TUTORIALS IN
OPERATIONS RESEARCH
2005
Emerging Theory, Methods, and Applications
Harvey J. Greenberg, Series Editor • J. Cole Smith, Tutorials Chair and Volume Editor
Presented at the INFORMS Annual Meeting, November 13-16, 2005

INFORMS Tutorials in Operations Research

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

An Introduction to Revenue Management

Garrett J. van RyzinKalyan T. Talluri

To cite this entry: Garrett J. van RyzinKalyan T. Talluri. An Introduction to Revenue Management. *In* INFORMS Tutorials in Operations Research. Published online: 14 Oct 2014; 142-194.
https://doi.org/10.1287/educ.1053.0019

TUTORIALS IN
OPERATIONS RESEARCH
INFORMS 2005

© 2005 INFORMS | ISBN 1-877640-21-2
DOI 10.1287/educ.1053.0019

An Introduction to Revenue Management

Garrett J. van Ryzin
Graduate School of Business, Columbia University, 3022 Broadway, New York, New York 10027,
gjv1@columbia.edu

Kalyan T. Talluri
Department of Economics and Business, Universitat Pompeu Fabra, Jaume I Building,
Ramon Trias Fargas, 25–27, 08005 Barcelona, Spain, kalyan.talluri@upf.edu

Abstract *Revenue management* (RM) refers to the collection of strategies and tactics firms use to scientifically manage demand for their products and services. It has gained attention recently as one of the most successful application areas of operations research (OR). The practice has grown from its origins as a relatively obscure practice among a handful of major airlines in the postderegulation era in the United States (circa 1978) to its status today as a mainstream business practice with a growing list of industry users, ranging from Walt Disney Resorts to National Car Rental. The economic impact of RM is significant, with increases in revenue of 5% or more reported in several industry applications of RM systems. Professional practice and research in the area is also expanding. There are now several major industry RM conferences each year, and published research on the methodology of RM has been growing rapidly. This chapter provides an introduction to this increasingly important subfield of OR. It is based on excerpts from our book *The Theory and Practice of Revenue Management* [75].

Keywords revenue management; dynamic pricing; optimization; demand modeling; demand management; capacity control; tutorial

1. Introduction

Every seller of a product or service faces a number of fundamental decisions. A child selling lemonade outside her house has to decide on which day to have her sale, how much to ask for each cup, and when to drop the price (if at all) as the day rolls on. A homeowner selling a house must decide when to list it, what the asking price should be, which offer to accept, and when to lower the listing price—and by how much—if no offers come in. Anyone who has ever faced such decisions knows the uncertainty involved. You want to sell at a time when market conditions are most favorable, but who knows what the future might hold? You want the price to be right—not so high that you put off potential buyers and not so low that you lose out on potential profits. You would like to know how much buyers value your product, but more often than not you must just guess at this number.

Businesses face even more complex selling decisions. For example, how can a firm segment buyers by providing different conditions and terms of trade that profitably exploit their different buying behavior or willingness to pay? How can a firm design products to prevent cannibalization across segments and channels? Once it segments customers, what prices should it charge each segment? If the firm sells in different channels, should it use the same price in each channel? How should prices be adjusted over time, based on seasonal factors and the observed demand to date for each product? If a product is in short supply, to which segments and channels should it allocate the products? How should a firm manage the pricing and allocation decisions for products that are complements (seats on two connecting airline flights) or substitutes (different car categories for rentals)?

RM is concerned with such *demand-management* decisions[1] and the methodology and systems required to make them. It involves managing the firm's "interface with the market," as it were—with the objective of *increasing revenues*. RM can be thought of as the complement of *supply chain management* (SCM), which addresses the *supply decisions* and processes of a firm with the objective (typically) of *lowering the cost* of production and delivery.

Other roughly synonymous names have been given to the practice over recent years—*yield management* (the traditional airline term), *pricing and revenue management*, *pricing and revenue optimization*, *revenue process optimization*, *demand management*, *demand-chain management* (favored by those who want to create a practice parallel to supply chain management)—each with its own nuances of meaning and positioning. However, we use the more standard term revenue management to refer to the wide range of techniques, decisions, methods, processes, and technologies involved in demand management.

1.1. Demand-Management Decisions

RM addresses three basic categories of demand-management decisions:

- Structural decisions: Which selling format to use (such as posted prices, negotiations, or auctions); which segmentation or differentiation mechanisms to use (if any); which terms of trade to offer (including volume discounts and cancellation or refund options); how to bundle products; and so on.
- Price decisions: How to set posted prices, individual-offer prices, and reserve prices (in auctions); how to price across product categories; how to price over time; how to mark down (discount) over the product lifetime; and so on.
- Quantity decisions: Whether to accept or reject an offer to buy; how to allocate output or capacity to different segments, products, or channels; when to withhold a product from the market and sale at later points in time; and so on.

Which of these decisions is most important in any given business depends on the context. The timescale of the decisions varies as well. Structural decisions about which mechanism to use for selling and how to segment and bundle products are normally strategic decisions taken relatively infrequently. Firms may also have to commit to certain price or quantity decisions, for example, by advertising prices in advance or deploying capacity in advance, which can limit their ability to adjust price or quantities on a tactical level. The ability to adjust quantities may also be a function of the technology of production—the flexibility of the supply process and the costs of reallocating capacity and inventory. For example, the use of capacity controls as a tactic in airlines stems largely from the fact that the different "products" an airline sells (different ticket types sold at different times and under different terms) are all supplied using the same homogeneous seat capacity. This gives airlines tremendous quantity flexibility, so quantity control is a natural tactic in this industry. Retailers, in contrast, often commit to quantities (initial stocking decisions) but have more flexibility to adjust prices over time. The ability to price tactically, however, depends on how costly price changes are, which can vary depending on the channel of distribution, such as online versus catalog.

Whether a firm uses quantity-based or price-based RM controls varies even across firms within a given industry. For instance, while most airlines commit to fixed prices and tactically allocate capacity, low-cost carriers tend to use price as the primary tactical variable.

Broadly speaking, RM addresses all three categories of demand-management decisions—structural, pricing, and quantity decisions. We qualify RM as being either *quantity-based RM* or *price-based RM* if it uses (inventory- or) capacity-allocation decisions or prices as the

[1] These can be referred to as either *sales decisions* (we are making decisions on where and when to sell and to whom and at what price) or *demand-management decisions* (we are estimating demand and its characteristics and using price and capacity control to "manage" demand). We use the latter consistently and use the shorter *demand management* whenever appropriate.

primary tactical tool, respectively, for managing demand. Both the theory and practice of RM differ depending on which control variable is used, and hence we use this dichotomy as necessary.

1.2. What's New About RM?

In one sense, RM is a very old idea. Every seller in human history has faced RM-type decisions. What price to ask? Which offers to accept? When to offer a lower price? And when to simply "pack up one's tent" as it were and try selling at a later point in time or in a different market. In terms of business practice, the problems of RM are as old as business itself.

In terms of theory, at a broad level the problems of RM are not new either. Indeed, the forces of supply and demand and the resulting process of price formation—the "invisible hand" of Adam Smith—lie at the heart of our current understanding of market economics. They are embodied in the concept of the "rational" (profit-maximizing) firm, and define the mechanisms by which market equilibria are reached. Modern economic theory addresses many advanced and subtle demand-management decisions, such as nonlinear pricing, bundling, segmentation, and optimizing in the presence of asymmetric information between buyers and sellers.

What *is* new about RM is not the demand-management decisions themselves, but rather *how* these decisions are made. The true innovation of RM lies in the *method* of decision making—a technologically sophisticated, detailed, and intensely operational approach to making demand-management decisions.

This new approach is driven by two complementary forces. First, scientific advances in economics, statistics, and operations research now make it possible to model demand and economic conditions, quantify the uncertainties faced by decision makers, estimate and forecast market response, and compute optimal solutions to complex decision problems. Second, advances in information technology provide the capability to automate transactions, capture and store vast amounts of data, quickly execute complex algorithms, and then implement and manage highly detailed demand-management decisions. This combination of science and technology applied to age-old demand management is the hallmark of modern RM.

Also, both the science and technology used in RM are quite new. Much of the science used in RM today (demand models, forecasting methods, optimization algorithms) is less than 50 years old, most of the information technology (large databases, personal computers, Internet) is less than 20 years old, and most of the software technology (Java, object-oriented programming) is less than 5 years old. Prior to these scientific developments, it would have been unthinkable to accurately model real-world phenomena and demand-management decisions. Without the information technology, it would be impossible to *operationalize* this science. These two capabilities *combined* make possible an entirely new approach to decision making—one that has profound consequences for demand management.

The first consequence is that science and technology now make it possible to manage demand on a *scale and complexity* that would be unthinkable through manual means (or would require a veritable army of analysts to achieve). A modern large airline, for example, can have thousands of flights a day and provide service between hundreds of thousands of origin-destination pairs, each of which is sold at dozens of prices—and this entire problem is replicated for hundreds of days into the future!

The second consequence of science and technology is that they make it possible to improve the *quality* of demand-management decisions. The management tasks that are involved—quantifying the risks and rewards in making demand-management decisions under uncertainty; working through the often subtle economics of pricing; accurately interpreting market conditions and trends and reacting to this information with timely, accurate, and consistent real-time decisions; optimizing a complex objective function subject to many constraints and business rules—are tasks that most humans, even with many years of experience, are simply not good at.

Of course, even with the best science and technology, there will always be decisions that are better left to human decision makers. Most RM systems recognize this fact and parse the decision-making task, with models and systems handling routine demand-management decisions on an automated basis and human analysts overseeing these decisions and intervening (based on flags or alerts from the system) when extraordinary conditions arise. Such man-machine interaction offers a firm the best of both human and automated decision making.

The process of managing demand decisions with science and technology—implemented with disciplined processes and systems, and overseen by human analysts (a sort of "industrialization" of the entire demand-management process)—defines modern RM.

1.3. The Origins of RM

Where did RM come from? In short, the airline industry. There are few business practices whose origins are so intimately connected to a single industry. Here we briefly review the history of airline RM and then discuss the implications of this history for the field.

The starting point for RM was the Airline Deregulation Act of 1978. With this act, the U.S. Civil Aviation Board (CAB) loosened control of airline prices, which had been strictly regulated based on standardized price and profitability targets. Passage of the act led to rapid change and a rash of innovation in the industry. Established carriers were now free to change prices, schedules, and service without CAB approval. At the same time, new low-cost and charter airlines entered the market. Because of their lower labor costs, simpler (point-to-point) operations, and no-frills service, these new entrants were able to profitably price much lower than the major airlines. They tapped into an entirely new and vast market for discretionary travel—families on a holiday, couples getting away for the weekend, college students visiting home—many of whom might otherwise have driven their cars or not traveled at all.

The potential of this market was embodied in the rapid rise of PeopleExpress, which started in 1981 with cost-efficient operations and fares 50% to 70% lower than the major carriers. By 1984, its revenues were approaching $1 billion, and for the year 1984 it posted a profit of $60 million, its highest profit ever (Cross [21]). These developments resulted in a significant migration of price-sensitive discretionary travelers to the new, low-cost carriers and the cumulative losses in revenue from the shift in traffic were badly damaging the profits of major airlines.

A strategy to recapture the leisure passenger was needed. However, for the majors, a head-to-head, across-the-board price war with the upstarts was deemed almost suicidal. Robert Crandall, American Airline's vice president of marketing at the time, is widely credited with the breakthrough in solving this problem. He recognized that his airline was already producing seats at a marginal cost near zero because most of the costs of a flight (capital costs, wages, fuel) are fixed. As a result, American could in fact afford to compete on cost with the upstarts using its surplus seats.

However, two problems had to be solved to execute this strategy. First, American had to have some way of identifying the surplus seats on each flight. Second, they had to ensure that American's business customers did not switch and buy the new low-price products it offered to discretionary, leisure customers.

American solved these problems using a combination of *purchase restrictions* and *capacity-controlled fares.* First, they designed discounts that had significant restrictions for purchase: They had to be purchased 30 days in advance of departure, were nonrefundable, and required a seven-day minimum stay. These restrictions were designed to prevent most business travelers from utilizing the new low fares. At the same time, American limited the number of discount seats sold on each flight: They *capacity-controlled* the fares. This combination provided the means to compete on price with the upstart airlines without damaging their core business-traveler revenues.

Initially, American's capacity controls were based on setting aside a fixed portion of seats on each flight for the new low-fare products. However, as they gained experience with its Super-Saver fares, American realized that not all flights were the same. Flights on different days and at different times had very different patterns of demand. A more intelligent approach was needed to realize the full potential of capacity-controlled discounts. American therefore embarked on the development of what became known as the *Dynamic Inventory Allocation and Maintenance Optimizer* system (DINAMO). These efforts on DINAMO represent, in many ways, the first large-scale RM system development in the industry.

DINAMO was implemented in full in January 1985 along with a new fare program entitled Ultimate Super-Saver Fares, which matched or undercut the lowest discount fares available in every market American served. DINAMO made all this possible. American could now be much more aggressive on price. It could announce low fares that spanned a large swath of individual flights, confident in its capability to accurately capacity-control the discounts on each individual departure. If a rival airline advertised a special fare in one of American's markets, American could immediately match the offer across the board, knowing that the DINAMO system would carefully control the availability of this fare on the thousands of departures affected by the price change. This feature of pricing aggressively and competitively at an aggregate, market level, while controlling capacity at a tactical, individual-departure level, still characterizes the practice of RM in the airline industry today.

The effect of this new capability was dramatic. PeopleExpress was especially hard-hit as American repeatedly matched or beat their prices in every market it served. PeopleExpress's annual profit fell from an all-time high in 1984 (the year prior to implementation of DINAMO) to a loss of over $160 million by 1986 (one year after DINAMO was implemented). It soon went bankrupt as a result of mounting losses, and in September 1986 the company was sold to Continental Airlines.

Donald Burr, CEO of PeopleExpress, summarized the reasons behind the company's failure (Cross [21]):

> We were a vibrant, profitable company from 1981 to 1985, and then we tipped right over into losing $50 million a month. We were still the same company. What changed was American's ability to do widespread Yield Management in every one of our markets. ... We did a lot of things right. But we didn't get our hands around Yield Management and automation issues. ... [If I were to do it again,] the number one priority on my list every day would be to see that my people got the best information technology tools. In my view, that's what drives airline revenues today more than any other factor—more than service, more than planes, more than routes.

This story was played out in similar fashion throughout the airline industry in the decades following deregulation, and airlines that did not have similar RM capabilities scrambled to get them.

As a result of this history, the practice of RM in the airline industry today is both pervasive and mature, and RM is viewed as critical to running a modern airline profitably. For example, American Airlines estimates that its RM practices generated $1.4 billion in additional incremental revenue over a three-year period starting around 1988 (Smith et al. [69]). Many other carriers also attribute similar improvements in their revenue due to RM.

1.4. Consequences of the Airline History

The intimate connection of RM to the airline industry is both a blessing and a curse for the field of RM. The blessing is that RM can point to a major industry in which the practice of RM is pervasive, highly developed, and enormously effective. Indeed, a large, modern airline today would just not be able to operate profitably *without* RM. By most estimates, the revenue gains from the use of RM systems are roughly comparable to many airlines' total profitability in a good year (about 4% to 5% of revenues).[2] In addition, the scale and complexity of RM at major airlines is truly mind-boggling. Therefore, the airline success story validates both the economic importance of RM and the feasibility of executing it reliably in a complex business environment. This is the good-news story for the field from the airline experience.

The bad news—the curse if you will—of the strong association of RM with airlines is that it has created a certain myopia inside the field. Many practitioners and researchers view RM solely in airline-specific terms, and this has at times tended to create biases that have hampered both research and implementation efforts in other industries.

A second problem with the airline-specific association of RM is that airline pricing has something of a bad reputation among consumers. While on the one hand customers love the very low fares made possible by RM practices, the fact that fares are complex, are available one minute and gone the next, and can be drastically different for two people sitting side by side on the same flight, has led to a certain hostility toward the way airlines price. As a result, managers outside the industry are at times, quite naturally, somewhat reluctant to try RM practices for fear of engendering a similar hostile reaction among their customers. However, the reality is that, in most cases, applying RM does *not* involve radically changing the structure of pricing and sales practices; rather, it is a matter of making more intelligent decisions.

1.5. A Conceptual Framework for RM

So, if airlinelike conditions aren't strictly necessary for RM, then exactly where *does* it apply? A short answer is: in any business where tactical demand management is important and the technology and management culture exists to implement it. More specifically, the following conditions generally favor the application of RM techniques:

- *Customer heterogeneity.* If all customers value a product identically and exhibit similar purchase behavior, there is less potential to exploit variations in willingness to pay, variations in preference for different products, and variations of purchase behavior over time. Therefore, the more heterogeneity in customers, the more potential there is to exploit this heterogeneity strategically and tactically to improve revenues.
- *Demand variability and uncertainty.* The more demand varies over time (due to seasonalities, shocks, and so on) and the more uncertainty one has about future demand, the more difficult the demand-management decisions become. Hence, the potential to make bad decisions rises, and it becomes important to have sophisticated tools to evaluate the resulting complex trade-offs.
- *Production inflexibility.* Joint production constraints and costs complicate the demand-management problem. If a firm can "absorb" variations in *demand* easily and costlessly through variations in *supply*, then the complexity of managing demand diminishes; you just supply enough to meet demand. However, the more inflexible the production—the more delays involved in producing units, the more fixed costs or economies of scale involved in production, the more the switch-over costs, the more capacity constraints—the more difficult or costly it becomes to match demand variations with supply variations. As a result, inflexibility leads to more interaction in the demand management at different points in time, between different segments of customers, across different products of a product line, and across different channels of distribution. The complexity increases and the consequences of poor decisions become more acute. Hence, RM becomes more beneficial.
- *Data and information systems infrastructure.* To operationalize RM requires data to accurately characterize and model demand. It also requires systems to collect and store the

[2] Many skeptics point to Southwest Airlines as a counterexample, but Southwest does use RM systems. However, because its pricing structure is simpler than most other airlines, the use of RM is less obvious to consumers and casual observers.

data and to implement and monitor the resulting real-time decisions. In most industries it is usually feasible—in theory, at least—to collect and store demand data and automate demand decisions. However, attempting to apply RM in industries that do not have databases or transactions systems in place can be a time consuming, expensive, and risky proposition. RM, therefore, tends to be more suited to industries where transaction-processing systems are already employed as part of incumbent business processes.

- *Management culture*. RM is a technically complex and demanding practice. There is a risk, therefore, that a firm's management may simply not have sufficient familiarity with—or confidence in—science and technology to make implementing an RM system a realistic prospect. The culture of the firm may not be receptive to innovation or may value more intuitive approaches to problem solving. This is often due to the culture of the industry and its managers: their educational backgrounds, their professional experiences and responsibilities en route to leadership positions, and the skills required to succeed in the industry.

1.6. Industry Adopters Beyond the Airlines

What do these conditions imply for adopters of RM technology? The production-inflexibility characteristics of airlines are shared by many other service industries, such as hotels, cruise ship lines, car rental companies, theaters and sporting venues, and radio/TV broadcasters, to name a few. Indeed, RM is strongly associated with service industries.

Retailers have recently begun to adopt RM, especially in the fashion apparel, consumer electronics, and toy sectors. Retail demand is highly volatile and uncertain, consumers' valuations change rapidly over time, and with short selling seasons and long production and distribution lead times, supply is quite inflexible. On the technology front, the introduction of bar codes and point-of-sale (POS) technology has resulted in a high degree of automation of sales transactions for most major retailers.

The energy sector has been a recent adopter of RM methods as well, principally in the area of managing the sale of pipeline capacity for gas transportation. Again, energy demands are volatile and uncertain, and the technology for generating and transmitting electricity and gas can be inflexible. Also, thanks to deregulation in the industry, there has been a lot of experimentation and innovation in the pricing practices of energy, gas, and transmission markets.

Manufacturing is potentially a vast market for RM methods, although to date relatively few instances of the practice have been documented. Enterprise resource planning (ERP), supply chain management (SCM), and customer relationship management (CRM) systems are commonplace in the industry, and most manufacturers have huge amounts of data and heavily automated business processes, which could form the foundations for RM. For example, in the auto industry, Ford Motor Corporation recently completed a high-profile implementation of RM technology (Coy [20]).

What about future adopters of RM? Given the criteria outlined above, one can argue that many industries are potential candidates. Almost all businesses must deal with demand variability, uncertainty, and customer heterogeneity. Most are subject to some sort of supply or production inflexibility. Finally, thanks largely to the wave of enterprise software and e-commerce innovation of late, many firms have now automated their business processes. All of these factors bode well for the future of RM.

Nevertheless, as with any technological and business-practice innovation, the case for RM ultimately boils down to a cost-benefit analysis for each individual firm. For some, the potential benefit will simply never justify the costs of implementing RM systems and business processes. However, we believe that for the majority of firms, RM will eventually be justified once the technology and methodology in their industry matures. Indeed, the history of RM in industries such as airlines, hotels, and retail suggests that once the technology gains a foothold in an industry, it spreads quite rapidly. As a result, we would not be surprised to see RM systems (or systems performing RM functions under a different label) become as ubiquitous as ERP, SCM, and CRM systems are today.

1.7. Overview of Topics

The number of topics the field spans is too large to cover adequately in a single chapter like this. Our book, *The Theory and Practice of Revenue Management* (Kluwer 2004) provides in-depth coverage of both quantity- and price-based RM as well as supporting topics such as demand modeling, economics, forecasting, and system implementation. Here, we only give a sample of two such topics: single-resource capacity control and deterministic dynamic pricing problems. While incomplete, this sampling serves to illustrate the types of modeling ideas and solution methods found in revenue management.

