Chapter 4

OVERBOOKING

Overbooking is somewhat distinct from the core pricing and capacity-control problems of revenue management. RM is mainly concerned with how best to price or allocate capacity—how to achieve the best *mix* of demand, in essence. In contrast, overbooking is concerned with increasing capacity utilization in a reservation-based system when there are significant cancellations.[1] Its focus is increasing the total *volume* of sales in the presence of cancellations rather than optimizing customer mix. The problems of optimizing demand mix and volume are quite related, however, and both are considered integral parts of RM.

Indeed, from a historical standpoint, overbooking is the oldest—and, in financial terms, among the most successful—of RM practices. In the airline industry it is estimated that approximately 50% of reservations result in cancellations or no-shows[2] and about 15% of all seats would go unsold without some form of overbooking. [477] This is to be compared to fare-class allocation, which by most estimates leads to incremental revenues to the order of 5%. Despite its economic importance, many researchers consider overbooking a somewhat mature area, and it has

[1]We note, however, that reservations are not used in all quantity-based RM industries. In certain advertising markets, for example, one advertiser is allowed to preempt another if it'is willing to pay more for the same ad slot. This effectively produces an auction in which the current highest bidder has claim to the capacity.

[2]A *cancellation* is defined as a reservation that is terminated by a customer strictly *prior* to the time of service. A no-show, in contrast, occurs when a customer does not cancel his reservation but rather just fails to show up at the time of service. The distinction is important because the firm has some opportunity to compensate for a cancellation by accepting more reservations after the fact, while no such opportunities exist when a customer no-shows.

received less attention in the recent research literature than fare-class allocation or pricing.

As a business practice, the biggest challenges in overbooking are managing the negative effects of denying service on customer relations and dealing with the resulting legal and regulatory issues. On a planning level, overbooking involves controlling the level of reservations to balance the potential risks of denied service against the rewards of increased sales. Theoretically, this involves controlling parameters of a probability distribution, which introduces somewhat unique methodology that is not encountered in other areas of RM.

4.1 Business Context and Overview

A reservation is essentially a forward contract between a customer and the firm. Reservations give customers the right to use a service in the future at a fixed price and often also the option to opt out (perhaps with a penalty) before the time of service.

Customers value reservations whenever the costs of unavailability at the desired time of consumption are higher than the costs of unavailability prior to the time of consumption. For instance, because customers travel to attend business meetings, visit family members, or take vacations, they must coordinate their travel with hotel arrangements, business appointments, scheduling of vacation days, and so on. Since it is generally more costly to change or renege on these contingent arrangements at the time of service than it is to change or renege on them in advance, customers value reservations for travel services.

Yet committing to purchasing in advance has its own risks. Uncertain future events (such as clients rescheduling meetings, illness, or more attractive vacation opportunities) may make it impossible or undesirable to use the service. Therefore, customers also value the option to cancel reservations. Indeed, a reservation with a cancellation option gives customers the best of both worlds—the benefit of locking-in availability in advance and the flexibility to renege should their plans or preferences change.

While advance reservations with a cancellation option are highly valued by customers, they require a firm to take a two-sided risk—to honor the reservation when customers show up (or provide suitable compensation if it cannot honor the reservation), and in cases where customers cancel or do not show up, to bear the opportunity cost of wasted capacity. Firms try to manage this risk through a combination of cancellation penalties and overbooking.

Cancellation and no-show penalties effectively allow customers and the firm to share the risks of cancellations. In practice, penalties for

cancellations and no-shows range from zero to full price and are most often implemented as a sale condition of the product. In fact, the potential for abuse without such penalties is substantial; customers may make multiple reservations to preserve various options and then cancel all of them except the one they want, not an uncommon practice in certain wholesale international air travel markets. Some minimal penalty is necessary to curb such abuses. (Though firms in a surprising number of industries—the restaurant industry for one—do not penalize for cancellations.) On the other hand, if the penalties are too large, the cancellation option has little value or effectively becomes a nonoption for customers.

