. Res.* **41**(2) 351–370.

Taaffe, K., J. Geunes, H. E. Romeijn. 2008. Target market selection and marketing effort under uncertainty: The selective newsvendor. *Eur. J. Oper. Res.* **189**(3) 987–1003.

Talluri, K. T., G. Van Ryzin. 2005. *The Theory and Practice of Revenue Management*. Springer Verlag, New York.

Tan, T., R. Gullu, N. Erkip. 2007. Modelling imperfect advance demand information and analysis of optimal inventory policies. *Eur. J. Oper. Res.* **127**(2) 897–923.

Taylor, S., G. Plenert. 1999. Finite capacity promising. *Production Inventory Management* **40**(3) 50–56.

Thas, O., J. C. W. Rayner. 2005. Smooth tests for the zero-inflated Poisson distribution. *Biometrics* **61**(3) 808–815.

Topkis, D. M. 1998. *Supermodularity and Complementarity*. Princeton University Press, Princeton, NJ.

Xu, S. H., L. Gao, J. Ou. 2007. Service performance analysis and improvement for a ticket queue with balking customers. *Management Sci.* **53**(6) 971–990.

Yang, J., Z. Qin. 2007. Capacitated production control with virtual lateral transshipments. *Oper. Res.* **55**(6) 1104–1119.

Yang, J., X. Qi, Y. Xia. 2005. A production-inventory system with Markovian capacity and outsourcing option. *Oper. Res.* **53**(2) 328–349.

Zhao, H., V. Deshpande, J. K. Ryan. 2005. Inventory sharing and rationing in decentralized dealer networks. *Management Sci.* **51**(4) 531–547.

Zhao, H., J. K. Ryan, V. Deshpande. 2008. Optimal dynamic production and inventory transshipment policies for a two-location make-to-stock system. *Oper. Res.* **56**(2) 400–410.