2. Single-Resource Capacity Control

In this section, we examine some basic results on the problem of quantity-based revenue management for a single resource; specifically, optimally allocating capacity of a resource to different classes of demand. Two prototypical examples are controlling the sale of different fare classes on a single flight leg of an airline and the sale of hotel rooms for a given date at different rate classes. This is to be contrasted with multiple-resource—or network—problems, in which customers require a bundle of different resources (such as two connecting flights or a sequence of nights at the same hotel). In reality, many quantity-based RM problems are network RM problems, but in practice, they are still frequently solved as a collection of single-resource problems (treating the resources independently). For this reason, it is important to study single-resource RM models. Moreover, single-resource models are useful as building blocks in heuristics for the network case.

We assume that the firm sells its capacity in n distinct classes[3] that require the same resource. In the airline and hotel context, these classes represent different discount levels with differentiated sale conditions and restrictions. In the early parts of this section, we assume that these products appeal to distinct and mutually exclusive segments of the market: The conditions of sale segment the market perfectly into n segments—one for each class. Customers in each segment are eligible for or can afford only the class corresponding to their segment. Later in the section, we look at models that do not assume that customers are perfectly segmented, but instead that they choose among the n classes.

The units of capacity are assumed to be homogeneous, and customers demand a single unit of capacity for the resource. The central problem of this section is how to optimally allocate the capacity of the resource to the various classes. This allocation must be done dynamically as demand materializes and with considerable uncertainty about the quantity or composition of future demand. The remainder of the section focuses on various models and methods for making these capacity-allocation decisions.

2.1. Types of Controls

In the travel industry, reservation systems provide different mechanisms for controlling availability. These mechanisms are usually deeply embedded in the software logic of the reservation system and, as a result, can be quite expensive and difficult to change. Therefore, the control mechanisms chosen for a given implementation are often dictated by the reservation system. Here, we focus on the control mechanisms themselves.

2.1.1. Booking Limits. *Booking limits* are controls that limit the amount of capacity that can be sold to any particular class at a given point in time. For example, a booking limit of 18 on class 2 indicates that at most 18 units of capacity can be sold to customers in class 2. Beyond this limit, the class would be "closed" to additional class 2 customers. This limit of 18 may be less than the physical capacity. For example, we might want to protect capacity for future demand from class 1 customers.

[3] In the case of airlines, these are called *fare classes*. Terms like *rate products*, *rate classes*, *revenue classes*, *booking classes*, and *fare products* are also used. We shall use the generic term *class* in this section.

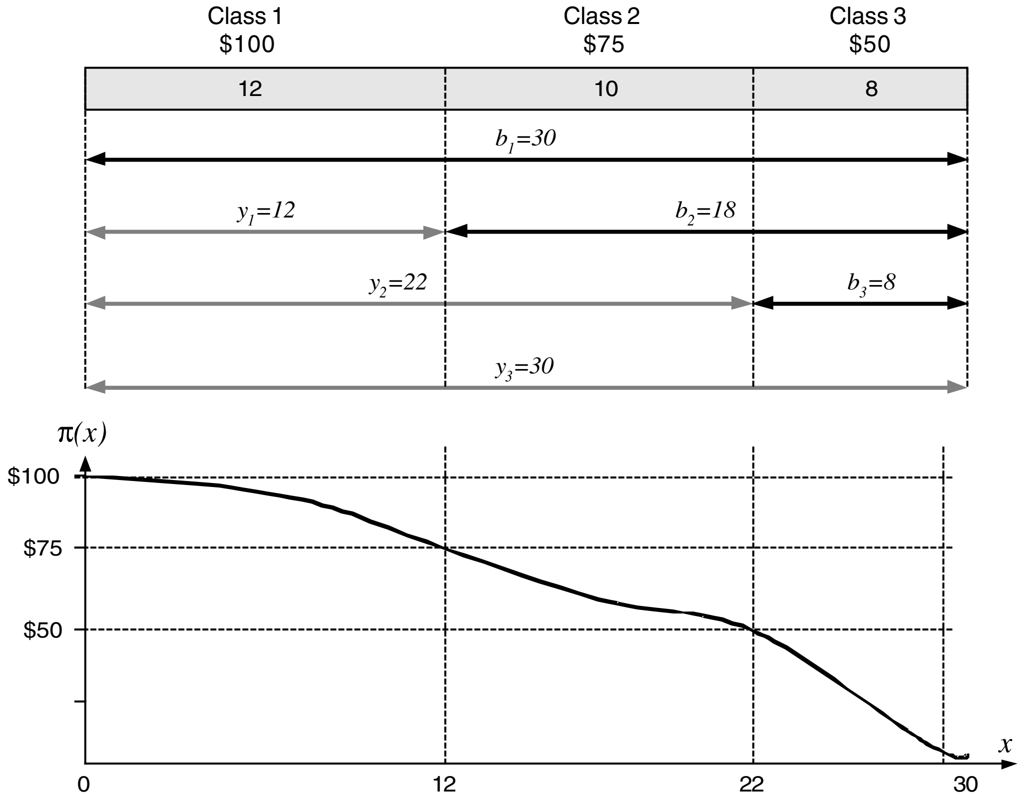
Booking limits are either *partitioned* or *nested*: A *partitioned booking limit* divides the available capacity into separate blocks (or *buckets*)—one for each class—that can be sold only to the designated class. For example, with 30 units to sell, a partitioned booking limit may set a booking limit of 12 units for class 1, 10 units for class 2, and 8 units for class 3. If the 12 units of class 1 capacity are used up, class 1 would be closed regardless of how much capacity is available in the remaining buckets. This could be undesirable if class 1 has higher revenues than do classes 2 and 3 and the units allocated to class 1 are sold out.

With a *nested booking limit*, the capacity available to different classes overlaps in a hierarchical manner—with higher-ranked classes having access to all the capacity reserved for lower-ranked classes (and perhaps more). Let the nested booking limit for class j be denoted b_j. Then b_j is the maximum number of units of capacity we are willing to sell to classes j and lower. So, in Figure 1, the nested booking limit on class 1 and lower (all classes) would be $b_1 = 30$ (the entire capacity), the nested booking limit on classes 2 and 3 combined would be $b_2 = 18$, and the nested booking limit on class 3 alone would be $b_3 = 8$. We would accept at most 30 bookings for classes 1, 2, and 3; at most 18 for classes 2 and 3 combined; and at most 8 for class 3 customers. Effectively, this logic simply allows any capacity "left over" after selling to low classes to become available for sale to higher classes.

Nesting booking limits in this way avoids the problem of capacity being simultaneously unavailable for a high class yet available for lower classes. Most reservations systems that use booking-limit controls quite sensibly use nested rather than partitioned booking limits for this reason.

2.1.2. Protection Levels. A *protection level* specifies an amount of capacity to reserve (protect) for a particular class or set of classes. Again, protection levels can be *nested* or *partitioned*. A partitioned protection level is trivially equivalent to a partitioned booking

FIGURE 1. The relationship between booking limits b_j, protection levels y_j, and bid prices $\pi(x)$.



limit; a booking limit of 18 on class 2 sales is equivalent to protecting 18 units of capacity for class 2.

In the nested case, protection levels are again defined for sets of classes—ordered in a hierarchical manner according to class order. Suppose class 1 is the highest class, class 2 the second highest, and so on. Then the protection level j, denoted y_j, is defined as the amount of capacity to save for classes $j, j-1, \ldots, 1$ combined—that is, for classes j and higher (in terms of class order). Continuing our example, we might set a protection level of 12 for class 1 (meaning 12 units of capacity would be protected for sale only to class 1), a protection level of 22 for classes 1 and 2 combined, and a protection level of 30 for classes 1, 2, and 3 combined (although frequently no protection level is specified for this last case because it is clear that all the capacity is available to at least one of the classes).

Figure 1 shows the relationship between protection levels and booking limits. The booking limit for class j, b_j is simply the capacity minus the protection level for classes $j-1$ and higher. That is,

$$b_j = C - y_{j-1}, \quad j = 2, \ldots, n,$$

where C is the capacity. For convenience, we define $b_1 = C$ (the highest class has a booking limit equal to the capacity) and $y_n = C$ (all classes combined have a protection level equal to capacity).

2.1.3. Standard vs. Theft Nesting.

The standard process for using booking limits or nested protection levels proceeds as follows. Starting with C units of capacity, we begin receiving bookings. A booking for class j is accepted provided (1) there is capacity remaining and (2) the total number of requests accepted for class j to date is less than the booking limit b_j (equivalently, the current capacity remaining is more than the protection level y_{j-1} for classes higher than j). This is called *standard nesting*, and it is the most natural and common way to implement nested-capacity controls.

Another alternative, which is less prevalent although still encountered occasionally in practice, is called *theft nesting*. In theft nesting, a booking in class j not only reduces the allocation for class j, but also "steals" from the allocation of all lower classes. Therefore, when we accept a request for class j, not only is the class j allocation reduced by one, but so are the allocations for classes $j+1, j+2, \ldots, n$. This is equivalent to keeping y_j units of capacity protected for *future* demand from class j and higher. In other words, even though we just accepted a request for class j, under theft nesting we continue to reserve y_j units for class j and higher, and to do so requires reducing the allocation for classes $j+1, j+2, \ldots, n$. Under standard nesting, in contrast, when we accept a request from class j we effectively reduce by one the capacity we protect for future demand from class j and higher.

The rationale for standard nesting is that the capacity protected for, say, class 1 is based on a forecast of future demand for class 1. Once we observe some demand for class 1, we then *reduce* our estimate of future demand—and hence the capacity we protect for class 1. Standard nesting does this by reducing the capacity protected for future class 1 demand on a one-for-one basis after each arriving request is accepted (and similarly for other classes as well). To illustrate, suppose in our example that demand for class 1 is deterministic and equal to the protection level $y_1 = 12$. Then if we receive 5 requests for class 1, we know for certain that future demand for class 1 will be only 7, and hence that it makes sense to reduce the capacity we protect for future demand from 12 to 7, which is precisely what standard nesting does. Theft nesting, in contrast, intuitively corresponds to an assumption of "memorylessness" in demand. In other words, it assumes the demand to date for class 1 does not affect our estimate of *future* demand for class 1. Therefore, we continue to protect y_1 units of capacity for class 1 (and hence must reduce the allocation for classes $2, 3, \ldots, n$).

The two forms of nesting are in fact equivalent if demand arrives strictly in low-to-high class order; that is, the demand for class n arrives first, followed by the demand for class $n-1$, and so on.[4] This is what the standard (static) single-resource models assume, so for these static models, the distinction is not important. However, in practice demand rarely arrives in low-to-high order, and the choice of standard versus theft nesting matters. With mixed order of arrivals, theft nesting protects more capacity for higher classes (equivalently, allocates less capacity to lower classes). Again, however, standard nesting is the norm in RM practice.

2.1.4. Bid Prices. What distinguishes bid-price controls from both booking limits and protection levels is that they are revenue-based, rather than class-based, controls. Specifically, a bid-price control sets a threshold price (which may depend on variables such as the remaining capacity or time), such that a request is accepted if its revenue exceeds the threshold price and rejected if its revenue is less than the threshold price. Bid-price controls are, in principle, simpler than booking-limit or protection-level controls because they require storing only a single threshold value at any point in time—rather than a set of capacity numbers, one for each class. However, to be effective, bid prices must be updated after each sale—and possibly also with time as well—and this typically requires storing a table of bid-price values indexed by the current available capacity, current time, or both.

Figure 1 shows how bid prices can be used to implement the same nested-allocation policy as booking limits and protection levels. The bid price $\pi(x)$ is plotted as a function of the remaining capacity x. When there are 12 or fewer units remaining, the bid price is over \$75 but less than \$100, so only class 1 demand is accepted. With 13 to 22 units remaining, the bid price is over \$50 but less than \$75, so only classes 1 and 2 are accepted. With more than 22 units of capacity available, the bid price drops below \$50, so all three classes are accepted.

Bid-price control is criticized by some as being “unsafe”—the argument being that having a threshold price as the only control means that the RM system will sell an unlimited amount of capacity to any class whose revenues exceed the bid-price threshold. However, this is true only if the bid price is not updated. As shown in Figure 1, if the bid price is a function of the current remaining capacity, then it performs exactly like a booking limit or protection level, closing off capacity to successively higher classes as capacity is consumed. Without this ability to make bid prices a function of capacity, however, a simple static threshold is indeed a somewhat dangerous form of control.

One potential advantage of bid-price controls is their ability to discriminate based on revenue rather than class. Often a number of products with different prices are booked in a single class. RM systems then use an average price as the price associated with a class. However, if actual revenue information is available for each request, then a bid-price control can selectively accept only the higher revenue requests in a class, whereas a control based on class designation alone can only accept or reject all requests of a class. Of course, if the exact revenue is not observable at the time of reservation, then this advantage is lost.

2.2. Displacement Cost

While the mathematics of optimal capacity controls can become complex, the overriding logic is simple. First, capacity should be allocated to a request if and only if its revenue is greater than the value of the capacity required to satisfy it. Second, the value of capacity should be measured by its (expected) *displacement cost*—or *opportunity cost*—which is the expected loss in future revenue from using the capacity now rather than reserving it for future use.

Theoretically, the displacement-cost idea is captured by using a *value function*, $V(x)$, that measures the optimal expected revenue as a function of the remaining capacity x. The displacement cost then is the difference between the value function at x and the value function at $x-1$, or $V(x)-V(x-1)$. Much of the theoretical analysis of the capacity controls boils down to analyzing this value function, but conceptually, the logic is simply to compare revenues to displacement costs to make the accept or deny decision.

[4] It is easy to convince oneself of this fact by tracing out the accept/deny decisions under both forms of nesting, and doing so is an instructive exercise.

2.3. Static Models

In this section, we examine one of the first models for quantity-based RM, the so-called *static*[5] single-resource models.

The static model makes several assumptions that are worth examining in some detail. The first is that demand for the different classes arrives in nonoverlapping intervals in the order of increasing prices of the classes.[6] In reality, demand for the different classes may overlap in time. However, the nonoverlapping-intervals assumption is a reasonable approximation (for example, advance-purchase discount demand typically arrives before full-fare coach demand in the airline case). Moreover, the optimal controls that emerge from the model can be applied—at least heuristically—even where demand comes in arbitrary order (using either bid prices or the nesting policies, for example). As for the strict low-before-high assumption, this represents something of a worst-case scenario; for instance, if high-revenue demand arrives before low-revenue demand, the problem is trivial because we simply accept demand first come, first serve.

The second main assumption is that the demands for different classes are independent random variables. Largely, this assumption is made for analytical convenience, because to deal with dependence in the demand structure would require introducing complex state variables on the history of observed demand. We can make some justification for the assumption by appealing to the forecast inputs to the model. That is, to the extent that there are systematic factors affecting *all* demand classes (such as seasonalities), these are often reflected in the forecast and become part of the *explained* variation in demand in the forecasting model (for example, as the differences in the forecasted means and variance on different days). The randomness in the single-resource model is then only the residual, unexplained variation in demand. So, for example, the fact that demand for all classes may increase on peak flights does not in itself cause problems, provided the increase is predicted by the forecasting method. Still, one has to worry about possible residual dependence in the *unexplained variation* in demand, and this is a potential weakness of the independence assumption.

A third assumption is that demand for a given class does not depend on the capacity controls; in particular, it does not depend on the availability of other classes. Its only justification is if the multiple restrictions associated with each class are so well designed that customers in a high-revenue class will not buy down to a lower class, and if the prices are so well separated that customers in a lower class will not buy up to a higher class if the lower class is closed. However, neither is really true in practice. There is considerable porousness (imperfect segmentation) in the design of the restrictions, and the price differences between the classes are rarely that dispersed. The assumption that demand does not depend on the capacity controls is therefore a weakness, although in §2.5 we look at models that handle imperfect segmentation.

Fourth, the static model suppresses many details about the demand and control process within each of the periods. This creates a potential source of confusion when relating these models to actual RM systems. In particular, the static model assumes that an aggregate quantity of demand arrives in a single stage and the decision is simply how much of this demand to accept. However, in a real reservation system, we typically observe demand sequentially over time, or it may come in batch downloads. The control decision has to be made knowing only the demand observed to date and is usually implemented in the form of prespecified controls uploaded to the reservation system. These details are essentially ignored in the static model. However, fortunately (and perhaps surprisingly), the form of the

[5] The term *static* is somewhat of a misnomer here because demand does arrive sequentially over time, albeit in stages ordered from low-revenue to high-revenue demand. However, this term is now standard and helps distinguish this class of models from *dynamic* models that allow arbitrary arrival orders.

[6] Robinson [67] generalizes the static model to the case where demand from each class arrives in nonoverlapping intervals, but the order is not necessarily from low to high revenue.

optimal control is not sensitive to this assumption and can be applied quite independently of how the demand is realized within a period (all at once, sequentially, or in batches). The simplicity and robustness of the optimal control is, in fact, a central result of the theory for this class of models.

A fifth assumption of the model is that either there are no groups, or if there are group bookings, they can be partially accepted.

Finally, the static models assume risk neutrality. This is a reasonable assumption in practice, because a firm implementing RM typically makes such decisions for a large number of products sold repeatedly (for example, daily flights, daily hotel room stays, and so on). Maximizing the average revenue, therefore, is what matters in the end. While we do not cover this case here, some researchers have recently analyzed the single-resource problem with risk-averse decision makers (Feng and Xiao [30]).

We start with the simple two-class model to build some basic intuition, and then examine the more general n-class case.

2.3.1. Littlewood's Two-Class Model. The earliest single-resource model for quantity-based RM is due to Littlewood [55]. The model assumes two product classes, with associated prices $p_1 > p_2$. The capacity is C, and we assume there are no cancellations or overbooking. Demand for class j is denoted D_j, and its distribution is denoted by $F_j(\cdot)$. Demand for class 2 arrives first. The problem is to decide how much class 2 demand to accept before seeing the realization of class 1 demand.

The two-class problem is similar to the classic *newsboy problem* in inventory theory, and the optimal decision can be derived informally using a simple marginal analysis: Suppose that we have x units of capacity remaining and we receive a request from class 2. If we accept the request, we collect revenues of p_2. If we do not accept it, we will sell unit x (the marginal unit) at p_1 if and only if demand for class 1 is x or higher. That is, if and only if $D_1 \geq x$. Thus, the expected gain from reserving the xth unit for class 1 (the *expected marginal value*) is $p_1 P(D_1 \geq x)$. Therefore, it makes sense to accept a class 2 request as long as its price exceeds this marginal value or, equivalently, if and only if

$$p_2 \geq p_1 P(D_1 \geq x). \tag{1}$$

Note that the right-hand side of (1) is decreasing in x. Therefore, there will be an optimal protection level, denoted y_1^*, such that we accept class 2 if the remaining capacity exceeds y_1^* and reject it if the remaining capacity is y_1^* or less. Formally, y_1^* satisfies

$$p_2 < p_1 P(D_1 \geq y_1^*) \quad \text{and} \quad p_2 \geq p_1 P(D_1 \geq y_1^* + 1).$$

If a continuous distribution $F_1(x)$ is used to model demand (as is often the case), then the optimal protection level y_1^* is given by the simpler expressions

$$p_2 = p_1 P(D_1 > y_1^*), \quad \text{equivalently,} \quad y_1^* = F_1^{-1}\left(1 - \frac{p_2}{p_1}\right), \tag{2}$$

which is known as *Littlewood's rule*. Setting a protection level of y_1^* for class 1 according to Littlewood's rule is an optimal policy. Equivalently, setting a booking limit of $b_2^* = c - y_1^*$ on class 2 demand is optimal. Alternatively, we can use a bid-price control with the bid price set at $\pi(x) = p_1 P(D_1 > x)$.

We omit a rigorous proof of Littlewood's rule because it is a special case of a more general result proved below. However, to gain some insight into it, consider the following example:

Example 1. Suppose D_1 is normally distributed with mean μ and standard deviation σ. Then by Littlewood's rule, $F_1(y_1^*) = 1 - p_2/p_1$, which implies that the optimal protection level can be expressed as

$$y_1^* = \mu + z\sigma,$$

where $z = \Phi^{-1}(1 - p_2/p_1)$ and $\Phi(\cdot)^{-1}$ denotes the inverse of the standard normal c.d.f. Thus, we reserve enough capacity to meet the mean demand for class 1, μ, plus or minus a factor that depends both on the revenue ratio and the demand variation σ. If $p_2/p_1 > 0.5$, the optimal protection level is less than the mean demand; and if $p_2/p_1 < 0.5$, it is greater than the mean demand. In general, the lower the ratio p_2/p_1, the more capacity we reserve for class 1. This makes intuitive sense because we should be willing to take very low prices only when the chances of selling at a high price are lower.

2.3.2. n-Class Models. We next consider the general case of $n > 2$ classes. Again, we assume that demand for the n classes arrives in n stages, one for each class, with classes arriving in increasing order of their revenue values. Let the classes be indexed so that $p_1 > p_2 > \cdots > p_n$. Hence, class n (the lowest price) demand arrives in the first stage (stage n), followed by class $n-1$ demand in stage $n-1$, and so on, with the highest price class (class 1) arriving in the last stage (stage 1). Because there is a one-to-one correspondence between stages and classes, we index both by j. Demand and capacity are most often assumed to be discrete, but occasionally we model them as continuous variables when it helps simplify the analysis and optimality conditions.

Dynamic programming formulation. This problem can be formulated as a dynamic program in the stages (equivalently, classes), with the remaining capacity x being the state variable. At the start of each stage j, the demand $D_j, D_{j-1}, \ldots, D_1$ has not been realized. Within stage j, the model assumes that the following sequence of events occurs:

(1) The realization of the demand D_j occurs, and we observe its value.

(2) We decide on a quantity u of this demand to accept. The amount accepted must be less than the capacity remaining, so $u \le x$. The optimal control u^* is therefore a function of the stage j, the capacity x, and the demand D_j, $u^* = u^*(j, x, D_j)$, although we often suppress this explicit dependence on j, x, and D_j in what follows.