To further reduce the costs of cancellations, a firm may also adopt a strategy of accepting more reservations than it has capacity to serve, taking the chance that the number of surviving reservations will be within capacity. This is the essence of planned overbooking.

A firm that chooses a strategy of planned overbooking is immediately faced with several important problems. One is confronting the legal and regulatory implications of failing to honor the reservation contract. Even if the firm is on safe legal ground, it must have operational policies and procedures in place to deal with the situation in which service must be denied. Once these basic structural and policy elements are determined, it must develop methodology to control the level of overbooking on an operational basis. We look at each of these issues in turn.

4.1.1 A History of Legal Issues in Airline Overbooking

Legally, overbooking involves the risk of failing to deliver on a contract to provide service. While there are somewhat different legal requirements in each industry in this regard, it is instructive to look at the evolution of airline overbooking regulations in the United States as an example of the legal issues involved.

Prior to 1961, intentional overbooking was practiced somewhat clandestinely by U.S. airlines and was not acknowledged publicly. Despite this fact, Rothstein [449] reports that as director of Operations Research at American Airlines he, "found much publicly available evidence that all the major airlines were deliberately overbooking." In 1961, the Civil Aeronautics Board (CAB) reported a no-show rate of 1 out of every 10 passengers booked among the 12 leading carriers at that time. The CAB acknowledged that this situation created real economic problems for the airlines. As a result, the CAB implemented a no-show penalty of 50% of the ticket price. At the same time, they explicitly required airlines to pay a penalty of 50% of the ticket price to passengers who

were denied boarding. However, the CAB still did not officially sanction overbooking practices. The no-show penalty was abandoned in 1963 largely because airline management felt that the penalties created ill will among passengers and might be discouraging air travel in general.

The CAB conducted another study of overbooking during 1965-66. They found that the denied-boarding rate at that time was approximately 7.69 per 10,000 passengers boarded [119]. Their conclusion was issued in a 1967 docket [119]:

> There is a substantial reservation turnover before flight time from cancellations and no-shows. The airlines are engaging in deliberate or controlled overbooking to compensate for it. Through carefully controlled overbooking, the airlines can reduce the number of empty seats and at the same time serve the public interest by accommodating more passengers.
>
> The present reservation systems of the carriers greatly benefit the traveling public. The Board is not prepared, therefore, to require changes in these systems.

Thus, as of 1965, overbooking was an officially sanctioned practice, provided it was "carefully controlled," a criterion that was never precisely defined by the CAB.

In parallel, the CAB also increased the denied-boarding penalty to 100% of the coupon. Airlines controlled the percentage of denied boardings, and the CAB carefully monitored the denied-boarding performance of each airline. The involuntary denied-boarding rate is still carefully monitored in the United States by the Department of Transportation (DOT) and currently hovers around 0.5 to 1.5 involuntary and 15 to 20 voluntary denied boardings per 10,000 passengers (see Table 4.1).

Despite this progress in formalizing the practice of planned overbooking, the traveling public was still largely unaware of its existence. This was to change in 1972 when Ralph Nader, the well-known U.S. consumer advocate, was denied boarding on an Allegheny Airlines flight. Rather than accept the standard compensation, he sued Allegheny, won, and was awarded $25,000 in punitive damages. The judge's ruling was based on the fact that Allegheny did not advise passengers of its practice of deliberate overbooking. The case was appealed all the way to the U.S. Supreme Court, but the ruling was upheld. As Rothstein [448] noted at the time:

> ... public policy may very well force the airlines into the position that a reservation *involves a definite, legal claim on a seat.* And if this happens, most of the operations research carried out on this problem will have to be discarded and eventually redone.