(3) The revenue $p_j u$ is collected, and we proceed to the start of stage $j-1$ with a remaining capacity of $x - u$.

This sequence of events is assumed for analytical convenience; we derive the optimal control u^* "as if" the decision on the amount to accept is made *after* knowing the value of demand D_j. In reality, of course, demand arrives sequentially over time, and the control decision has to be made *before* observing all the demand D_j. However, it turns out that optimal decisions do not use the prior knowledge of D_j as we show below. Hence, the assumption that D_j is known is not restrictive.

Let $V_j(x)$ denote the value function at the start of stage j. Once the value D_j is observed, the value of u is chosen to maximize the current stage j revenue plus the revenue to go, or

$$p_j u + V_{j-1}(x-u),$$

subject to the constraint $0 \le u \le \min\{D_j, x\}$. The value function entering stage j, $V_j(x)$, is then the expected value of this optimization with respect to the demand D_j. Hence, the Bellman equation is[7]

$$V_j(x) = E\left[\max_{0 \le u \le \min\{D_j, x\}} \{p_j u + V_{j-1}(x-u)\}\right], \tag{3}$$

with boundary conditions

$$V_0(x) = 0, \quad x = 0, 1, \ldots, C.$$

The values u^* that maximize the right-hand side of (3) for each j and x form an optimal control policy for this model.

[7] Readers familiar with dynamic programming may notice that this Bellman equation is of the form $E[\max\{\cdot\}]$ and not $\max E[\cdot]$ as in many standard texts. Essentially, however, the $\max E[\cdot]$ form can be recovered by considering the demand D_j to be a state variable along with x. While the two forms can be shown to be equivalent, the $E[\max\{\cdot\}]$ is simpler to work with in many RM problems. In our case, this leads to the modeling assumption that we optimize "as if" we observed D_j.

Optimal policy: Discrete demand and capacity. We first consider the case where demand and capacity are discrete. To analyze the form of the optimal control in this case, define

$$\Delta V_j(x) \equiv V_j(x) - V_j(x-1).$$

$\Delta V_j(x)$ is the *expected marginal value of capacity* at stage j—the expected incremental value of the xth unit of capacity. A key result concerns how these marginal values change with capacity x and the stage j:

Proposition 1. *The marginal values $\Delta V_j(x)$ of the value function $V_j(x)$ defined by* (3) *satisfy $\forall x, j$:*
(i) $\Delta V_j(x+1) \le \Delta V_j(x)$ *and*
(ii) $\Delta V_{j+1}(x) \ge \Delta V_j(x)$.

That is, at a given stage j the marginal value is decreasing in the remaining capacity, and at a given capacity level x the marginal value increases in the number of stages remaining. These two properties are intuitive and greatly simplify the control. To see this, consider the optimization problem at stage $j+1$. From (3) and the definition of $\Delta V_j(x)$, we can write

$$V_{j+1}(x) = V_j(x) + E\left[\max_{0\le u\le \min\{D_{j+1},x\}} \left\{\sum_{z=1}^{u}(p_{j+1} - \Delta V_j(x+1-z)\right\}\right],$$

where we take the summation above to be empty if $u=0$. Because $\Delta V_j(x)$ is decreasing in x by Proposition 1(i), it follows that the terms in the sum $p_{j+1} - \Delta V_j(x+1-z)$ are decreasing in z. Thus, it is optimal to increase u (keep adding terms) until the terms $p_{j+1} - \Delta V_j(x+1-z)$ become negative or the upper bound $\min\{D_{j+1}, x\}$ is reached, whichever comes first.

The resulting optimal control can be expressed in terms of optimal protection levels y_j^* for $j, j-1, \ldots, 1$ (class j and higher in the revenue order) by

$$y_j^* \equiv \max\{x : p_{j+1} < \Delta V_j(x)\}, \quad j = 1, \ldots, n-1. \tag{4}$$

(Recall the optimal protection level $y_n^* \equiv C$ by convention.) The optimal control at stage $j+1$ is then

$$u^*(j+1, x, D_{j+1}) = \min\{(x - y_j^*)^+, D_{j+1}\}, \tag{5}$$

where the notation $z^+ = \max\{0, x\}$ denotes the positive part of z. The quantity $(x - y_j^*)^+$ is the remaining capacity in excess of the protection level, which is the maximum capacity we are willing to sell to class $j+1$. The situation is shown in Figure 2.

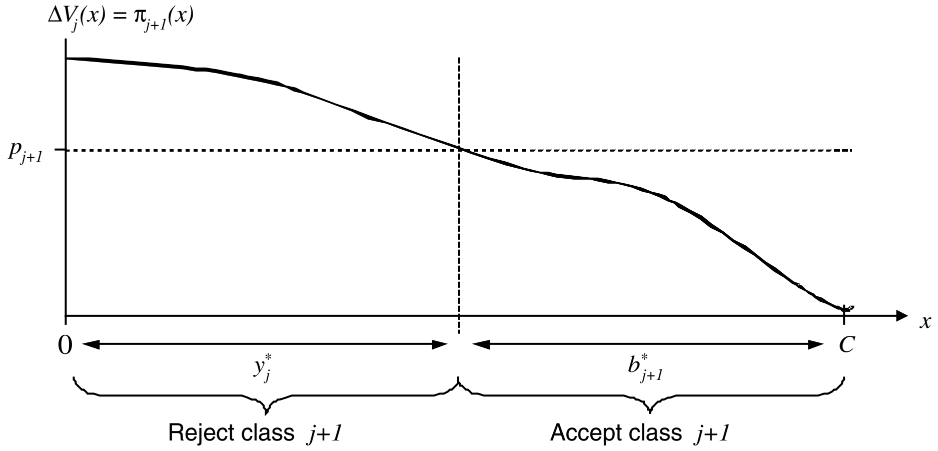
In practice, we can simply post the protection level y_j^* in a reservation system and accept requests first come, first serve until the capacity threshold y_j^* is reached or the stage ends, whichever comes first. Thus, the optimal protection-level control at stage $j+1$ requires no information about the demand D_{j+1}, yet it produces the same optimal decision "as if" we knew D_{j+1} exactly at the start of stage $j+1$. The reason for this is that knowledge of D_{j+1} does not affect the future value of capacity, $V_j(x)$. Deciding to accept or reject each request simply involves comparing current revenues to the marginal cost of capacity, and this comparison does not depend on how many stage-$(j+1)$ requests there are in total.

Proposition 1(ii) implies the nested protection structure

$$y_1^* \le y_2^* \le \cdots \le y_n^*.$$

This fact is easily seen from Figure 2. If p_{j+1} increases with j and the curve $\Delta V_j(x)$ decreases with j, then the optimal protection level y_j^* will shift to the left (decrease). Together, this ordering produces the nested protection-level structure.

Figure 2. The optimal protection level y_j^* in the static model.



One can also use booking limits in place of protection levels to achieve the same control. Optimal nested booking limits are defined by

$$b_j^* \equiv C - y_{j-1}^*, \quad j = 2, \ldots, n, \tag{6}$$

with $b_1^* \equiv C$. The optimal control in stage $j+1$ is then to accept

$$u^*(j+1, x, D_{j+1}) = \min\{(b_{j+1} - (C-x))^+, D_{j+1}\}.$$

Note that $C - x$ is the total capacity sold prior to stage $j+1$ and b_{j+1} is the booking limit for class $j+1$, so $(b_{j+1} - (C-x))^+$ is the remaining capacity available for class $j+1$. The optimal booking limit is also shown in Figure 2.

Finally, the optimal control can also be implemented through a table of bid prices. Indeed, if we define the stage $j+1$ bid price by

$$\pi_{j+1}(x) \equiv \Delta V_j(x), \tag{7}$$

then the optimal control is

$$u^*(j+1, x, D_{j+1}) = \begin{cases} 0 & \text{if } p_{j+1} < \pi_{j+1}(x) \\ \max\{z : p_{j+1} \geq \pi_{j+1}(x-z)\} & \text{otherwise.} \end{cases}$$

In words, we accept the zth request in stage $j+1$ if the price p_{j+1} exceeds the bid price value $\pi_{j+1}(x-z)$ of the zth unit of capacity that is allocated. In practice, we can store a table of bid prices and process requests by sequentially comparing the price of each product to the table values corresponding to the remaining capacity.

We summarize these results in the following theorem:

Theorem 1. *For the static model defined by* (3), *the optimal control can be achieved using either*

(i) *nested protection levels defined by* (4),
(ii) *nested booking limits defined by* (6), *or*
(iii) *bid-price tables defined by* (7).

Optimality conditions for continuous demand. Next, consider the case where capacity is continuous and demand at each stage has a continuous distribution. In this case, the dynamic program is still given by (3); however, D_j, x, and u are now continuous quantities. The analysis of the dynamic program is slightly more complex than it is in the discrete-demand case, but many of the details are quite similar. Hence, we briefly describe only the key differences.

The main change is that the marginal value $\Delta V_j(x)$ is now replaced by the derivative of $V_j(x)$ with respect to x, $\frac{\partial}{\partial x}V_j(x)$. This derivative is still interpreted as the marginal expected value of capacity, and an argument nearly identical to that in the proof of Proposition 1 shows that the marginal value $\frac{\partial}{\partial x}V_j(x)$ is decreasing in x (equivalently, $V_j(x)$ is concave in x).

Therefore, the optimal control in stage $j+1$ is to keep increasing u (keep accepting demand) as long as

$$p_{j+1} \geq \frac{\partial}{\partial x}V_j(x-u)$$

and to stop accepting once this condition is violated or the demand D_{j+1} is exhausted, whichever comes first. Again, this decision rule can be implemented with optimal protection levels, defined by

$$y_j^* \equiv \max\left\{x : p_{j+1} < \frac{\partial}{\partial x}V_j(x)\right\}, \quad j = 1, \ldots, n-1.$$

One of the chief virtues of the continuous model is that it leads to simplified expressions for the optimal vector of protection levels $\mathbf{y}^* = (y_1^*, \ldots, y_n^*)$. We state the basic result without proof (see Brumelle and McGill [17] for a proof).

First, for an arbitrary vector of protection levels $\mathbf{y}$ and vector of demands $\mathbf{D} = (D_1, \ldots, D_n)$, define the following $n-1$ *fill events*

$$B_j(\mathbf{y}, \mathbf{D}) \equiv \{D_1 > y_1, D_1 + D_2 > y_2, \ldots, D_1 + \cdots + D_j > y_j\}, \quad j = 1, \ldots, n-1. \tag{8}$$

$B_j(\mathbf{y}, \mathbf{D})$ is the event that demand to come in stages $1, 2, \ldots, j$ exceeds the corresponding protection levels. A necessary and sufficient condition for $\mathbf{y}^*$ to be an optimal vector of protection levels is that it satisfy the $n-1$ equations

$$P(B_j(\mathbf{y}^*, \mathbf{D})) = \frac{p_{j+1}}{p_1}, \quad j = 1, 2, \ldots, n-1. \tag{9}$$

That is, the jth fill event should occur with probability equal to the ratio of class $(j+1)$ revenue to class 1 revenue. As it should, this reduces to Littlewood's rule (2) in the $n = 2$ case, because $P(B_1(\mathbf{y}^*, \mathbf{D})) = P(D_1 > y_1^*) = p_2/p_1$.

Note that

$$B_j(\mathbf{y}, \mathbf{D}) = B_{j-1}(\mathbf{y}, \mathbf{D}) \cap \{D_1 + \cdots + D_j > y_j\},$$

so the event $B_j(\mathbf{y}, \mathbf{D})$ can occur only if $B_{j-1}(\mathbf{y}, \mathbf{D})$ occurs. Also, if $y_j = y_{j-1}$ then $B_j(\mathbf{y}, \mathbf{D}) = B_{j-1}(\mathbf{y}, \mathbf{D})$. Thus, if $p_j < p_{j-1}$, we must have $y_j^* > y_{j-1}^*$ to satisfy (9). Thus, the optimal protection levels are strictly increasing in j if the revenues are strictly decreasing in j.

2.3.3. Heuristics. As we have seen, computing optimal controls for the static single-resource model is not particularly difficult. Despite this fact, exact optimization models are not widely used in practice. Indeed, most single-resource airline RM systems use one of several heuristics to compute booking limits and protection levels.

There are two main reasons for this state of affairs. The first is simply a case of practice being one step ahead of the underlying theory. As mentioned, in the airline industry the practice of using capacity controls to manage multiple classes quickly gained popularity following deregulation in the mid-1970s. However, this predates the theory of optimal controls by more than a decade. The only known optimal controls in the 1970s were Littlewood's results for the two-class problem. As a result, heuristics were developed for the general n-class problem. During the decade following deregulation, RM software embedded these heuristics, and people grew accustomed to thinking in terms of them. The inertia generated from this early use of the heuristics is one reason for their continued popularity today.

Heuristics are also widely used because they are simpler to code, quicker to run, and generate revenues that in many cases are close to optimal. Indeed, many practitioners in the airline industry simply believe that even the modest effort of computing optimal controls is not worth the benefit they provide in improved revenue performance. Proponents of heuristics argue that the potential improvement from getting better revenue data and improving demand forecasts swamps the gains from using optimal controls—reflecting the philosophy that it is better to be "approximately right" than it is to be "precisely wrong."

While these points are well taken, such criticisms are somewhat misdirected. For starters, using optimal controls does not mean one has to give up on improvements in other areas, such as forecasting. These activities are not mutually exclusive, although an understaffed development group might very well consider refining optimization modules a low-priority task. Still, given the very modest cost of coding and computing optimal controls, the strong objections to the use of optimal controls are often not entirely rational.

Regardless of one's view on the use of heuristics, it is important to understand them. They remain widely used in practice and can also help develop useful intuition.

We next look at the two most popular heuristics: EMSR-a and EMSR-b, both of which are attributed to Belobaba [3, 4, 5]. Both heuristics are based on the n-class, static, single-resource model defined above in §2.3. They differ only in how they approximate the problem. Static model assumptions apply: Classes are indexed so that $p_1 > p_2 > \cdots > p_n$, $F_j(x)$ denotes the c.d.f. of class j demand, and low-revenue demand arrives before high-revenue demand in stages that are indexed by j as well. Moreover, for ease of exposition we assume that capacity and demand are continuous and that the distribution functions $F_j(x)$, $j = 1, \ldots, n$, are continuous as well, although these assumptions are easily relaxed.

EMSR-a. EMSR-a (*expected marginal seat revenue–version a*) is the most widely publicized heuristic for the single-resource problem. Despite this fact, it is less popular in practice than its close cousin, EMSR-b, which surprisingly is not well documented in the literature. Generally, EMSR-b provides better revenue performance, and it is certainly more intuitive, although EMSR-a is important to know just the same.

EMSR-a is based on the idea of adding the protection levels produced by applying Littlewood's rule to successive pairs of classes. Consider stage $j+1$, in which demand of class $j+1$ arrives with price p_{j+1}. We are interested in computing how much capacity to reserve for the remaining classes, $j, j-1, \ldots, 1$; that is, the protection level, y_j, for classes j and higher. To do so, let us consider a single class k among the remaining classes $j, j-1, \ldots, 1$ and compare k and $j+1$ *in isolation.* Considering only these two classes, we would use Littlewood's rule (2) and reserve capacity y_k^{j+1} for class k, where

$$P(D_k > y_k^{j+1}) = \frac{p_{j+1}}{p_k}. \tag{10}$$

Repeating for each future class $k = j, j-1, \ldots, 1$, we could likewise compute how much capacity to reserve for each such class k in isolation. The idea of EMSR-a, then, is simply to add up these individual protection levels to approximate the total protection level y_j for classes j and higher. That is, set the protection level y_j as

$$y_j = \sum_{k=1}^{j} y_k^{j+1}, \tag{11}$$

where y_k^{j+1} is given by (10). One then repeats this same calculation for each stage j.

EMSR-a is certainly simple and has an intuitive appeal. For a short while it was even believed to be optimal, but this notion was quickly dispelled once the published work on optimal controls appeared.

EMSR-b. EMSR-b is again based on an approximation that reduces the problem at each stage to two classes, but in contrast to EMSR-a, the approximation is based on aggregating *demand* rather than aggregating *protection levels*. Specifically, the demand from future classes is aggregated and treated as one class with a revenue equal to the weighted-average revenue.

Consider stage $j+1$, in which we want to determine protection level y_j. Define the aggregated future demand for classes $j, j-1, \ldots, 1$ by

$$S_j = \sum_{k=1}^{j} D_k,$$

and let the weighted-average revenue from classes $1, \ldots, j$, denoted $\bar{p}_j$, be defined by

$$\bar{p}_j = \frac{\sum_{k=1}^{j} p_k E[D_k]}{\sum_{k=1}^{j} E[D_k]}. \tag{12}$$

Then the EMSR-b protection level for class j and higher, y_j, is chosen by Littlewood's rule (2) so that

$$P(S_j > y_j) = \frac{p_{j+1}}{\bar{p}_j}. \tag{13}$$

It is common when using EMSR-b to assume that demand for each class j is independent and normally distributed with mean μ_j and variance σ_j^2, in which case

$$y_j = \mu + z_\alpha \sigma,$$

where $\mu = \sum_{k=1}^{j} \mu_k$ is the mean and $\sigma^2 = \sum_{k=1}^{j} \sigma_k^2$ is the variance of the aggregated demand to come at stage $j+1$ and $z_\alpha = \Phi^{-1}(1 - p_{j+1}/\bar{p}_j)$ (recall $\Phi^{-1}(x)$ is the inverse of the standard normal c.d.f.). Again, one repeats this calculation for each j.

In practice EMSR-b is more popular and generally seems to perform better than EMSR-a, although studies comparing the two have at times shown mixed results. Belobaba [6] reports studies in which EMSR-b is consistently within 0.5% of the optimal revenue, whereas EMSR-a can deviate by nearly 1.5% from the optimal revenue in certain cases, although with mixed order of arrival and frequent reoptimization, he reports that both methods perform well. However, another recent study by Polt [66] using Lufthansa airline data showed more mixed performance, with neither method dominating the other.

2.4. Dynamic Models

Dynamic models relax the assumption that the demand for classes arrives in a strict low-to-high revenue order. Instead, they allow for an arbitrary order of arrival, with the possibility of interspersed arrivals of several classes. While at first this seems like a strict generalization of the static case, the dynamic models require the assumption of Markovian (such as Poisson) arrivals to make them tractable. This puts restrictions on modeling different levels of variability in demand. Indeed, this limitation on the distribution of demand is the main drawback of dynamic models in practice. In addition, dynamic models require an estimate of the pattern of arrivals over time (called the *booking curve*), which may be difficult to estimate in certain applications. Thus, the choice of dynamic versus static models essentially comes down to a choice of which set of approximations is more acceptable and what data are available in any given application.

Other assumptions of the static model are retained. Demand is assumed to be independent between classes and over time and also independent of the capacity controls. The firm is again assumed to be risk neutral. The justifications (or criticisms) for these assumptions are the same as in the static-model case.

2.4.1. Formulation and Structural Properties. In the simplest dynamic model, we have n classes as before, with associated prices $p_1 \geq p_2 \geq \cdots \geq p_n$. There are T total periods and t indexes the periods, with the time index running forward ($t = 1$ is the first period, and $t = T$ is the last period; this is in contrast to the static dynamic program, where the stages run from n to 1 in the dynamic programming recursion). Because there is no longer a one-to-one correspondence between periods and classes, we use separate indices—t for periods and j for classes.

In each period we assume, by a sufficiently fine discretization of time, that at most one arrival occurs.[8] The probability of an arrival of class j in period t is denoted $\lambda_j(t)$. The assumption of at most one arrival per period implies that we must have

$$\sum_{j=1}^{n} \lambda_j(t) \leq 1.$$

In general, the periods need not be of the same duration. For example, early in the booking process when demand is low, we might use a period of several days, whereas during periods of peak booking activity we might use a period of less than an hour. Note also that the arrival probabilities may vary with t, so the mix of classes that arrive may vary over time. In particular, we do not require lower classes to arrive earlier than higher classes.

Dynamic program. As before, let x denote the remaining capacity and $V_t(x)$ denote the value function in period t. Let $R(t)$ be a random variable, with $R(t) = p_j$ if a demand for class j arrives in period t, and $R(t) = 0$ otherwise. Note that $P(R(t) = p_j) = \lambda_j(t)$. Let $u = 1$ if we accept the arrival (if there has been one), and $u = 0$ otherwise. (We suppress the period subscript t of the control as it should be clear from the context.) We want to maximize the sum of current revenue and the revenue to go, or

$$R(t)u + V_{t+1}(x - u).$$

The Bellman equation is therefore

$$\begin{aligned} V_t(x) &= E\Big[\max_{u \in \{0,1\}} \{R(t)u + V_{t+1}(x-u)\}\Big] \\ &= V_{t+1}(x) + E\Big[\max_{u \in \{0,1\}} \{(R(t) - \Delta V_{t+1}(x))u\}\Big], \end{aligned} \tag{14}$$

where $\Delta V_{t+1}(x) = V_{t+1}(x) - V_{t+1}(x-1)$ is the expected marginal value of capacity in period $t+1$. The boundary conditions are[9]

$$V_{T+1}(x) = 0, \quad x = 0, 1, \ldots, C,$$

and

$$V_t(0) = 0, \quad t = 1, \ldots, T.$$

Optimal policy. An immediate consequence of (14) is that if a class j request arrives, so that $R(t) = p_j$, then it is optimal to accept the request if and only if

$$p_j \geq \Delta V_{t+1}(x).$$

Thus, the optimal control can be implemented using a bid-price control where the bid price is equal to the marginal value,

$$\pi_t(x) = \Delta V_t(x). \tag{15}$$

[8] The assumption of one arrival per period can be generalized as shown by Lautenbacher and Stidham [51], but it is a convenient assumption both theoretically and computationally.

[9] The second boundary condition can be eliminated if we use the control constraint $u \in \{0, \min\{1, x\}\}$ instead of $u \in \{0, 1\}$. However, it is simpler conceptually and notationally to use the $x = 0$ boundary conditions instead.

Revenues that exceed this threshold are accepted; those that do not are rejected.

As in the static case, an important property of the value function is that it has decreasing marginal value $\Delta V_t(x) = V_t(x) - V_t(x-1)$.

Proposition 2. *The increments* $\Delta V_t(x)$ *of the value function* $V_t(x)$ *defined by* (14) *satisfy* $\forall x, t$:

(i) $\Delta V_t(x+1) \le \Delta V_t(x)$, *and*
(ii) $\Delta V_{t+1}(x) \le \Delta V_t(x)$.

This theorem is natural and intuitive because one would expect the value of additional capacity at any point in time to have a decreasing marginal benefit and the marginal value at any given remaining capacity x to decrease with time (because as time elapses, there are fewer opportunities to sell the capacity).

As a consequence, the optimization on the right-hand side of (14) can also be implemented as a nested-allocation policy, albeit one that has time-varying protection levels (or booking limits). Specifically, we can define time-dependent optimal protection levels

$$y_j^*(t) = \max\{x : p_{j+1} < \Delta V_{t+1}(x)\}, \quad j = 1, 2, \ldots, n-1 \tag{16}$$

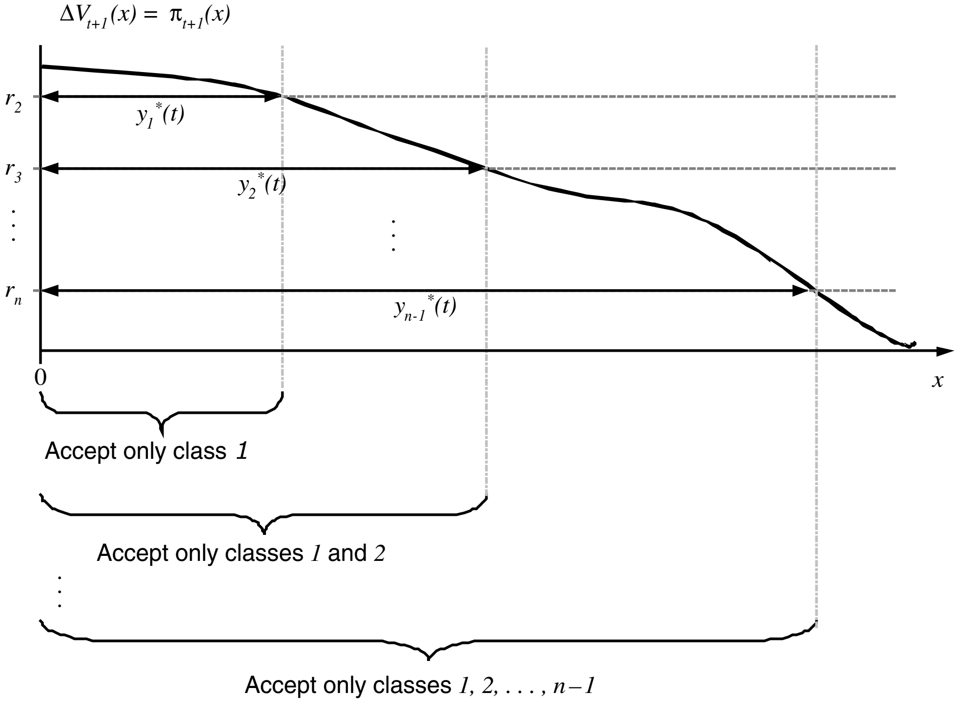
that have the usual interpretation that $y_j^*(t)$ is the capacity we protect for classes $j, j-1, \ldots, 1$. Then the protection levels are nested, $y_1^*(t) \le y_2^*(t) \le \cdots \le y_{n-1}^*(t)$, and it is optimal to accept class j if and only if the remaining capacity exceeds $y_{j-1}^*(t)$. The situation is illustrated in Figure 3.

Time-dependent nested booking limits can also be defined as before by

$$b_j^*(t) \equiv C - y_{j-1}^*(t), \quad j = 2, \ldots, n, \tag{17}$$

That the booking limits and protection levels depend on time in this case essentially stems from the fact that the demand to come varies with time in the dynamic model. The change

FIGURE 3. Optimal protection level $y_j^*(t)$ in the dynamic model.



in demand to come as time evolves affects the opportunity cost, and therefore the resulting booking limit and protection levels.

As a practical matter, because the value function is not likely to change much over short periods of time, fixing the protection levels or booking limits computed by a dynamic model and updating them periodically (as is done in most RM systems in practice) is usually close to optimal. Still, the time-varying nature of the protection levels remains a key distinction between static and dynamic models.

We summarize these results in the following theorem:

Theorem 2. *For the dynamic model defined by* (14)*, the optimal control can be achieved using either:*

(i) *time-dependent nested protection levels defined by* (16)*,*
(ii) *time-dependent nested booking limits defined by* (17)*, or*
(iii) *bid-price tables defined by* (15)*.*

2.5. Customer-Choice Behavior

A key assumption in the models that we have described thus far is that demand for each of the classes is completely independent of the capacity controls being applied by the seller. That is, it is assumed that the likelihood of receiving a request for any given class does not depend on which other classes are available at the time of the request. Needless to say, this is a somewhat unrealistic assumption. For example, in the airline case the likelihood of selling a full-fare ticket may very well depend on whether a discount fare is available at the same time, and the likelihood that a customer buys at all may depend on the lowest available fare. When customers buy a higher fare when a discount is closed it is called *buy-up* (from the firm's point of view, this is also called *sell-up*); when they choose another flight when a discount is closed it is called *diversion*.

Clearly, such customer behavior could have important RM consequences and ought to be considered when making control decisions. We next look at some heuristic and exact methods for incorporating customer-choice behavior in single-resource problems.

2.5.1. Buy-Up Factors. One approach to modeling customer-choice behavior that works with the two-class model is to include buy-up probabilities—also called *buy-up factors*—in the formulation.

The approach works as follows. Consider the simple two-class static model, and recall that Littlewood's rule (2) (slightly restated) is to accept demand from class 2 if and only if

$$p_2 \geq p_1 P(D_1 > x), \tag{18}$$

where x is the remaining capacity—that is, if the revenue from accepting class 2 exceeds the marginal value of the unit of capacity required to satisfy the request. Now suppose that there is a probability q that a customer for class 2 will buy class 1 if class 2 is closed. The net benefit of accepting the request is still the same, but now rather than losing the request when we reject it, there is some chance the customer will buy up to class 1. If so, we earn a net benefit of $p_1 - p_1 P(D_1 > x)$ (the class 1 revenue minus the expected marginal cost). Thus, it is optimal to accept class 2 now if $p_2 - p_1 P(D_1 \geq x) \geq q p_1 (1 - P(D_1 > x))$ or, equivalently, if

$$p_2 \geq (1-q) p_1 P(D_1 > x) + q p_1. \tag{19}$$

Note that the right-hand side of the modified rule (19) is strictly larger than the right-hand side in Littlewood's rule (18), which means that the modified rule (19) is more likely to reject class 2 demand. This is intuitive because with the possibility of customers upgrading to class 1, we should be more eager to close class 2.

The difficulty with this approach is that it does not extend to more than two classes—at least not in an exact way—because the probability that a customer buys class i given that class j is closed depends not only on i and j, but also on which other classes are also available. In other words, with more than two classes the customer faces a *multinomial* choice rather than a *binary* choice.

However, one can at least heuristically extend the buy-up factor idea to EMSR-a or EMSR-b, because these heuristics approximate the multiclass problem using the two-class model.

For example, EMSR-b can be extended to allow for a buy-up factor by modifying the equation for determining the protection level y_j, (13), as follows:

$$p_{j+1} = (1 - q_{j+1})\bar{p}_j P(S_j > y_j) + q_{j+1}\hat{p}_{j+1}, \tag{20}$$

where q_{j+1} is the probability that a customer of class $j+1$ buys up to one of the classes $j, j-1, \ldots, 1$; $\bar{p}_j$ is the weighted-average revenue from these classes as defined by (12); and $\hat{p}_{j+1} > p_{j+1}$ is an estimate of the average revenue received given that a class $j+1$ customer buys up to one of the classes $j, j-1, \ldots, 1$ (for example, $\hat{p}_{j+1} = p_j$ if customers are assumed to buy up to the next-highest price class). Again, the net result of this change is to increase the protection level y_j and close down class $j+1$ earlier than one would do under the traditional EMSR-b rule.[10]

While this modification to EMSR-b provides a simple heuristic way to incorporate choice behavior, it is a somewhat ad hoc adjustment to an already heuristic approach to the problem. Beyond the limitations of the model and its assumptions, there are some serious difficulties involved in estimating the buy-up factors. Indeed, in current applications of the model, they are often simply made-up, reasonable-sounding numbers. Moreover, the assumptions of the model can clash with unconstraining and recapture procedures that are subsequently applied, resulting in double counting of demand. Despite these limitations, buy-up factors have proved useful as a rough-cut approach for incorporating choice behavior in practice.

2.5.2. Discrete-Choice Models.

We next look at a single-resource problem in which customer-choice behavior is explicitly modeled using a general discrete-choice model. In contrast to the heuristic approach of buy-up factors, this model provides a more theoretically sound approach to incorporating choice behavior. It also provides insights into how choice behavior affects the optimal availability controls. The theory is first developed for the general choice model case and then applied to some special demand models.

Model definition. As in the traditional dynamic model of §2.4, time is discrete and indexed by t, with the indices running forward in time ($t = T$ is the period of resource usage). In each period there is at most one arrival. The probability of arrival is denoted by λ, which we assume, for ease of exposition, is the same for all time periods t. There are n classes, and we let $\mathcal{N} = \{1, \ldots, n\}$ denote the entire set of classes. We let choice index 0 denote the no-purchase choice; that is, the event that the customer does not purchase any of the classes offered. Each class $j \in \mathcal{N}$ has an associated price p_j, and without loss of generality we index classes so that $p_1 \geq p_2 \geq \cdots \geq p_n \geq 0$. We let $p_0 = 0$ denote the revenue of the no-purchase choice.

Customer purchase behavior is modeled as follows. In each period t, the seller chooses a subset $S_t \subseteq \mathcal{N}$ of classes to offer. When the set of classes S_t is offered in period t, the probability that a customer chooses class $j \in S_t$ is denoted $P_j(S_t)$. $P_0(S_t)$ denotes the no-purchase probability.

The probability that a sale of class j is made in period t is therefore $\lambda P_j(S_t)$, and the probability that no sale is made is $\lambda P_0(S_t) + (1 - \lambda)$. Note that this last expression reflects the fact that having no sales in a period could be due either to no arrival at all or an

[10] That it increases the protection level about the usual EMSR-b value can be seen by noting that $p_{j+1} = \bar{R}_j P(S_j > y_j)$ in the usual EMSR-b case and $\hat{p}_{j+1} > p_{j+1}$; thus, y_j has to increase to satisfy the equality (20).

TABLE 1. Fare-product revenues and restrictions for Example 2.

Fare product (class)	SA stay	21-day adv.	Revenue
Y	No	No	\$800
M	No	Yes	\$500
K	Yes	Yes	\$450

TABLE 2. Segments and their characteristics for Example 2.

		Qualifies for restrictions?		Willing to buy?	
Segment	Prob.	SA stay	21-day adv.	Y class	M class
Bus. 1	0.1	No	No	Yes	Yes
Bus. 2	0.2	No	Yes	Yes	Yes
Leis. 1	0.2	No	Yes	No	Yes
Leis. 2	0.2	Yes	Yes	No	Yes
Leis. 3	0.3	Yes	Yes	No	No

arrival that does not purchase. This leads to an incomplete-data problem when estimating the model.

The only condition we impose on the choice probabilities $P_j(S)$ is that they define a proper probability function. That is, for every set $S \subseteq \mathcal{N}$, the probabilities satisfy

$$P_j(S) \geq 0, \quad \forall j \in S$$
$$\sum_{j\in S} P_j(S) + P_0(S) = 1.$$

This includes most choice models of practical interest (see Ben-Akiva and Lerman [8]) and even some rather pathological cases.[11] The following running example will be used to illustrate the model and analysis:

Example 2. An airline offers three fare products—Y, M, and K. These products differ in terms of revenues and conditions, as shown in Table 1. The airline has five segments of customers—two business segments and three leisure segments. The segments differ in terms of the restrictions that they qualify for and the fares they are willing to pay. The data describing each segment are given in Table 2. The second column of Table 2 gives the probability that an arriving customer is from each given segment.

Given this data for Example 2, the first four columns of Table 3 give the choice probabilities that would result.[12]

This particular method of generating choice probabilities is only for illustration. Other choice models could be used, and in general any proper set of probabilities could be used to populate Table 3.

Formulation. As before, let C denote the total capacity, T the number of time-periods, t the current period, and x the number of remaining inventory units. Define the value function

[11] For example, some psychologists have shown that customers can be overwhelmed by more choices, and they may become more reluctant to purchase as more options are offered (see Iyengar and Lepper [43]). Such cases would be covered by a suitable choice of $P_j(S)$ that results in the total probability of purchase, $\sum_{j\in S} P_j(S)$, being decreasing in S.

[12] To see how the probabilities in Table 3 are derived, consider the set $S = \{Y, K\}$. If $S = \{Y, K\}$ is offered, segments Business 1 and Business 2 buy the Y fare because they cannot qualify for both the SA stay and 21-day advance-purchase restrictions on K, so $P_Y = 0.1 + 0.2 = 0.3$. Similarly, Leisure 1 cannot qualify for the SA stay restriction of K and is not willing to purchase Y, so these customers do not purchase at all. Segments Leisure 2 and 3, however, qualify for both restrictions on K and purchase K. Hence, $P_K = 0.2 + 0.3 = 0.5$. Class M is not offered, so $P_M = 0$. The other rows of Table 3 are filled out similarly.

Table 3. Choice probabilities $P_j(S)$, probability of purchase $Q(S)$, and expected revenue $R(S)$ for Example 2.

S	$P_Y(S)$	$P_M(S)$	$P_K(S)$	$Q(S)$	$R(S)$	*Efficient?*
$\{Y\}$	0.3	0	0	0.3	240	Yes
$\{M\}$	0	0.4	0	0.4	200	No
$\{K\}$	0	0	0.5	0.5	225	No
$\{Y,M\}$	0.1	0.6	0	0.7	380	No
$\{Y,K\}$	0.3	0	0.5	0.8	465	Yes
$\{M,K\}$	0	0.4	0.5	0.9	425	No
$\{Y,M,K\}$	0.1	0.4	0.5	1	505	Yes

Note. Efficient sets are defined in §2.5.2.

$V_t(x)$ as the maximum expected revenue obtainable from periods $t, t+1, \ldots, T$ given that there are x inventory units remaining at time t. Then the Bellman equation for $V_t(x)$ is

$$\begin{aligned} V_t(x) &= \max_{S\subseteq\mathcal{N}}\left\{\sum_{j\in S}\lambda P_j(S)(p_j + V_{t+1}(x-1)) + (\lambda P_0(S) + 1 - \lambda)V_{t+1}(x)\right\} \\ &= \max_{S\subseteq\mathcal{N}}\left\{\sum_{j\in S}\lambda P_j(S)(p_j - \Delta V_{t+1}(x))\right\} + V_{t+1}(x), \end{aligned} \tag{21}$$

where $\Delta V_{t+1}(x) = V_{t+1}(x) - V_{t+1}(x-1)$ denotes the marginal cost of capacity in the next period, and we have used the fact that for all S,

$$\sum_{j\in S} P_j(S) + P_0(S) = 1.$$

The boundary conditions are

$$V_{T+1}(x) = 0, \quad x = 0, 1, \ldots, C \tag{22}$$

$$V_t(0) = 0, \quad t = 1, \ldots, T. \tag{23}$$

Note one key difference in this formulation compared to our analysis of the traditional independent-class models of §§2.3.2 and 2.4—we assume the seller *precommits* to the open set of classes S in each period, while in the traditional models, we assume the seller observes the class of the request and then makes an accept or deny decision based on the class. The reason for the difference is that in the traditional models the class of an arriving request is completely independent of the controls, so it does not matter whether we precommit to the set of open classes or not. However, in the choice-based model, the class that an arriving customer chooses depends (through the choice model $P_j(S)$) on which classes S we report as being open. Hence, the formulation (21) reflects this fact (we are taking $\max E[\cdot]$ in 21 instead of $E[\max(\cdot)]$); we must choose S prior to seeing the realization of the choice decision.

Structure of the optimal policy. The problem (21) at first seems to have very little structure, but a sequence of simplifications provides a good characterization of the optimal policy. The first simplification is to write (21) in more compact form as

$$V_t(x) = \max_{S\subseteq\mathcal{N}}\{\lambda(R(S) - Q(S)\Delta V_{t+1}(x))\} + V_{t+1}(x), \tag{24}$$

where

$$Q(S) = \sum_{j\in S} P_j(S) = 1 - P_0(S)$$

denotes the total probability of purchase, and

$$R(S) = \sum_{j\in S} P_j(S) p_j$$

denotes the total expected revenue from offering set S. Table 3 gives the values $Q(S)$ and $R(S)$ for our Example 2. For theoretical purposes, we also consider allowing the seller to randomize over the sets S that are offered at the beginning of each time period, but this relaxation is not strictly needed because there is always at least one set S that achieves the maximum in (24).

The second simplification is to note that not all $2^n - 1$ subsets need to be considered when maximizing the right-hand side of (24). Indeed, the search can be reduced to only those sets that are efficient as defined below:

Definition 1. A set T is said to be *inefficient* if there exist probabilities $\alpha(S), \forall S \subseteq \mathcal{N}$ with $\sum_{S\subseteq\mathcal{N}} \alpha(S) = 1$ such that

$$Q(T) \geq \sum_{S\subseteq\mathcal{N}} \alpha(S) Q(S) \quad \text{and} \quad R(T) < \sum_{S\subseteq\mathcal{N}} \alpha(S) R(S).$$

A set is said to be *efficient* if no such probabilities $\alpha(S)$ exist.

In words, a set T is inefficient if we can use a randomization of other sets S to produce an expected revenue that is strictly greater than $R(T)$ with no increase in the probability of purchase $Q(T)$.

The significance of inefficient sets is that they can be eliminated from consideration:

Proposition 3. *An inefficient set is never an optimal solution to* (21).

The proof is omitted, but the fact that such sets should be eliminated from consideration is quite intuitive from (24); an inefficient set T provides strictly less revenue $R(T)$ than do other sets and incurs at least as high a probability of consuming capacity $Q(T)$ (and hence incurs at least as high an opportunity cost $Q(S)\Delta V_{t+1}(x)$ in (24)).

For Example 2, Table 3 shows which sets are efficient—namely, the sets $\{Y\}$, $\{Y,K\}$, and $\{Y,K,M\}$. That these sets are efficient follows from inspection of Figure 4, which shows a scatter plot of the values $Q(S)$ and $R(S)$ for all subsets S. Note from this figure and Definition 1 that an efficient set is a point that is on the "efficient frontier" of the set of points $\{Q(S), R(S)\}, S \subseteq \mathcal{N}$. Here, "efficiency" is with respect to the trade-off between expected revenue $R(S)$ and probability of sale $Q(S)$.

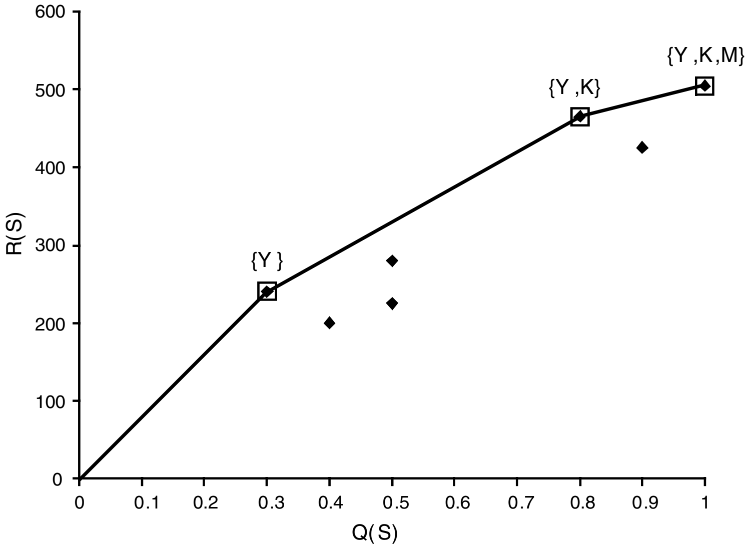
The third simplification is to note that the efficient sets can be easily ordered. Indeed, let m denote the number of efficient sets. These sets can be indexed $S_1, \ldots, S_m$ such that both the revenues and probabilities of purchase are monotone increasing in the index. That is, if the collection of m efficient sets is indexed such that $Q(S_1) \leq Q(S_2) \leq \cdots \leq Q(S_m)$, then $R(S_1) \leq R(S_2) \leq \cdots \leq R(S_m)$ as well. The proof of this fact is again omitted, but it is easy to see intuitively from Figure 4. Note from Table 3 that there are $m = 3$ efficient sets $\{Y\}$, $\{Y,K\}$, and $\{Y,K,M\}$. These can be ordered $S_1 = \{Y\}$, $S_2 = \{Y,K\}$, and $S_3 = \{Y,K,M\}$, with associated probabilities of purchase $Q_1 = 0.3$, $Q_2 = 0.8$, and $Q_3 = 1$ and prices $p_1 =$ \$240, $p_2 =$ \$465, and $p_3 =$ \$505 as claimed.

Henceforth, we assume the efficient sets are denoted $S_1, \ldots, S_m$ and are indexed in increasing revenue and probability order. Also, to simplify notation we let $R_k = R(S_k)$ and $Q_k = Q(S_k)$ and note that R_k and Q_k are both increasing in k. Therefore, the Bellman equation can be further simplified to

$$V_t(x) = \max_{k=1,\ldots,m} \{\lambda(R_k - Q_k \Delta V_{t+1}(x))\} + V_{t+1}(x). \tag{25}$$

The final simplification is to show that when expressed in terms of the sequence $S_1, \ldots, S_m$ of efficient sets, the optimal policy has a simple form as stated in the following theorem:

FIGURE 4. Scatter plot of $Q(S)$ and $R(S)$ for Example 2 (efficient points are enclosed in squares and labeled).



Theorem 3. *An optimal policy for* (21) *is to select a set k^* from among the m efficient, ordered sets $\{S_k : k = 1, \ldots, m\}$ that maximizes* (25)*. Moreover, for a fixed t, the largest optimal index k^* is increasing in the remaining capacity x, and for any fixed x, k^* is increasing in time t.*

The proof of this theorem is involved, but derives from the fact that the marginal value $\Delta V_{t+1}(x)$ is decreasing in x and the fact that the optimal index k^* is decreasing in this marginal value.

This characterization is significant for several reasons. First, it shows that the optimal sets can be reduced to only those that are efficient, which in many cases significantly reduces the number of sets we need to consider. Moreover, it shows that this limited number of sets can be sequenced in a natural way, and that the more capacity we have (or the less time remaining), the higher the set we should use in this sequence.

For example, applying Theorem 3 to Example 2, we see that the efficient sets $S_1 = \{Y\}$, $S_2 = \{Y, K\}$, and $S_3 = \{Y, K, M\}$ would be used as follows. With very large amounts of capacity remaining, S_3 is optimal: All three fare classes are opened. As capacity is consumed, at some point we switch to only offering S_2: Class M is closed, and only Y and K are offered. As capacity is reduced further, at some point we close class K and offer only class Y (set S_1 is used).

Note what is odd here; it can be optimal to offer the highest fare Y and the lowest fare K, but not the middle fare M. This is because opening M causes some buy-down from Y to M, whereas K is sufficiently restricted to prevent buy-down. Only when capacity is plentiful is M opened.

Optimality of nested-allocation policies. The optimization results above also have important implications for the optimality of nested-allocation policies. Indeed, Definition 1 and Theorem 3 can be used to provide a complete characterization of cases in which nested-allocation policies are optimal. They also can be used to provide conditions under which the optimal nesting is by revenue order.

We begin with a precise definition of a nested-allocation policy in the context of the choice model:

Definition 2. A control policy is called a *nested policy* if there is an increasing family of subsets $S_1 \subseteq S_2 \subseteq \cdots \subseteq S_m$ and an index $k_t(x)$ that is increasing in x, such that set $S_{k_t(x)}$ is chosen at time t when the remaining capacity is x.

Although this is a somewhat abstract definition of a nested policy, it is in fact a natural generalization of nested allocations from the traditional single-resource models of §§2.3.2 and 2.4 and implies an ordering of the classes based on when they first appear in the increasing sequence of sets S_k. That is, class i is considered "higher" than class j in the nesting order if class i appears earlier in the sequence. Returning to Example 2, we see that the efficient sets are indeed nested according to this definition because $S_1 = \{Y\}$, $S_2 = \{Y, K\}$, and $S_3 = \{Y, K, M\}$ are increasing. Class Y would be considered the highest in the nested order, followed by class K and then class M.

If the optimal policy is nested in this sense, then we can define optimal protection levels $y_k^*(t), k = 1, \ldots, m$, such that classes lower in the nesting order than those in S_k are closed if the remaining capacity is less than $y_k^*(t)$, just as in the traditional single-resource case. The optimal protection levels for $k = 1, 2, \ldots, m-1$ are defined by

$$y_k^*(t) = \max\{x : R_k - Q_k \Delta V_{t+1}(x) > R_{k+1} - Q_{k+1} \Delta V_{t+1}(x)\}.$$

Nested booking limits can also be defined in the usual way, $b_k(t) = C - y_{k-1}(t)$.

We again return to Example 2 to illustrate this concept. Table 4 shows the objective function value $R_k - Q_k \Delta V_{t+1}(x)$ for each of the three efficient sets $k = 1, 2, 3$, for a particular marginal value function $\Delta V_{t+1}(x)$, which we assume is given in this example. Capacities are in the range $x = 1, 2, \ldots, 20$. The last column of Table 4 gives the index, $k_t^*(x)$, of the efficient set that is optimal for each capacity x.

Note that for capacities $1, 2$, and 3, the set $S_1 = \{Y\}$ is the optimal set, so class Y is the only open fare. Once we reach four units of remaining capacity, set $S_2 = \{Y, K\}$ becomes optimal and we open class K in addition to class Y. When the remaining capacity reaches 13, set $S_3 = \{Y, K, M\}$ becomes optimal, and we open M in addition to Y and K. As a result, the optimal protection level for set S_1, is $y_1^* = 3$, and the protection level for set S_2 is $y_2^* = 12$. S_3 has a protection level equal to capacity.

Table 4. Illustration of nested policy for Example 2.

		$R_k - Q_k \Delta V_{t+1}(x)$			
x	$\Delta V_{t+1}(x)$	$k=1$	$k=2$	$k=3$	$k_t^*(x)$
1	780.00	6.00	−159.00	−275.00	1
2	624.00	52.80	−34.20	−119.00	1
3	520.00	84.00	49.00	−15.00	1
4	445.71	106.29	108.43	59.29	2
5	390.00	123.00	153.00	115.00	2
6	346.67	136.00	187.67	158.33	2
7	312.00	146.40	215.40	193.00	2
8	283.64	154.91	238.09	221.36	2
9	260.00	162.00	257.00	245.00	2
10	240.00	168.00	273.00	265.00	2
11	222.86	173.14	286.71	282.14	2
12	208.00	177.60	298.60	297.00	2
13	195.00	181.50	309.00	310.00	3
14	183.53	184.94	318.18	321.47	3
15	173.33	188.00	326.33	331.67	3
16	164.21	190.74	333.63	340.79	3
17	156.00	193.20	340.20	349.00	3
18	148.57	195.43	346.14	356.43	3
19	141.82	197.45	351.55	363.18	3
20	135.65	199.30	356.48	369.35	3

3. Dynamic Pricing

In this section, we look at settings in which prices rather than quantity controls are the primary variables used to manage demand. While the distinction between quantity and price controls is not always sharp (for instance, closing the availability of a discount class can be considered equivalent to raising the product's price to that of the next-highest class), the techniques we look at here are distinguished by their explicit use of price as the control variable and their explicit modeling of demand as a price-dependent process.

In terms of business practice, varying prices is often the most natural mechanism for revenue management. In most retail and industrial trades, firms use various forms of dynamic pricing—including personalized pricing, markdowns, display and trade promotions, coupons, discounts, clearance sales, and auctions and price negotiations (request for proposals and request for quotes—RFP/RFQ processes) to respond to market fluctuations and uncertainty in demand. Exactly how to make such price adjustments in a way that maximizes revenues (or profits, in the case where variable costs are involved) is the subject of this section.

Dynamic pricing is as old as commerce itself. Firms and individuals have always resorted to price adjustments (such as haggling at the bazaar) in an effort to sell their goods at a price that is as high as possible, yet acceptable to customers. However, the last decade has witnessed an increased application of scientific methods and software systems for dynamic pricing, both in the estimation of demand functions and the optimization of pricing decisions.

3.1. Price-Based vs. Quantity-Based RM

Some industries use price-based RM (retailing), whereas others use quantity-based RM (airlines). Even in the same industry, firms may use a mixture of price- and quantity-based RM. For instance, many of the RM practices of the new low-cost airlines more closely resemble dynamic pricing than the quantity-based RM of the traditional carriers. What explains these differences?

It is hard to give a definitive answer, but in essence it boils down to a question of the extent to which a firm is able to vary quantity or price in response to changes in market conditions. This ability, in turn, is determined by the commitments a firm makes (to price or quantity), its level of flexibility in supplying products or services, and the costs of making quantity or price changes.

Consider airlines, for example. While arguably less true today than in the past, airlines normally commit to prices for their various fare products in advance of taking bookings. This is due to advertising constraints (such as the desire to publish fares in print media and fare tariff books), distribution constraints, and a desire to simplify the task of managing prices. For these marketing and administrative reasons, most airlines advertise and price fare products on an aggregate origin-destination market level for a number of flights over a given interval of time, and do not price on a departure-by-departure basis. This limits their ability to use price to manage the demand on any given departure, demand that varies considerably by flight and is quite uncertain at the time of the price posting. At the same time, the supply of the various classes is almost perfectly flexible between the products (subject to the capacity constraint of the flight) because all fare products sold in the same cabin of service share a homogeneous seat capacity. It is this combination of price commitments together with flexibility on the supply side that make quantity-based RM an attractive tactic in the airline industry. Hotels, cruise ships, and rental cars—other common quantity-based RM industries—share many of these same attributes.

In other cases, however, firms have more price flexibility than quantity flexibility. In apparel retailing, for example, firms commit to order quantities well in advance of a sales season—and may even commit to certain stocking levels in each store. Often, it is impossible (or very costly) to reorder stock or reallocate inventory from one store to another. At the same time, it is easier (though not costless) for most retailers to change prices, as this may require only changing signage and making data entries into a point-of-sale system. Online retailers in particular enjoy tremendous price flexibility because changing prices is almost costless. Business-to-business sales are often conducted through a RFP/RFQ process, which allows firms to determine prices on a transaction-by-transaction basis. In all these situations, price-based RM is therefore a more natural practice. Of course, the context could dictate a different choice even in these industries. For example, if a retailer commits to advertised prices in different regional markets yet retains a centralized stock of products, it might then choose to manage demand by tactically allocating its supply to these different regions—a quantity-based RM approach.

However, given the choice between price- and quantity-based RM, one can argue that price-based RM is the preferred option. The argument is as follows (see Gallego and van Ryzin [35]). Quantity-based RM operates by rationing the quantity sold to different products or to different segments of customers. However, rationing, by its very nature, involves reducing sales by *limiting* supply. If one has price flexibility, however, rather than reducing sales by *limiting supply*, we can reduce sales by *increasing price*. This achieves the same quantity-reducing function as rationing, but does it more profitably because by increasing price we both reduce sales *and* increase revenue at the same time. In short, price-based "rationing" is simply a more profitable way to limit sales than quantity-based rationing.

In practice, of course, firms rarely have the luxury of choosing price and quantity flexibility. Therefore, practical business constraints dictate which tactical response—price- or quantity-based RM (or a mixture of both)—is most appropriate in any given business context.

3.2. Industry Overview

To give a sense of the scope of activity in the area of dynamic pricing, we next review pricing innovations in a few industries.

3.2.1. Retailing. Retailers, especially in apparel and other seasonal-goods sectors, have been at the forefront in deploying science-based software for pricing, driven primarily by the importance of pricing decisions to retailers' profits. For example, Kmart alone wrote off $400 million due to markdowns in one quarter of 2001, resulting in a 40% decline in its net income (Friend and Walker [33]).

Several software firms specializing in RM in retailing have recently emerged. Most of this software is currently oriented toward optimizing markdown decisions. Demand models fit to historical point-of-sale data, together with data on available inventory, serve as inputs to optimization models that recommend the timing and magnitude of markdown decisions.

Major retailers—including Gymboree, J. C. Penney, L. L. Bean, Liz Claiborne, Safeway, ShopKo, and Walgreen's—are experimenting with this new generation of software (Friend and Walker [33], Girard [38], Johnson [45], and Merrick [59]). Many have reported significant improvements in revenue from using pricing models and software. For example, ShopKo reported a 24% improvement in gross margins as a result of using its model-based pricing software (Johnson [45]), and other retailers report gains in gross margins of 5% to 15% (Friend and Walker [33]). Academic studies based on retail data have also documented significant improvements in revenues using model-based markdown recommendations (Bitran et al. [15] and Heching et al. [40]).

3.2.2. Manufacturing. Scientific approaches to pricing are gaining acceptance in the manufacturing sector as well. For example, Ford Motor Co. reported a high-profile implementation of pricing-software technology to support pricing and discounts for its products (Coy [20]). The project, which started in 1995, focused on identifying features that customers were most willing to pay for and changing salesforce incentives to focus on profit margins rather than unit-sale volumes. Ford then applied pricing models developed by an outside consulting firm to optimize prices and dealer and customer incentives across its various product lines. In 1998, Ford reported that the first five U.S. sales regions using this new pricing approach collectively beat their profit targets by $1 billion, while the 13 that used their old methods fell short of their targets by about $250 million (Coy [20]).

3.2.3. E-business. E-commerce has also had a strong influence on the practice of pricing (van Ryzin [79]). Companies such as eBay and Priceline have demonstrated the viability of using innovative pricing mechanisms that leverage the capabilities of the Internet. E-tailers can discount and mark down on the fly based on customer loyalty and click-stream behavior. Because a large e-tailer like Amazon.com has to make a large number of such pricing decisions based on real-time information, automating decision making is a natural priority. The success of these e-commerce companies—inconsistent and volatile as it may appear at times—is at least partly responsible for the increased interest among traditional retailers in using more innovative approaches to pricing.

On the industrial side, e-commerce pricing has been influenced by the growth of business-to-business (B2B) exchanges and other innovations in using the Internet to gain trading efficiencies. While this sector too has had its ebbs and flows, it has produced an astounding variety of new pricing and trading mechanisms, some of which are used regularly for the sale products such as raw materials, generic commodity items, and excess inventory. For example, Freemarkets has had significant success in providing software and service for industrial-procurement auctions, and as of this writing claims to have facilitated over $30 billion in trade since its founding in 1999. Covisint—an exchange jointly funded by Daimler-Chrysler, Ford Motor Company, and General Motors—while slow to develop, looks nevertheless to become a permanent feature of the auto-industry procurement market. Most infrastructure software for B2B exchanges—sold by firms such as Ariba, i2, IBM, and Commerce One—also has various forms of dynamic pricing capabilities built in.

For all these reasons, e-commerce has given price-based RM a significant boost in recent years.

3.3. Examples of Dynamic Pricing

We next examine three specific examples of dynamic pricing and the qualitative factors driving price changes in each case.

3.3.1. Style-Goods Markdown Pricing. Retailers of style and seasonal goods use markdown pricing to clear excess inventory before the end of the season. This type of price-based RM is most prevalent in apparel, sporting goods, high-tech, and perishable-foods retailing. The main incentive for price reductions in such cases is that goods perish or have low salvage values once the sales season is over; hence, firms have an incentive to sell inventory while they can, even at a low price, rather than salvage it.

However, apart from inventory considerations, there are other proposed explanations for markdown pricing. One explanation, proposed by Lazear [52] and investigated empirically in Pashigan [63] and Pashigan and Bowen [64], is that retailers are uncertain about which products will be popular with customers. Therefore, firms set high prices for all items initially. Products that are popular are the ones for which customers have high reservation prices, so these sell out at the high initial price. The firm then identifies the remaining items as low-reservation-price products and marks them down. In this explanation, markdown pricing serves as a form of demand learning.

A second explanation for markdowns is that customers who purchase early have higher willingness to pay, either because they can use the product for a full season (a bathing suit at the start of summer) or because there is some cache to being the first to own it (a new dress style or electronic gadget). Markdown pricing then serves as a segmentation mechanism to separate price-insensitive customers from those price-sensitive customers willing to defer consumption to get a lower price.

Warner and Barsky [80] give yet another explanation, with empirical evidence, for markdown pricing. On holidays and during peak shopping periods (such as before Christmas), customers can search for the lowest prices more efficiently because they are actively engaged in the search, making many shopping trips over a concentrated period of time. Even those customers who normally do not spend much time searching for the best price change their behavior during these peak shopping periods and become more vigilant. The result is that demand during peak periods is more price sensitive and retailers respond by running "sales" during these periods.

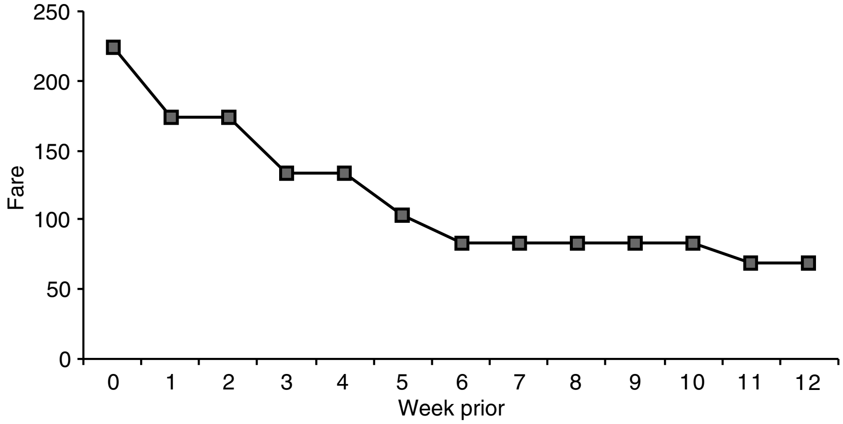
3.3.2. Discount Airline Pricing. Not all dynamic pricing involves price reductions, however. As we mentioned earlier, discount airlines use primarily price-based RM, but with prices often going up over time. These airlines (some examples are easyJet and Ryanair in Europe and jetBlue in the United States) typically offer only one type of ticket on each flight, a nonrefundable, one-way fare without advance-purchase restrictions. However, they offer these tickets at different prices for different flights, and moreover, during the booking period for each flight, vary prices dynamically based on capacity and demand for that specific departure. To quote from one practitioner of this type of dynamic pricing (Easyjet website, 2003):

> The way we structure our fares is based on supply and demand, and prices usually increase as seats are sold on every flight. So, generally speaking, the earlier you book, the cheaper the fare will be. Sometimes, however, due to market forces our fares may be reduced further. Our booking system continually reviews bookings for all future flights and tries to predict how popular each flight is likely to be.

Figure 5 shows the evolution of prices for a particular European discount airline flight as a function of the number of weeks prior to departure. Note that prices are highest in the last few weeks prior to departure.

There are some fundamental differences between air travel and style- and seasonal-goods products that explain this increasing price pattern. For one, the value of air travel to customers does not necessarily go down as the deadline approaches. Conversely, the value of a ticket earlier on is lower for customers as customers multiply the value by the probability that they will indeed use the ticket (especially for a nonrefundable ticket). Somewhat related to these points, additionally, although customers purchase tickets at different points of time, all customers consume the product (fly the flight) at the same time. Therefore, two factors come into play. Customers who purchased early may get upset to see prices drop while they are still holding a reservation; indeed, many airlines give a price guarantee to refund the

FIGURE 5. Prices as a function of weeks prior to departure at a European low-cost discount air carrier.



difference if there is a price drop (to encourage passengers to book early), making it costly for the firms to lower prices. Also, in the travel business, high-valuation high-uncertainty customers tend to purchase closer to the time of service. Hence, demand is less price sensitive close to the time of service.

3.3.3. Consumer Packaged-Goods Promotions. In contrast to markdown and discount airline pricing, promotions are short-run, temporary price reductions. Promotions are the most common form of price-based RM in the consumer packaged-goods (CPG) industry (soap, diapers, coffee, yogurt, and so on).

The fact that customers purchase CPG products repeatedly has important implications for pricing and promotions. Specifically, customers are aware of past prices and past promotions, so running promotions too frequently may condition customers to view the brand as a frequently discounted product, cutting into brand equity in the long run. Because customers are aware of past prices, promotions impact their subjective "reference price"—or sense of the "fair" price—for products. Also, customers may *stockpile* products, so short-run increases in demand due to promotions may come at the expense of reduced future demand.

The institutional structure of promotions is also more complicated. There are three parties involved—manufacturers, retailers, and end customers. Promotions are run either by a manufacturer as discounts to retailers (trade promotions), which may or may not be passed on to the customers by the retailers (retailer pass-thru), or by retailers (retail promotions or consumer promotions). In some forms of promotion (e.g., mail-in coupons) manufacturers give a discount directly to the end customer.

The motivations of the manufacturer and the retailer are different as well. While a manufacturer is interested in increasing sales or profits for its brand, retailers are interested in overall sales or profits for a category constituting multiple brands from multiple manufacturers. For a retailer, discounting a particular brand may increase sales for that brand but dilute overall category profits as customers switch from high-margin brands to the discounted brand. Therefore, in designing optimal promotions structures, one has to consider complex incentive compatibility constraints.

3.4. Modeling Dynamic Price-Sensitive Demand

Any dynamic-pricing model requires a model of how demand—either individual or aggregate—responds to changes in price. The basic theory of consumer choice and the resulting market response models from economics and marketing are used here. However, in dynamic-pricing problems some additional factors must be considered. The first concerns how individual customers behave over time—what factors influence their purchase decisions and how sophisticated their decision-making process is, and so on. The second concerns the state of market conditions—specifically the level of competition and the size of the customer population. We next look at each of these assumptions qualitatively.

3.4.1. Myopic-Customer vs. Strategic-Customer Models. One important demand-modeling assumption concerns the level of sophistication of customers. Most of the models we consider in this section assume *myopic customers*—those who buy as soon as the offered price is less than their willingness to pay. Myopic customers do not adopt complex buying strategies, such as refusing to buy in the hope of lower prices in the future. They simply buy the first time the price drops below their willingness to pay. Models that incorporate *strategic customers*, in contrast, allow for the fact that customers will optimize their own purchase behavior in response to the pricing strategies of the firms.

Of course, the strategic-customer model is more realistic. However, such a demand model makes the pricing problem essentially a strategic game between the customers and the firm, and this significantly complicates the estimation and analysis of optimal pricing strategies—often making the problem intractable. In contrast, the myopic-customer model is much more tractable, and hence is more widely used. The issue in practice is really a matter of how "bad" the myopic assumption is in any given context. In many situations, customers are sufficiently spontaneous in making decisions that one can ignore their strategic behavior. Moreover, customers often do not have sufficient time or information to behave very strategically. However, the more expensive and durable the purchase, the more important it becomes to model strategic-customer behavior (for example, automobile buyers waiting to purchase at the end of a model year).

One common defense of the myopic assumption is the following. The forecasting models that use observations of past customer behavior in a sense reflect the effects of our customers' strategic behavior. For example, if the customers who are most price sensitive tend to adopt a strategy of postponing their purchases until end-of-season clearance sales, then the estimated price sensitivity in these later periods will tend to appear much higher than in earlier periods. Therefore, even though we do not model the strategic behavior directly, our forecasting models indirectly capture the correct price response.

This view is plausible if the pricing strategies obtained from a model are roughly similar to past policies, so that they can be viewed as "perturbations" or "fine-tuning" of a historical pricing strategy—a strategy that customers have already factored into their behavior. On the other hand, if optimized pricing recommendations are radically different in structure from past pricing strategies, then it is reasonable to expect that customers will adjust their buying strategies in response. If this happens, the predictions of myopic models that are fit to historical data may be very bad indeed.

Yet even when the myopic approach works (in the sense of correctly predicting price responses), it runs the risk of reinforcing "bad equilibrium" pricing strategies. For example, a myopic model fit to past data may reconfirm the "optimality" of lowering prices significantly at the end of a sales season or running periodic holiday sales because it estimates, based on historical data, that demand is especially price sensitive in these periods. However, this price sensitivity may be due to the fact that customers have learned not to buy at other times because they know prices will be cut at the end of the season or during holidays. If the firm was to adopt a constant price strategy—and customers were convinced that the firm was sticking to this strategy—then the observed price sensitivity might shift. The resulting equilibrium might be more profitable, but it is one that the firm would not discover using a myopic-customer model.

Despite these limitations and potential pitfalls of the myopic model, it is practical, is widely used, and provides useful insight into dynamic pricing. We therefore focus on the myopic case for the most part in this section. However, we consider strategic customers in §3.7.2.

3.4.2. Infinite-Population vs. Finite-Population Models. Another important assumption in demand modeling is whether the population of potential customers is finite or infinite. Of course, in reality, every population of customers is finite; the question is really a matter of whether the number and type of customers that have already bought changes one's estimate of the number or type of future customers.

In an infinite-population model, we assume that we are sampling *with replacement* when observing customers. As a result, the distribution of the number of customers and the distribution of their willingness to pay is not affected by the past history of observed demand. This is often termed the *nondurable-goods assumption* in economics because we can view this as a case where customers immediately consume their purchase and then reenter the population of potential customers (say, for a can of Coke). This assumption is convenient analytically because one does not need to retain the history of demand (or a suitable sufficient statistic) as a state variable in a pricing-optimization problem.

The finite-population model assumes a random process *without replacement*. That is, there is a finite (possibly random) number of customers with heterogeneous willingness-to-pay values. If one of the customers in the population purchases, the customer is removed from the population of potential customers, and therefore future purchases only occur from the remaining customers. This is termed the *durable-goods assumption* in economics because we can consider it as a case where the good being purchased is consumed over a long period of time (for example, an automobile), and hence once a customer purchases, he effectively removes himself from the population of potential customers.

For example, suppose we assume that a price $p(t)$ is offered in period t and all customers who value the item at more than $p(t)$ purchase in period t (myopic behavior). Then, under a finite-population model, we know that after period t, the remaining customers all have valuations less than $p(t)$. In particular, the future distribution of willingness to pay is conditioned on the values being less than $p(t)$. As a result, in formulating a dynamic-pricing problem, we have to keep track of past pricing decisions and their effect on the residual population of customers.

Which of these models is most appropriate depends on the context. While often the infinite-population model is used simply because it is easier to deal with analytically, the key factors in choosing one model over the other are the number of potential customers relative to the number that actually buy and the type of good (durable versus nondurable). Specifically, the infinite-population model is a reasonable approximation when there is a large population of potential customers and the firm's demand represents a relatively small fraction of this population, because in such cases the impact of the firm's past sales on the number of customers and the distribution of their valuations is negligible. It is also reasonable for consumable goods. However, if the firm's demand represents a large fraction of the potential pool of customers or if the product is a durable good, then past sales will have a more significant impact on the statistics of future demand, and the finite-population assumption is more appropriate.

Qualitatively, the two models lead to quite different pricing policies. Most notably, finite-population models typically lead to *price skimming* as an optimal strategy, in which prices are lowered over time in such a way that high-valuation customers pay higher prices earlier, while low-valuation customers pay lower prices in later periods. Effectively, this creates a form of second-degree price discrimination, segmenting customers with different values for the good and charging differential prices over time. In infinite-population models, there is no such price-skimming incentive. Provided the distribution of customer valuations does not shift over time, the same price that yields a high revenue in one period will yield a high revenue in later periods, and thus a firm has no incentive to deviate from this revenue-maximizing price.

3.4.3. Monopoly, Oligopoly, and Perfect-Competition Models. Another key assumption in dynamic-pricing models concerns the level of competition the firm faces. Many pricing models used in RM practice are *monopoly models*, in which the demand a firm faces is assumed to depend only on its own price and not on the price of its competitors. Thus, the model does not explicitly consider the competitive reaction to a price change. Again, one makes this assumption primarily for tractability, and it is not always realistic.

As with the myopic-customer model, the monopoly model can be partly justified on empirical grounds—namely, that an observed historical price response has embedded in it the effects of competitors' responses to the firm's pricing strategy. So, for instance, if a firm decides to lower its price, the firm's competitors might respond by lowering their prices. With market prices lower, the firm and its competitors see an increase in demand. The observed increase in demand is then measured empirically and treated as the "monopoly" demand response to the firm's price change in a dynamic-pricing model—even though competitive effects are at work.

Again, while such a view is pragmatic and reflects the conventional wisdom behind the pricing models used in practice, there are some dangers inherent in it, paralleling those of the myopic-customer model. The price-sensitivity estimates may prove wrong if the optimized strategy deviates significantly from past strategies because then the resulting competitive response may be quite different from the historical response. Also, the practice runs the risk of reinforcing "bad" equilibrium responses. Despite these risks, monopoly models have still proved to be valuable for decision support.

It is worth noting that oligopoly models, in which the equilibrium-price response of competitors is explicitly modeled and computed, also have their pitfalls. Most notably, the assumption that firms behave rationally (or quasi-rationally, if heuristics are used in place of optimal strategies) may result in a poor predictor of their actual price response. These potential modeling errors, together with the increased complexity of analyzing oligopoly models and the difficulty in collecting competitor data to estimate the models accurately have made them less popular in practice. Shugan [68] provides a good summary of this point of view; he notes that "the strong approximating assumption of no competitive response is sometimes better than the approximating assumption of pre-existing optimal behavior." However, properly designed and validated, oligopoly models can provide valuable insights on issues of pricing strategy.

Finally, one can also consider perfectly competitive models—in which many competing firms supply an identical commodity. The output of each firm is assumed to be small relative to the market size, and this, combined with the fact that each firm is offering identical commodities, means that a firm cannot influence market prices.[13] Therefore, each firm is essentially a *price taker*—able to sell as much as it wants at the prevailing market price, but unable to sell anything at higher prices. Despite the importance of perfect-competition models in economic theory, the assumption that firms have no pricing power means that the results are not that useful for price-based RM. Nevertheless, they do play a role in quantity-based RM. For example, one can interpret the capacity-control models of Section 2 as stemming from competitive, price-taking models; firms take the price for their various products as given (set by competitive market forces), and control only the quantity they supply (the availability or allocation) at these competitive prices. As our focus in this section is on price-based RM, we do not consider this model of competition further in this section.

3.5. Basic Single-Product Dynamic Pricing Without Replenishment

The first problem we look at is dynamic pricing of a single product over a finite sales horizon given a fixed inventory at the start of the sales horizon. We assume that the firm is a monopolist, customers are myopic, and there is no replenishment of inventory.

The models are representative of the type used in style and seasonal-goods retail RM. For such retailers, production and ordering cycles are typically much larger than the sales season, and the main challenge is to determine the price path of a particular style at a particular store location, given a fixed set of inventory at the beginning of the season.

At one level, such models are simplistic: They consider only a single product in isolation and assume customers are myopic, and therefore demand is a function solely of time and the current price (although other factors such as inventory depletion are sometimes included). They therefore ignore competition, the impact of substitution, and the possible strategic behavior of customers over time. Despite these simplifications, the models provide good rough-cut approximations and are useful in practice. In addition, by decomposing the problem and treating products independently, it is possible to solve such models efficiently even when there are hundreds of thousands of product-location combinations. Finally, even with the simplifying assumptions, the analysis can still become complex if we allow stochastic demand and put constraints on prices.

[13] This is in contrast to the Cournot model of quantity competition, in which there is only a small number of firms whose quantity decisions do affect the market price. Roughly speaking, Cournot competition approaches perfect competition, as the number of firms in the industry tends to infinity.

Because we consider only a single product, there is a single (scalar) price decision at each time t, denoted $p(t)$, which induces a unique (scalar) demand rate $d(t,p)$. The set of allowable prices is denoted Ω_p, and Ω_d denotes the set of achievable demand rates. We assume that these functions satisfy the following regularity conditions:

- The demand functions are continuously differentiable and strictly decreasing, $d'(t,p) < 0$, on Ω_p. Hence, they have an inverse, denoted $d(t,p)$.
- The demand functions are bounded above and below and tend to zero for sufficiently high prices—namely,

$$\inf_{p\in\Omega_p} d(t,p) = 0.$$

- The revenue functions $r(t,p) = pd(t,p)$ (equivalently $r(t,d) = dp(t,d)$) are finite for all $p \in \Omega_p$ and have a finite maximizer interior to Ω_p.
- The marginal revenue as a function of demand, d, defined by

$$J(t,d) \equiv \frac{\partial}{\partial d} r(t,d) = p(t,d) + dp'(t,d),$$

is strictly decreasing in d.

The demand function can also be expressed as $d(t,p) = N_t(1 - F(t,p))$, where N_t is the market-size parameter and $F(t,p)$ is the fraction of the market with willingness to pay less than p. We let $x(t)$ denote the inventory at time $t = 1,\dots,T$, where T is the number of periods in the sale horizon. The initial inventory is $x(0) = C$.

3.5.1. The Model. The simplest deterministic pricing model is formulated in discrete time as follows. Given an initial inventory $x(0) = C$, select a sequence of prices $p(t)$ (inducing demand rates of $d(t,p(t))$) that maximize total revenues. Formulating the problem in terms of the demand rates $d(t)$, the optimal rates $d^*(t)$ must solve

$$\begin{aligned} \max \sum_{t=1}^{T} r(t,d(t)) \\ \text{s.t.} \sum_{t=1}^{T} d(t) \le C \\ d(t) \ge 0. \end{aligned} \tag{26}$$

Let π^* be the Lagrange multiplier on the inventory constraint, and recall that $J(t,d) = \frac{\partial}{\partial d} r(t,d)$ denotes the marginal revenue. Then the first-order necessary conditions for the optimal rates $d^*(t)$ and multiplier π^* are

$$J(t,d^*(t)) = \pi^*, \tag{27}$$

subject to the complementary slackness condition

$$\pi^*\left(C - \sum_{t=1}^{T} d^*(t)\right) = 0 \tag{28}$$

and the multiplier nonnegativity constraint $\pi^* \ge 0$. Assuming that $J(t,d)$ is decreasing in d, $r(t,d)$ is concave; hence, these conditions are also sufficient.

The optimality conditions are quite intuitive. The Lagrange multiplier π^* has the interpretation as the marginal opportunity cost of capacity. The condition $J(t,d^*(t)) = \pi^*$ says that the marginal revenue should equal the marginal opportunity cost of capacity in each period. This makes sense, because if marginal revenues and costs are not balanced, we can increase revenues by reallocating sales (by adjusting prices) from a period of low marginal

TABLE 5. Allocations of capacity between periods 1 and 2 and the marginal values and total revenue.

d_1	d_2	$J(1,d_1)$	$J(2,d_2)$	r
22	18	56	42	2634
23	17	54	43	2646.5
24	16	52	44	2656
25	15	50	45	2662.5
26	14	48	46	2666
27	**13**	**46**	**47**	**2666**.5
28	12	44	48	2664
29	11	42	49	2658.5
30	10	40	50	2650
31	9	38	51	2638.5
32	8	36	52	2624
33	7	34	53	2606.5

revenue to a period of higher marginal revenue. Finally, the complementary slackness condition says that the opportunity cost cannot be positive if there is an excess of stock. If the opportunity cost is zero ($\pi^* = 0$), then if we maximize revenue without a constraint in every period (pricing to the point where marginal revenue is zero), we will still not exhaust the supply. This means it can be optimal—even in the absence of any costs for capacity—not to sell all the available supply.

Note that this problem is essentially equivalent to the problem of optimal third-degree price discrimination if we consider customers in each period t to be different segments who are offered discriminatory prices $p(t)$. Another way of viewing the above argument is that the firm, faced with a capacity constraint, decides how much to sell in each period, and its optimal allocation of capacity occurs when the marginal revenue in all the periods is the same. The following example illustrates the idea:

Example 3. Consider a two-period selling horizon, where during the first period demand is given by $d_1 = -p_1 + 100$ and in period 2 demand is given by $d_2 = -2p_2 + 120$. (Customers in the second period are more price sensitive than those in the first period.) Purchase behavior is assumed to be myopic. Considered separately, the revenue-maximizing price for the first period (maximizing $r_1 = p_1(-p_1 + 100)$) is given by $p_1^* = 50$ and $d_1^* = 50$, and in the second period by $p_2^* = 30, d_2^* = 60$ (maximizing $r_2 = p_2(-2p_2 + 120)$).

Intertemporal effects come into play if the firm has only a limited number of items to sell (less than 50+60). Suppose the firm's capacity is 40. How should it divide the sale between the two periods?

Note that here, $J(1,d_1) = -2d_1 + 100$ and $J(2,d_2) = -d_2 + 60$. Consider the table of marginal values, Table 5, at various allocations and the corresponding revenues. The total revenue is maximized at the point where the marginal values for the two periods are approximately the same (when $d_1 = 27, d_2 = 13$), conforming to our intuition; if they were not equal, the firm would reallocate capacity to the higher marginal-value period.

To see qualitatively how prices will change over time, we can write the optimality condition (27) as

$$\frac{p^*(t) - \pi^*}{p^*(t)} = \frac{1}{|\epsilon(t,p^*)|},$$

where $\epsilon(t,p)$ is the elasticity of demand in period t, defined by

$$\epsilon(t,p) \equiv \frac{p}{d(t,p)} \frac{\partial d(t,p)}{\partial p}.$$

Thus, more elastic demand in period t implies a lower optimal price $p^*(t)$.

For example, if customers that buy toward the end of the sales horizon are more price sensitive than those that buy early, then optimal prices will decline over time. If customers early on are price sensitive, and those buying later are less price sensitive, then optimal prices will increase over time. This observation offers one explanation for why in some industries (such as apparel retailing) prices tend to decline over time, while in others (such as airlines) prices increase over time.

Discrete Prices. Often, in practice, we would like to choose prices from a discrete set. For example, prices close to convenient whole dollar amounts (such as $24.99 or $149.99), or fixed percentage markdowns (such as 25% off or 50% off) are often used because they are familiar to customers and easy to understand. In such cases, it may be desirable as a matter of policy to constrain prices to a finite set of k discrete price points, so that $p(t) \in \Omega_p$, where $\Omega_p = \{p_1, \ldots, p_k\}$. Equivalently, the sales rate $d(t)$ is constrained to a discrete set $d(t) \in \Omega_d(t)$ (time varying in this case if the demand function is time varying), where $\Omega_d(t) = \{d_1(t), \ldots, d_k(t)\}$, and $d_i(t) = d(t, p_i)$ denotes the sales rate at time t when using the price p_i.

The discreteness of the prices imposes technical complications when attempting to solve the dynamic pricing problem (26) because the problem is no longer continuous or convex. However, one can overcome this difficulty by relaxing the problem to allow the use of convex combinations of the discrete prices (or demand rates). In most periods, the optimal solution will be to use only one of the discrete prices; in the remaining periods, the solution has the interpretation of allocating a fraction of time to each of several prices.

To see this, define a vector of new variables $\alpha_i(t)$ for each t, $\boldsymbol{\alpha}(t) = (\alpha_1(t), \ldots, \alpha_k(t))$, which represent convex weights: They are nonnegative and sum to one. Next, in each period replace the variable $d(t)$ with the convex combination

$$d(t) = \sum_{i=1}^{k} \alpha_i(t) d_i(t),$$

and replace the constraint $d(t) \in \Omega_d(t)$ with the constraint

$$\boldsymbol{\alpha}(t) \in W \equiv \left\{ \boldsymbol{\alpha} \in \Re^k : \sum_{i=1}^{k} \alpha_i = 1, \alpha \geq 0 \right\}.$$

The optimization problem is then

$$\begin{aligned} &\max_{\boldsymbol{\alpha}(t) \in W} \sum_{t=1}^{T} \sum_{i=1}^{k} r_i(t) \alpha_i(t) \\ &\text{s.t.} \quad \sum_{t=1}^{T} \sum_{i=1}^{k} \alpha_i(t) d_i(t) \leq C, \end{aligned} \tag{29}$$

where $r_i(t) = p_i d_i(t)$ is the revenue rate at price p_i. This is a linear program in the variables $\boldsymbol{\alpha}(t)$, so it is easy to solve numerically.

To relate the solution to the unconstrained price case, introduce a dual variable π^* on the capacity constraint as before. The optimal solution $\boldsymbol{\alpha}^*(t)$ in each period is then characterized by solving

$$\max_{\boldsymbol{\alpha}(t) \in W} \left\{ \sum_{i=1}^{k} \alpha_i(t)(r_i(t) - \pi^* d_i(t)) \right\}, \tag{30}$$

where $\pi^* \geq 0$ and $\boldsymbol{\alpha}^*(t)$ are convex weights satisfying the complementary slackness condition

$$\pi^* \left(\sum_{t=1}^{T} \sum_{i=1}^{k} \alpha_i^*(t) d_i(t) - C \right) = 0. \tag{31}$$

Because the objective function of (30) is linear in $\alpha(t)$, if there is a unique index i^* for which $r_{i^*}(t) - \pi^* d_{i^*}(t)$ is greatest, then the optimal solution is simply $\alpha_{i^*}(t) = 1$, which corresponds to using the discrete price p_{i^*}. If there is more than one such value i^*, then there will be multiple solutions to (30), and determining which is optimal can be resolved by appealing to the complementary slackness condition (31). Of course, such a choice could result in a fractional solution in which $\alpha_i(t) > 0$ for two or more values i. However, this can be interpreted as saying that we should use the price i for a fraction $\alpha_i(t)$ of period t. Hence, the solution of (29) can be converted in practice into a discrete-price recommendation.

Inventory-depletion effect. Another practical factor affecting dynamic pricing in many retailing contexts is the adverse effects of low inventory levels. This is sometimes referred to in retailing as a *broken-assortment effect*. For example, if the inventory-pricing model is applied at an aggregate item level, where an item contains several SKUs—such as color-size combinations in apparel retailing—then when inventories run low, certain SKUs may be out of stock even though there is a positive inventory for the item as a whole (for example, if a color or size runs out). The resulting reduction in alternatives naturally reduces the sales rate at any given price. Indeed, empirical studies have confirmed a positive correlation between inventory levels and sales rates (Bhat [12]).

These inventory-depletion effects can be modeled by making the demand rate a function of inventory as well as of price and time, so that the demand rate becomes $d(t, p(t), x(t))$. We can use a variety of functional forms to represent this inventory-depletion effect. For example, one proposed model is the following multiplicative form (Smith and Achabal [70]):

$$\hat{d}(t, x(t)) = d(t)g(x(t)), \tag{32}$$

where $g(\cdot)$ is a depletion-effect term. We will call $d(t)$ the *unadjusted sales rate* (the rate of sales if inventory were unlimited) and $\hat{d}(t, x(t))$ the *adjusted sales rate* (the rate adjusted for inventory-depletion effects). One choice for g is

$$g(x) = 1 - \gamma \max\{0, 1 - x/x_0\},$$

where x_0 is the minimum *full-fixture inventory* and $0 \le \gamma \le 1$ is a sensitivity parameter. Both x_0 and γ can be estimated from historical data. Note that $g(x)$ is concave in x.

Another possible form is

$$g(x) = e^{-\gamma \max\{0, 1 - x/x_0\}},$$

where γ and x_0 have the same interpretation (see Smith and Achabal [70]).

For this model with inventory depletion one must keep track of the inventory at each time t in the optimization problem. For example, assuming the multiplicative inventory-depletion model of (32) and formulating the problem in terms of the unadjusted sales rate $d(t)$, the inventory evolves according to the state equation

$$x(t+1) = x(t) - d(t)g(x(t)),$$

and the revenue-maximization problem can be formulated as

$$\begin{aligned} \max_{d(t) \ge 0} \; & \sum_{t=1}^{T} r(t, d(t)) g(x(t)) \\ \text{s.t.} \; & x(t+1) = x(t) - d(t)g(x(t)), \quad t = 1, \ldots, T \\ & x(T) \ge 0, \\ & x(0) = C, \end{aligned} \tag{33}$$

where $r(t, d(t)) = p(t, d(t))d(t)$ is the unadjusted revenue-rate function.

While somewhat more complex than the case without inventory-depletions effects, this is still a relatively simple nonlinear program to solve because the objective function is separable and the constraints are linear. (The objective function, however, is not necessarily jointly concave even if $r(t, d(t))$ and $g(x)$ are both concave.)

One qualitative impact of this inventory-depletion phenomenon is that optimal prices may decline over time even though the unadjusted revenue-rate function is time invariant. (Recall that in the problem without inventory-depletion effects, a time-invariant revenue-rate function implied a time-invariant optimal price.) For example, Smith and Achabal [70] show, for the continuous-time version of this model, that if the unadjusted revenue-rate function is constant and the inventory-depletion effect is multiplicative, then optimal prices decline over time in such a way that the adjusted sales rate $g(x(t))d(t)$ is constant; that is, as inventory depletion reduces demand, the optimal prices fall to exactly compensate for the drop in sales due to inventory depletion.

3.6. Multiproduct, Multiresource Pricing

Multiproduct, multiresource—or network—versions of dynamic-pricing problems arise in many applications. Two fundamental factors typically link the pricing decisions for multiple products. First, demand for products may be correlated. For example, when products are substitutes or complements, the price charged for one product affects the demand for other related products. Then, a firm jointly managing the pricing of a family of such products must consider these cross-elasticity effects when determining its optimal pricing policy. Second, products may be linked by joint capacity constraints. For example, two products may require the same resource, which is available in limited supply. Even if there are no cross-elasticity effects between the two products, the pricing decision for one product will need to account for the joint effect on demand for the other product that uses the limited resource.

As in the case of capacity controls, most problems in real life are multiproduct problems, either because of cross-elasticity effects or because of joint capacity constraints, or both. For example, a grocery store that is pricing brands in a food category—say, salty snacks—needs to consider the cross-elasticity effects of its pricing decision for all products in the category. An increase in the price of a packet of potato chips will not just cause a drop in demand for potato chips, but will likely also increase the demand for corn chips. At the same time, these products may occupy the same limited shelf space, so stocking more of one product may require stocking less (or none) of other products.

We can model such situations using multiproduct demand functions and joint capacity constraints on resources. However, like the network problems of capacity control, such formulations quickly become difficult to analyze and solve, which is the reason that many commercial applications of dynamic-pricing models make the simplifying assumption of unrelated products and independent demands and solve a collection of single-product models as an approximation.

However, in cases where cross-elasticity or resource-constraint effects are strong—for example, when products are only slightly differentiated, customers are very price sensitive, or joint capacity constraints are tight—then ignoring multiproduct effects can be severely suboptimal. In such cases, we must solve a pricing problem incorporating these effects—or at least approximating them in some fashion. In this section, we look as such multiproduct, multiresource models and methods.

3.6.1. A Basic Deterministic Model Without Replenishment. Under a deterministic demand assumption, it is relatively straightforward to formulate a multiproduct, multiresource version of dynamic pricing similar to those described in §3.5. There are n products, indexed by j, and m resources, indexed by i. There is a horizon of T periods, with each period indexed by t. Let $\mathbf{d} = (d_1, \ldots, d_n)$ denote the demand rate for the n products and $\mathbf{p}(t, \mathbf{d})$ denote the inverse-demand function in period t. We further assume that the revenue-rate function $r(t, \mathbf{d})$ is bounded and jointly concave in $\mathbf{d}$.

Product j uses a quantity a_{ij} of resource i. The matrix $\mathbf{A} = [a_{ij}]$ therefore describes the *bill of materials* for all n products. We assume there are limited capacities $\mathbf{C} = (C_1, \ldots, C_m)$ of the m resources.

The dynamic-pricing problem can then be formulated as finding a sequence of demand vectors $d^*(t)$ that maximizes the firm's total revenue subject to the capacity constraints C:

$$\begin{aligned} \max \; & \sum_{t=1}^{T} r(t, \mathbf{d}(t)) \\ \text{s.t.} \; & \sum_{t=1}^{T} \mathbf{A}\mathbf{d}(t) \le \mathbf{C} \\ & \mathbf{d}(t) \ge 0, \quad t = 1, \ldots, T. \end{aligned} \tag{34}$$

We assume $r(t, \mathbf{d})$ is concave in $\mathbf{d}$, and therefore, the following Kuhn-Tucker conditions are necessary and sufficient for characterizing an optimal solution $\mathbf{d}^*(t)$ to (34):

$$J(t, \mathbf{d}^*(t)) = \mathbf{A}^\top \boldsymbol{\pi}^* \tag{35}$$

$$\boldsymbol{\pi}^{*\top} \left(\mathbf{C} - \sum_{t=1}^{T} \mathbf{A}\mathbf{d}(t) \right) = 0 \tag{36}$$

$$\boldsymbol{\pi}^* \ge 0, \tag{37}$$

where $J(t, \mathbf{d}) = \nabla_d r(t, \mathbf{d})$ is the marginal-value vector and $\boldsymbol{\pi}^*$ is the optimal dual price on the joint-capacity constraints, having the usual interpretation as the vector of marginal opportunity costs (marginal values) for the m resources. Condition (35) says that at the optimal sales rate, the marginal revenue for each product j should equal the marginal opportunity cost of the resources used by product j, or $\boldsymbol{\pi}^{*\top} \mathbf{A}_j$. Condition (36) says that the marginal opportunity cost of resource i can be positive only if the corresponding capacity constraint for resource i is binding. Finally, (37) requires that the marginal opportunity costs be nonnegative.

The nonlinear program (34) is relatively easy to solve numerically because the objective function is concave and the constraints are linear. (See Bertsekas [9, 10] for specific techniques.)

Example 4. Consider the six-node airline network shown in Figure 6. Nodes 2 and 3 are "hub" nodes. (Leg seat capacities are as indicated in the figure.) For a given path j on the network, the revenue function is time homogeneous and log linear

$$d_j(p_j) = a_j e^{-\epsilon_j (p_j / \bar{p}_j - 1)},$$

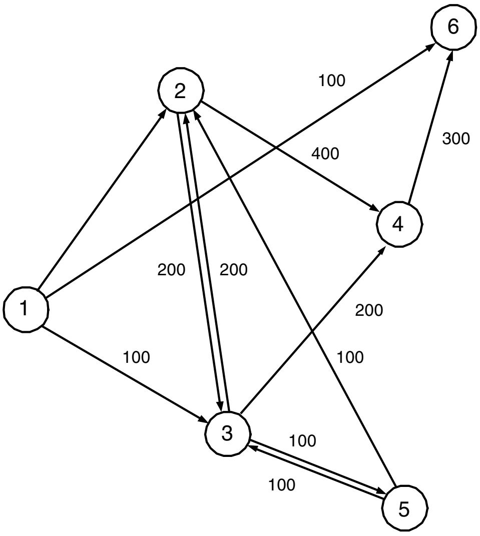
where $\bar{p}_j$ is interpreted as a reference price for itinerary j, a_j is the demand rate at the reference price, and ϵ_j is the magnitude of the elasticity of demand at the reference price. Demand-function parameters for all O-D pairs are shown in Table 6, along with the path (itinerary) used by each O-D pair.

Because the demand functions are time homogeneous, optimal prices are constant over time. The optimal O-D prices and demand are shown in the last two columns in Table 6. The solution gives a total revenue of \$661,200 across all O-D pairs.

3.6.2. Action-Space Reductions. One simplification that is useful for multiproduct dynamic-pricing problems is to express the problem in terms of resource-consumption rates rather than the demand rates $\mathbf{d}$. This yields an equivalent formulation with an often greatly reduced dimensionality that can be much easier to solve. The approach is due to Maglaras and Meissner [56].

To illustrate the main idea, consider the case of the deterministic model (34), where there is only $m = 1$ resource but $n > 1$ products. For example, this could be a situation similar

Figure 6. A six-node, two-hub airline network.



to the traditional single-resource problem of §2, but one in which we control the demand for each product j, d_j, by adjusting its price p_j. The deterministic problem (34) in this case is then

$$\begin{aligned} &\max \sum_{t=1}^{T} r(t, \mathbf{d}(t)) \\ &\text{s.t.} \ \sum_{t=1}^{T}\sum_{j=1}^{n} d_j(t) \leq C \\ &\qquad \mathbf{d}(t) \geq 0, \quad t = 1, \ldots, T. \end{aligned} \tag{38}$$

Table 6. Demand-function parameters, itineraries, and optimal solution for Example 4.

Market		Demand function				Optimal solution	
O	D	a_j	ϵ_j	$\bar{p}_j$	Path	d_j^*	p_j^*
1	2	300	1.0	220	1–2	135	\$396.62
1	3	300	1.2	220	1–3	67	\$495.86
1	4	300	2.0	400	1–2–4	165	\$520.11
1	5	300	1.0	250	1–3–5	33	\$752.04
1	6	300	0.8	200	1–6	100	\$525.58
2	3	300	1.0	230	2–3	168	\$364.28
2	4	300	0.9	200	2–4	143	\$365.74
2	5	300	2.0	200	2–3–5	32	\$423.79
2	6	300	1.0	200	2–4–6	92	\$436.80
3	2	300	1.0	200	3–2	200	\$281.76
3	4	300	2.0	230	3–4	131	\$325.30
3	5	300	2.0	120	3–5	35	\$249.51
3	6	300	2.0	150	3–4–6	14	\$378.60
4	6	300	1.0	150	4–6	162	\$243.30
5	2	300	1.0	200	5–2	100	\$420.39
5	3	300	2.0	150	5–3	47	\$289.90
5	4	300	1.0	160	5–3–4	21	\$585.20
5	6	300	1.0	230	5–3–4–6	32	\$748.50

To reduce the dimensionality of this problem, we express the problem in terms of the aggregate-demand rate rather than the individual demand rates $\mathbf{d}$. To this end, define the aggregate-demand rate

$$\hat{d} = \sum_{j=1}^{n} d_j,$$

and for a given $\hat{d}$ define the maximized revenue-rate function by

$$\begin{aligned} \hat{r}(t,\hat{d}) = \max\; & r(t,\mathbf{d}) \\ \text{s.t.}\; & \sum_{j=1}^{n} d_j = \hat{d} \\ & \mathbf{d} \geq 0. \end{aligned} \tag{39}$$

That is, $\hat{r}(t,\hat{d})$ is the instantaneous maximum revenue rate given that the total demand rate (equivalently, the resource *consumption rate*) is constrained to be $\hat{d}$. It is easy to show that if $r(t,\mathbf{d})$ is jointly concave in $\mathbf{d}$, then $\hat{r}(t,\hat{d})$ will be concave in $\hat{d}$.

Using these new variables, we can then formulate (38) as

$$\begin{aligned} \max\; & \sum_{t=1}^{T} \hat{r}(t,\hat{d}(t)) \\ \text{s.t.}\; & \sum_{t=1}^{T} \hat{d}(t) \leq C \\ & \hat{d}(t) \geq 0 \quad t = 1,\ldots,T. \end{aligned} \tag{40}$$

Note that this is now a problem that is equivalent to a single-product pricing problem of the same form as (26) with a scalar demand rate $\hat{d}$ and revenue-rate functions $\hat{r}(t,\hat{d})$. Once we solve for the optimal demand rates $\hat{d}^*(t)$, we can then convert these into optimal vectors of demand rates $\mathbf{d}^*(t)$ by inserting $\hat{d}^*(t)$ into the optimization problem (39). Thus, the solution proceeds in two steps: First, solve (40) to determine the optimal aggregate-sales rate, and then solve (39) at each time t to disaggregate this optimal aggregate rate into an optimal vector of sales rates (equivalently prices) for each product. This same action-space-reduction approach also works for stochastic versions of this problem, and it extends to the general multiproduct ($m > 1$), multiresource problem (34) as well.

3.7. Finite-Population Models and Price Skimming

We next consider what effect a finite-population assumption has on an optimal dynamic-pricing policy. Recall that a finite-population model assumes that we sample customers without replacement from a finite number of potential customers. Thus, the history of demand (how many customers have purchased, how much they paid, and so on) affects the distribution of both the number and valuations of the remaining customers.

Because the finite-population assumption is more complex, we focus on deterministic models of this situation. However, we consider both a myopic and strategic customer version of the problem.

3.7.1. Myopic Customers. Recall that a myopic customer is assumed to purchase the first time the current price $p(t)$ drops below his valuation v. Combined with the finite-population assumption, this behavior can be exploited by the firm to achieve *price skimming*—a version of classical second-degree price discrimination.

Assume for simplicity that there is a finite population size N and that customers in this population have valuations v that are uniformly distributed on the interval $[0,\bar{v}]$. As an approximation, we assume that sales can occur in fractions, so the population can be regarded as continuous. The important point to note is that the fraction of customers who purchased until time t leave the population of customers for the remaining sale period.

As a result of the myopic-customer assumption, if the firm offers a price p, $N(1-\bar{v}/p)$, customers will buy. Also, by the finite-population assumption, there will then be $N\bar{v}/p$ remaining customers, with valuations uniformly distributed on the interval $[0,p]$.

Now, consider a firm that sells a fixed capacity C of a product to this population over T time periods. The firm is free to set different prices in each period. What is the optimal pricing strategy?

First, it is not hard to see that the optimal prices are decreasing over time, because (by the myopic-customer assumption) the only customers left at time t are those with values less than the minimum price offered in periods $1,\ldots,t-1$. Hence, the firm will sell nothing if it posts a price in period t that is higher than the minimum price offered in the past. This observation, applied inductively, shows that the optimal prices must decline over time. Moreover, note that if $p(t)\le p(t-1)$ for all t, the revenue generated in period t is given by

$$p(t)\frac{N}{\bar{v}}(p(t-1)-p(t)),$$

where we define $p(0)=\bar{v}$. This is because $(N/\bar{v})(p(t-1)-p(t))$ is the number of customers with valuations greater than $p(t)$ but less than the lowest previous price $p(t-1)$.

To see the effect the decreasing price schedule has on the optimal pricing policy, assume for simplicity that $C>N$, so the capacity constraint is never binding. In this case, the firm must solve

$$\max\sum_{t=1}^{T}\frac{N}{\bar{v}}p(t)(p(t-1)-p(t)) \tag{41}$$

$$\text{s.t. } p(t)\le p(t-1),\quad t=1,\ldots,T \tag{42}$$

$$p(0)=\bar{v}, \tag{43}$$

$$p(t)\ge 0. \tag{44}$$

Note that the objective function is jointly concave in $p(t), t=1,\ldots,T$. It is not hard to see that the constraints (42) are redundant, because the objective function (41) will penalize the use of a price $p(t)>p(t-1)$. Therefore, ignoring constraints (42) and defining $p(T+1)=0$, the first-order conditions imply that the optimal unconstrained solution must satisfy

$$p(t)=\frac{p(t-1)-p(t+1)}{2},\quad t=1,\ldots,T.$$

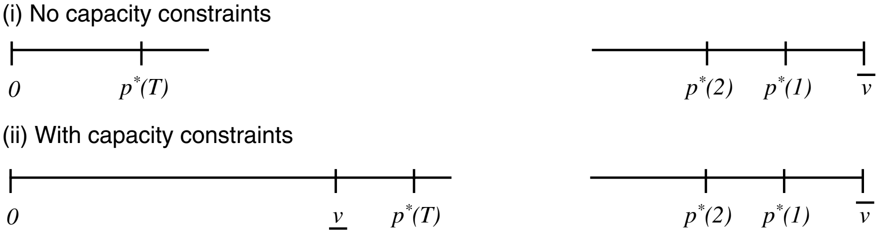
One can easily verify that the solution

$$p^*(t)=\bar{v}\left(1-\frac{t}{T+1}\right) \tag{45}$$

satisfies these first-order conditions. Because the optimization problem (41–d) is strictly concave and (45) satisfies the inequality constraints $p(t)\le p(t-1)$ for all t, it is in fact the unique optimal solution for (41–d). This solution is illustrated in Figure 7(i).

The optimal pricing strategy effectively exploits the myopic behavior of customers to segment them into $T+1$ groups based on their valuations, and then price discriminates based on this segmentation. Specifically, as shown in Figure 7, segment t consists of those customers whose valuations are in the range $[p^*(t),p^*(t-1)]$, and these segments pay a declining price $p^*(t)$ given by (45). Segment $T+1$ has values in the range $[0,\bar{v}/(T+1)]$ and is not served at all.

FIGURE 7. Optimal price-skimming solution for myopic customers: (i) no capacity constraints, (ii) with capacity constraints.



There are several interesting observations about this solution. First, note that we can write the optimal price in period t as

$$p^*(t) = \frac{p^*(t-1)}{2} + \frac{\bar{v}}{2}\left(1 - \frac{t}{T+1}\right).$$

The first term on the right, $p^*(t-1)/2$, is simply the single-period revenue-maximizing price, which follows from the fact that the remaining customers in period t have values uniformly distributed on $[0, p^*(t-1)]$. Therefore, the optimal price in period t is higher than the single-period revenue-maximizing price for period t (except in the last period $t = T$, where they are equal). Intuitively, this occurs because there is an additional benefit to the firm of raising its price in period t in the multiperiod setting; namely, it will have more customers to sell to in the future.

Second, note that the price changes over time not because the distribution of valuations changes over time—as in the infinite-population model of demand—but because the firm seeks to price discriminate among the finite population of customers. For example, in an equivalent infinite-population model (essentially, the model of §3.5.1 with a linear demand function), the distribution of values of customers is unaffected by past demand, and hence the distribution would still be uniform over $[0, \bar{v}]$ in each period. In this case, the optimal price to charge in each period would be a constant $\bar{v}/2$ rather than the declining price (45). Therefore, a finite population of customers creates an incentive to offer dynamically decreasing prices to achieve price discrimination, an incentive that is not present in infinite-population models.

Finally, note that if the number of periods T increases, the firm's revenues increase because one can show (after some algebra) that the optimal total revenue for T periods is

$$\sum_{t=1}^{T} p^*(t)\frac{N}{\bar{v}}(p^*(t) - p^*(t-1)) = \frac{N\bar{v}}{2}\left(\frac{T}{T+1}\right).$$

Indeed, as T tends to infinity, the firm achieves perfect price discrimination and captures the entire consumer surplus $N\bar{v}/2 = \int_0^{\bar{v}}(N/\bar{v})dv$; each customer ends up paying a price arbitrarily close to his valuation. In particular, a continuous-time model of this problem can achieve perfect price discrimination because the firm can continuously lower prices from $\bar{v}$ down to zero over the interval $[0, T]$. A number $dp(N/\bar{v})$ of customers with values $[p, p+dp]$ will buy when the price is p, so the firm achieves a revenue of $\int_0^{\bar{v}} p(N/\bar{v})dp = N\bar{v}/2$, which is the entire consumer surplus.

3.7.2. Strategic Customers. One might question why customers would behave myopically when faced with a price-skimming strategy. Indeed, knowing that prices will decline over time, rational customers could do better (increase their net utility) by deviating from myopic behavior and delaying purchase until the price is much lower than their valuation. Such behavior is quite plausible and is a valid criticism of the myopic-customer model, but it complicates the analysis of the firm's optimal-pricing policy considerably. Here we focus on the effect of strategic customers on the price-skimming strategy alone. Throughout this section we consider only the case where the firm has no capacity constraint ($C > N$).

To proceed, one first has to make assumptions about whether the firm can credibly commit to a schedule of prices over time or whether the firm must follow a subgame-perfect equilibrium-pricing strategy. In our case, requiring a subgame-perfect equilibrium means that the strategy for the firm at each time t has to be an equilibrium for the residual revenue-maximization game over the horizon $t, t+1, \ldots, T$, given whatever state the firm and customers were in period t.

For example, if the firm can commit to a price schedule, then a rational customer will simply look at the schedule of prices and (assuming no discounting of utility) decide to purchase in the period with the lowest price, and only customers with valuations above this lowest price will decide to purchase. So, effectively, it is only the lowest price among the T periods that matters to customers. Given this fact (and ignoring capacity constraints), the firm will then set this minimum price as the single-period revenue-maximizing price, which, in the case where customer valuations are uniformly distributed on $[0, \bar{v}]$, is just $\bar{v}/2$. The firm will then set arbitrary but higher prices in the other periods. Which period the firm chooses for the minimum price does not matter unless revenues are discounted, in which case the firm would prefer collecting revenues sooner rather than later and would choose period 1. The total revenue the firm receives is then $N\bar{v}^2/4$, which is just the product of the price $\bar{v}/2$ and the number of customers willing to pay that price, $N\bar{v}/2$. One can formalize this reasoning and show that this is indeed the equilibrium strategy in the case where the firm has to commit to a price schedule.

Note that the fact that customers are rational has eliminated the ability of the firm to price discriminate; the firm is forced to offer a single uniform price to all customers. Moreover, the firm's revenue is strictly worse under this model. This is to be expected; the firm ought to do worse when customers are "smarter."

However, the single-period strategy outlined above is not always subgame-perfect. To see why, suppose this lowest price $\bar{v}/2$ occurs in period 1. Then in period 2, there will be a population of customers with values less than $\bar{v}/2$ who have not purchased. If the firm has any remaining supply after period t, it would rather sell the remaining stock at some positive price than let it go unsold. Thus, it has an incentive to lower the price in period 2 to capture some of the remaining customers. However, rational customers realize the firm faces this temptation after period 1 and, anticipating the price drop, do not purchase in period 1, so offering the lowest price in period 1 cannot be a subgame-perfect equilibrium.

Besanko and Winston [11] analyze the subgame-perfect pricing strategy. The equilibrium is for the firm to lower prices over time, similar to the price-skimming strategy of §3.7.1. In the case where revenues are not discounted, this equilibrium results in the firm setting a declining sequence of prices, where the price in the last period T is simply the single-period optimal price $\bar{v}/2$; all customers buy only in the last period. This case is essentially equivalent to the case where the firm can commit to a schedule of prices, with the exception that the firm is forced to offer the lowest price only in the last period.

The situation is somewhat more interesting if revenues and customer utility are discounted at the same rate. In this case, the subgame-perfect equilibrium has customers with high values buying in the early periods and those with lower values buying in later periods, again, as in the price-skimming case of §3.7.1. However, unlike the price-skimming case, the equilibrium price in each period is *lower* than the single-period revenue-maximizing price for the customers remaining in that period. In particular, in period 1 the equilibrium price is less than $\bar{v}/2$, and the equilibrium price declines in subsequent periods. Thus, the firm is strictly worse off than when it can commit to a price schedule. This is because when the firm can commit to its price schedule, it can force all customers to purchase in period 1 by simply offering very high prices in periods $t > 1$ while setting a price of exactly $\bar{v}/2$ in period 1. All customers will then buy in period 1 at a price of $\bar{v}/2$.

Besanko and Winston [11] show that with strategic customers, the firm is always better off with fewer periods; that is, the firm's equilibrium revenue is decreasing in the number of periods. This is because the inability of the firm to commit to prices in later periods hurts it, and the more periods, the more often the firm falls victim to the temptation to lower prices. That is, it discounts early and often. This is to be contrasted with the case of myopic customers, where the firm's revenues are increasing in the number of periods. Thus, although the strategy looks like price skimming, rational customers create a qualitatively different situation for the firm than do myopic customers.

4. Summary and Conclusions

The notion that a firm's demand should be actively managed—and that scientific methods can help improve demand decisions—is the essence of revenue management. In this chapter, we have given an overview of the field, its origins and applications, and a sampling of the models and methods used. Still, there is much more to the subject than what we have presented here. Estimating and forecasting market response, for example, is a vast and important element of revenue management on which we have not touched. There are also many relevant concepts from economics on related topics such as price discrimination, peak-load pricing, mechanism design, and oligopoly pricing that augment the operational theory surveyed here. System implementation is also a vital part of RM practice. These and other topics are covered in our book, *The Theory and Practice of Revenue Management* (Kluwer 2004), for those readers interested in more depth and detail. What makes the subject fascinating is that it is in many ways the quintessential OR topic, combining a vitally important business application with sophisticated techniques and concepts from economics, statistics, and optimization.

In addition, the future of revenue management looks equally bright. The domain of industrial application is spreading rapidly beyond the transportation and hospitality industry. Retailing is already a major industry user of RM. Manufacturing, advertising, energy, and financial services applications are growing. With each new industry application, one encounters new challenges in modeling, forecasting, and optimization, so research in the area is also blossoming. It is an exciting field to follow and be a part of, and will likely stay that way for many years to come.

Appendix: Notes and Sources

Introduction and Overview Articles

The 1997 book by Robert Cross, *RM: Hard Core Tactics for Market Domination* (Cross [21]) was influential in popularizing the story of airline RM and introducing the concept of RM to the general business community. Several other books on RM have been published recently; Ingold et al. [42] focuses primarily on the hotel industry, and Daudel and Vialle [24] focuses on air transportation. The book by Nagel and Holden [61] provides a comprehensive overview of many managerial issues involved in pricing, and is useful reading.

Several survey articles provide general coverage of RM. The *Handbook of Airline Economics* edited by Jenkins [44] provides several good practice-oriented articles on RM in the airline industry. Kimes [49] provides a conceptual introduction to RM with a hotel RM focus. Smith et al. [69] provide a nice description of the practice of RM at American Airlines and the DINAMO system.

As for guides to the research literature, Weatherford and Bodily [81] propose a taxonomy for classifying the sets of assumptions used in many traditional RM models, although the taxonomy itself is little used. McGill and van Ryzin [58] provide a comprehensive overview and annotated bibliography of the published academic literature in the field through 1998. Elmaghraby and Keskinocak [27] provide a survey on research in the area of dynamic pricing.

Single-Resource Capacity Control

The notion of theft versus standard nesting is not well-documented and is part of the folklore of RM practice. Our understanding, however, greatly benefited from discussions with our colleagues Peter Belobaba, Sanne de Boer, and Craig Hopperstad.

The earliest paper on the static models of §2.3 is Littlewood [55]. Another early applied paper is Bhatia and Parekh [13]. However, there are close connections to earlier work on the stock-rationing problem in the inventory literature by Kaplan [47] and Topkis [78]; see also Gerchak and Parlar [36], Gerchak et al. [37], and Ha [39]. Indeed, Topkis's [78] results can be used to show the optimality of nested-allocation policies.

Optimal policies for the $n > 2$ case were obtained in close succession (using slightly different methods and assumptions) in papers by Brumelle and McGill [17], Curry [22], Robinson [67], and Wollmer [84]. See also McGill's thesis [57]. Robinson [67] also analyzed the case where the order of arrival is not the same as the revenue order. Brumelle et al. [19] analyzed a two-class static model with dependent demand.

The dynamic model of §2.4 was first analyzed by Lee and Hersh [53]. Lautenbacher and Stidham [51] provide a unified analysis of both the static and dynamic single-resource models. Walczak and Brumelle [18] relate this problem to a dynamic-pricing problem using a Markov model of demand that allows for partial information on the revenue values or customer types. See Liang [54] for an analysis of a continuous-time version of the dynamic model.

The EMSR-a and EMSR-b heuristics are both due to Belobaba. The most detailed coverage of EMSR-a is contained in Belobaba's 1987 thesis [3], but see also the published articles from it (Belobaba [4, 5]). EMSR-b was introduced in Belobaba [6]; see also Belobaba and Weatherford [7].

The buy-up heuristics in §2.5.1 are due to Belobaba [3, 4, 5]. See also Belobaba and Weatherford [7], Weatherford et al. [82], and the simulation study of Bohutinsky [16]. See Titze and Griesshaber [77] for a discussion of passenger behavior in the simple two-class model. The material on choice-based models in §2.5.2 is from Talluri and van Ryzin [74]; see also Algers and Besser [1] and Andersson [2] for an application of discrete-choice models at SAS. For a good reference on discrete-choice modeling, see Ben-Akiva and Lerman [8]. De Boer [25] is another recent work that addresses customer choice in a single-resource problem.

Dynamic Pricing

The book by Nagle [61] provides a good general-management overview of pricing decisions. Elmaghraby and Keskinocak [27] provide a survey on research in the area of dynamic pricing. As for the connection between pricing- and capacity-allocation decisions, see Walczak and Brumelle [18].

Smith and Achabal [70] study a continuous-time version of the problem with inventory-depletion effect, as in §3.5.1. Heching et al. [40] provide revenue estimates based on a regression test of this same type of deterministic model on data from an apparel retailer.

Gallego and van Ryzin [34] analyzed a continuous-time, time-homogeneous stochastic model, providing monotonicity properties of the optimal price, an exact solution in the exponential demand case, and proving the asymptotic optimality of the deterministic policy. Bitran and Mondschein [14] analyze a discrete-time model of this problem and test it on apparel retail data. Zhao and Zheng [87] analyze the continuous-time model with a time-varying demand function and provide an alternative proof of monotonicity of the marginal values. See also Kincaid and Darling [50] and Stadje [71]. Das Varmand and Vettas [23] analyze the problem of selling a finite supply over an infinite horizon with discounted revenues, where the discounting provides an incentive to sell items sooner rather than later and there is no hard deadline on the sales season.

Stochastic models with discrete price changes are analyzed in the continuous-time case in a series of papers by Feng and Gallego [28, 29] and Feng and Xiao [31, 32]. The problems differ in terms of whether there are two prices or more than two prices, whether the price changes are reversible or one-way changes. Feng and Gallego [29] extend the analysis also to the interesting case where demand is Markovian and may depend on the current inventory level—for example, as in the classical Bass model of new-product diffusion. The notion of the *maximum concave envelope* of prices is due to Feng and Xiao [31]. See also You [85] for a discrete-time analysis of the problem.

There is an extensive literature on production-pricing problems, which we have not covered in this chapater. Eliashberg and Steinberg [26] provide of review of joint pricing and production models. Single-period, convex-cost problems under demand uncertainty are analyzed by Karlin and Carr [48], Mills [60], and the early paper of Whitin [83]. The literature on single-period pricing under demand uncertainty (the price-dependent newsvendor problem) is surveyed by Petruzzi and Dada [65]. Multiperiod, convex-cost models are analyzed by Hempenius [41], Thowsen [76], and Zabel [86].

Multiproduct, multiresource dynamic-pricing problems were analyzed in Gallego and van Ryzin [35], including bounds on the relationship between the stochastic and deterministic versions of the problem. The action-space-reduction approach is a recent result due to Maglaras and Meissner [56]. A related network pricing we have omitted is congestion pricing for communications service; see, for example, Pashalidis and Tsitiklis [62].

Stokey [72] analyzes a model of intertemporal price discrimination similar to that presented in §3.7.1. See also Kalish [46]. Stokey [73] analyzes a price-skimming model with rational customers under the assumption that the firm can commit to a price schedule. The material in §3.7.2 on the subgame-perfect pricing equilibrium for a firm faced with strategic customers is from Besanko and Winston [11].

Acknowledgment

This tutorial chapter was published originally as *The Theory and Practice of Revenue Management* by Kalyan T. Talluri and Garrett J. van Ryzin [75]. Portions of the original are reprinted here with permission of Springer Science + Business Media.

References

[1] S. Algers and M. Besser. Modeling choice of flight and booking class: A study using stated preference and revealed preference data. *International Journal of Services Technology and Management* 2:28–45, 2001.

[2] S. E. Andersson. Operational planning in airline business—Can science improve efficiency? Experiences from SAS. *European Journal of Operations Research* 43:3–12, 1989.

[3] P. P. Belobaba. Air travel demand and airline seat inventory management. Ph.D. thesis, Flight Transportation Laboratory, MIT, Cambridge, MA, 1987.

[4] P. P. Belobaba. Airline yield management: An overview of seat inventory control. *Transportation Science* 21:63–73, 1987.

[5] P. P. Belobaba. Application of a probabilistic decision model to airline seat inventory control. *Operations Research* 37:183–197, 1989.

[6] P. P. Belobaba. Optimal vs. heuristic methods for nested seat allocation. *ORSA/TIMS Joint National Meeting*, San Francisco, CA (November) 1992.

[7] P. P. Belobaba and L. R. Weatherford. Comparing decision rules that incorporate customer diversion in perishable asset revenue management situations. *Decision Sciences* 27:343–363, 1996.

[8] M. Ben-Akiva and S. Lerman. *Discrete-Choice Analysis: Theory and Application to Travel Demand*. MIT Press, Cambridge, MA, 1985.

[9] D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Athena Scientific, Belmont, MA, 1996.

[10] D. P. Bertsekas. *Nonlinear Programming*, 2nd ed. Athena Scientific, Belmont, MA, 1999.

[11] D. Besanko and W. L. Winston. Optimal price skimming by a monopolist facing rational consumers. *Management Science* 36:555–567, 1990.

[12] R. R. Bhat. *Managing the Demand for Fashion Items*. UMI Research Press, Ann Arbor, MI, 1985.

[13] A. V. Bhatia and S. C. Parekh. Optimal allocation of seats by fare. Presentation to AGIFORS Reservations Study Group, Trans World Airlines, 1973.

[14] G. R. Bitran and S. V. Mondschein. Periodic pricing of seasonal products in retailing. *Management Science* 43:61–79, 1997.

[15] G. R. Bitran, R. Caldentey, and S. V. Mondschein. Coordinating clearance markdown sales of seasonal products in retail chains. *Operations Research* 46:609–624, 1998.

[16] C. H. Bohutinsky. The sell-up potential of airline demand. Master's thesis, Flight Transportation Lab, MIT, Cambridge, MA, 1990.

[17] S. L. Brumelle and J. I. McGill. Airline seat allocation with multiple nested fare classes. *Operations Research* 41:127–137, 1993.

[18] S. Brumelle and D. Walczak. Dynamic airline revenue management with multiple semi-Markov demand. *Operations Research* 51:137–148, 2003.

[19] S. L. Brumelle, J. I. McGill, T. H. Oum, K. Sawaki, and M. W. Tretheway. Allocation of airline seat between stochastically dependent demands. *Transportation Science* 24:183–192, 1990.

[20] P. Coy. The power of smart pricing: Companies are fine-tuning their price strategies—and it's paying off. *Business Week* (April 10):160–164, 2000.

[21] R. G. Cross. *Revenue Management: Hardcore Tactics for Market Domination.* Broadway Books, New York, 1997.

[22] R. E. Curry. Optimal airline seat allocation with fare classes nested by origins and destinations. *Transportation Science* 24:193–204, 1990.

[23] G. Das Varma and N. Vettas. Optimal dynamic pricing with inventories. *Economics Letters* 72:335–340, 2001.

[24] S. Daudel and G. Vialle. *Yield Management: Applications to Air Transport and Other Service Industries.* Les Presses de L'Institut du Transport Aerien, Paris, France, 1994.

[25] S. V. de Boer. Advances in airline revenue management and pricing. Ph.D. thesis, Sloan School of Management, MIT, Cambridge, MA, 2003.

[26] J. Eliashberg and R. Steinberg. Marketing-production joint decision making. J. Eliashberg and J. D. Lilien, eds. *Management Science in Marketing, Handbooks in Operations Research and Management Science.* North Holland, Amsterdam, The Netherlands, 1991.

[27] W. J. Elmaghraby and P. Keskinocak. Dynamic pricing in the presence of inventory considerations: Research overview, current practices, and future directions. *Management Science* 49:1287–1309, 2003.

[28] Y. Feng and G. Gallego. Optimal starting times for end-of-season sales and optimal stopping times for promotional fares. *Management Science* 41:1371–1391, 1995.

[29] Y. Feng and G. Gallego. Perishable asset revenue management with Markovian time dependent demand intensities. *Management Science* 46:941–956, 2000.

[30] Y. Feng and B. Xiao. Maximizing revenue of perishable assets with a risk factor. *Operations Research* 47:337–341, 1999.

[31] Y. Feng and B. Xiao. A continuous-time yield management model with multiple prices and reversible price changes. *Management Science* 46:644–657, 2000.

[32] Y. Feng and B. Xiao. Optimal policies of yield management with multiple predetermined prices. *Operations Research* 48:332–343, 2000.

[33] S. C. Friend and P. H. Walker. Welcome to the new world of merchandising. *Harvard Business Review* 79 (November) 2001.

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

[35] G. Gallego and G. J. van Ryzin. A multi-product dynamic pricing problem and its applications to network yield management. *Operations Research* 45:24–41, 1997.

[36] Y. Gerchak and M. Parlar. A single period inventory problem with partially controlled demand. *Computers and Operations Research* 14:1–9, 1987.

[37] Y. Gerchak, M. Parlar, and T. K. M. Ye. Optimal rationing policies and production quantities for products with several demand classes. *Canadian Journal of Administration Science* 2:161–176, 1985.

[38] G. Girard. Revenue management: The price can't be right if the tools aren't. Technical report, AMR Research Inc., Boston, MA (September) 2000.

[39] A. Y. Ha. Stock-rationing policy for a make-to-stock production system with two priority classes and backordering. *Naval Research Logistics* 44:457–472, 1997.

[40] A. Heching, G. Gallego, and G. J. van Ryzin. Markdown pricing: An empirical analysis of policies and revenue potential at an apparel retailer. *Journal of Pricing and Revenue Management* 1:139–160, 2002.

[41] A. L. Hempenius. *Monopoly with Random Demand.* Rotterdam University Press, Rotterdam, The Netherlands, 1970.

[42] A. Ingold, U. McMahon-Beattie, and I. Yeoman, eds. *Yield Management: Strategies for the Service Sector*, 2nd ed. Continuum, London, UK, 2000.

[43] S. S. Iyengar and M. Lepper. When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology* 76:995–1006, 2000.

[44] D. Jenkins, ed. *Handbook of Airline Economics.* McGraw-Hill, New York, 1995.

[45] C. A. Johnson. Retail revenue optimization: Timely and rewarding. Technical report, Forrester Research Inc., Cambridge, MA (July 23) 2001.

[46] C. Kalish. Monopolist pricing with dynamic demand and production costs. *Marketing Science* 2:135–159, 1983.

[47] A. Kaplan. Stock rationing. *Management Science* 15:260–267, 1969.

[48] S. Karlin and C. R. Carr. Prices and optimal inventory policies. K. J. Arrow, S. Karlin, and H. Scarf, eds. *Studies in Applied Probability and Management Science*. Stanford University Press, California, CA, 1962.

[49] S. E. Kimes. Yield management: A tool for capacity-constrained service firms. *Journal of Operations Management* 8:348–363, 1989.

[50] W. M. Kincaid and D. Darling. An inventory pricing problem. *Journal of Mathematical Analysis and Applications* 7:183–208, 1963.

[51] C. J. Lautenbacher and S. J. Stidham. The underlying Markov decision process in the single-leg airline yield management problem. *Transportation Science* 34:136–146, 1999.

[52] E. P Lazear. Retail pricing and clearance sales. *American Economic Review* 76:14–32, 1986.

[53] T. C. Lee and M. Hersh. A model for dynamic airline seat inventory control with multiple seat bookings. *Transportation Science* 27:252–265, 1993.

[54] Y. Liang. Solution to the continuous time dynamic yield management model. *Transportation Science* 33:117–123, 1999.

[55] K. Littlewood. Forecasting and control of passenger bookings. *Proceedings of the 12th Annual AGIFORS Symposium*, Nathanya, Israel, 1972.

[56] C. Maglaras and J. Meissner. Dynamic pricing strategies for multi-product revenue management problems. Decision, Risk and Operations Research Division Working Paper DRO-2003-10, 2003.

[57] J. I. McGill. Optimization and estimation problems in airline yield management. Ph.D. thesis, Faculty of Commerce and Business Administration, University of British Columbia, Vancouver, Canada, 1989.

[58] J. I. McGill and G. J. van Ryzin. Revenue management: Research overview and prospects. *Transportation Science* 33:233–256, 1999.

[59] A. Merrick. Priced to move: Retailers attempt to get a leg up on markdowns with new software. *Wall Street Journal* (April 7), 2001.

[60] E. S. Mills. Uncertainty and price theory. *Quarterly Journal of Economics* 73:117–130, 1959.

[61] N. T. Nagle and R. K. Holden (contributor). *The Strategy and Tactics of Pricing: A Guide to Profitable Decision Making*, 2nd ed. Prentice-Hall, Englewood Cliffs, NJ, 1994.

[62] I. C. Paschalidis and J. N. Tsitiklis. Congestion-dependent pricing of network services. *IEEE/ACM Transactions on Networking* 8:171–184, 2000.

[63] P. P. B. Pashigan. Demand uncertainty and sales. *American Economic Review* 78:936–953, 1988.

[64] P. P. B. Pashigan and B. Bowen. Why are products sold on sale? Explanations of pricing regularities. *Quarterly Journal of Economics* 106:1015–1038, 1991.

[65] N. C. Petruzzi and M. Dada. Pricing and the newsvendor problem: A review with extensions. *Operations Research* 47:183–194, 1999.

[66] S. Polt. Back to the roots: New results on leg optimization. *1999 AGIFORS Reservations and Yield Management Study Group Symposium*, London, UK, 1999.

[67] L. W. Robinson. Optimal and approximate control policies for airline booking with sequential nonmonotonic fare classes. *Operations Research* 43:252–263, 1995.

[68] S. M. Shugan. Editorial: Marketing science, models, monopoly models, and why we need them. *Marketing Science* 21:223–228, 2002.

[69] B. C. Smith, J. F. Leimkuhler, and R. M. Darrow. Yield management at American Airlines. *Interfaces* 22:8–31, 1992.

[70] S. A. Smith and D. D. Achabal. Clearance pricing and inventory policies for retail chains. *Management Science* 44:285–300, 1998.

[71] W. Stadje. A full information pricing problem for the sale of several identical commodities. *Zeitschrift für Operations Research* 34:161–181, 1990.

[72] N. Stokey. Intertemporal price descrimination. *Quarterly Journal of Economics* 94:355–371, 1979.

[73] N. Stokey. Rational expectations and durable goods pricing. *Bell Journal of Economics* 12:112–128, 1981.

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

[75] K. T. Talluri and G. J. van Ryzin. *The Theory and Practice of Revenue Management*. Springer Science + Business Media, Berlin, Germany, 2004.

[76] G. T. Thowsen. A dynamic, nonstationary inventory problem for a price/quantity setting firm. *Naval Research Logistics* 22:461–476, 1975.

[77] B. Titze and R. Griesshaber. Realistic passenger booking behaviors and the simple low-fare/high-fare seat allotment model. *Proceedings of the 23rd Annual AGIFORS Symposium*, 1983.

[78] D. M. Topkis. Optimal ordering and rationing policies in a nonstationary dynamic inventory model with n demand classes. *Management Science* 15:160–176, 1968.

[79] G. J. van Ryzin. The brave new world of pricing. Survey: Mastering management. *Financial Times* (October 16) 2000.

[80] E. J. Warner and R. B. Barsky. The timing and magnitude of retail store markdowns: Evidence from weekends and holidays. *Quarterly Journal of Economics* 110:321–352, 1995.

[81] L. R. Weatherford and S. E. Bodily. A taxonomy and research overview of perishable-asset revenue management: Yield management, overbooking, and pricing. *Operations Research* 40:831–844, 1992.

[82] L. R. Weatherford, S. E. Bodily, and P. E. Pfeifer. Modeling the customer arrival process and comparing decision rules in perishable asset revenue management situations. *Transportation Science* 27:239–251, 1993.

[83] T. M. Whitin. Inventory control and price theory. *Management Science* 2:61–68, 1955.

[84] R. D. Wollmer. An airline seat management model for a single leg route when lower fare classes book first. *Operations Research* 40:26–37, 1992.

[85] P. S. You. Dynamic pricing in airline seat management for flights with multiple legs. *Transportation Science* 34:192–206, 1999.

[86] Zabel, E. Muti-Period Monopoly Under Uncertainty. *Journal of Economic Theory* **5** 524–546, 1972.

[87] W. Zhao and Y.-S. Zheng. Optimal dynamic pricing for perishable assets with nonhomogeneous demand. *Management Science* 46:375–388, 2000.