Table 4.1. U.S. major airline denied-boarding rates, 1990-2000.[a,b,c]

	1990	*1991*	*1992*	*1993*	*1994*	*1995*	*1996*	*1997*	*1998*	*1999*	*2000*	*2001*	*2002*
Boarded ($\times 10^6$)	421	428	445	449	457	460	481	503	514	523	543	498	467
Total DB ($\times 10^3$)	628	646	764	683	824	843	957	1,072	1,126	1,070	1,119	941	837
Voluntary	561	599	718	632	771	794	899	1,018	1,081	1,024	1,062	898	803
Involuntary	67	47	46	51	53	49	58	54	45	46	55	43	34
VDB rate (per 10^4*)*	14.92	15.09	17.17	15.21	18.03	18.33	19.90	21.31	21.03	19.58	20.60	18.89	17.92
IDB rate (per 10^4*)*	1.59	1.10	1.03	1.14	1.16	1.07	1.21	1.07	0.87	0.88	1.05	0.86	0.72

[a] Data are for nonstop scheduled service flights between points within the United States (including territories) by the 10 largest U.S. air carriers—that is, those with at least 1% of total domestic scheduled-service passenger revenues (Alaska, America West, American, Continental, Delta, Northwest, Southwest, TWA, United, and US Airways). Before 1994, carriers included both major and national airlines—that is, airlines with over $100 million in revenue.

[b] Statistics are the number of passengers who hold confirmed reservations and are denied boarding ("bumped") from a flight because it is oversold. These figures include only passengers whose oversold flight departs without them; they do not include passengers affected by cancelled, delayed, or diverted flights.

[c] *Source:* U.S. Department of Transportation, Office of the Secretary, *Air Travel Consumer Report* (Washington, DC: Annual March issues).

NOTICE: OVERBOOKING OF FLIGHTS

Airline flights may be overbooked, and there is a slight chance that a seat will not be available on a flight for which a person has a confirmed reservation. If the flight is overbooked, no one will be denied a seat until airline personnel first ask for volunteers willing to give up their reservation in exchange for a payment of the airline's choosing. If there are not enough volunteers, the airline will deny boarding to other persons in accordance with its particular boarding priority. With few exceptions, persons denied boarding involuntarily are entitled to compensation. The complete rules for the payment of compensation and each airline's boarding priorities are available at all airport ticket counters and boarding locations. Some airlines do not apply these consumer protections to travel from some foreign countries, although other consumer protections may be available. Check with your airline or your travel agent.

Figure 4.1. Overbooking notification statement required by Department of Transportation on domestic U.S. airline tickets [523].

Rothstein proposed that airlines charge for reservations as a possible solution:

> In other words, the *reservation* itself, as opposed to the physical seat on the plane, is to be considered of value ... the reservation itself is a commodity to be purchased for an amount of money and possibly to be relinquished for a different amount of money.

In the wake of Mr. Nader's suit—and after much debate—the CAB revised its rules concerning overbooking as follows:

- Denied-boarding compensation was doubled again to 200% of the coupon.
- Airlines were required to seek volunteers first before denying boarding to any passenger involuntarily.
- The traveling public was to be notified of the deliberate overbooking practices of the airlines.
- A statement warning passengers that their flight may be overbooked and informing them of their rights was to be printed on every ticket.

As a result of this ruling, the DOT requires an overbooking notification statement on all U.S. airline tickets (see Figure 4.1).

These basic regulations are still in existence today in the U.S. Since deregulation in 1974, airlines have increasingly relied on vouchers and payments to attract volunteers to give up their seats on oversold flights. As a result, involuntary denied boardings are much less frequent today than they were in the days when overbooking was a clandestine practice.

4.1.2 Managing Denied-Service Occurrences

Managing the compensation and selection of customers in the event of oversales can have a significant impact on denied-service costs and customer perceptions of overbooking. We next briefly look at the main issues involved in managing oversales.

4.1.2.1 Compensation for Denied Service

While legally mandated compensation often specifies payment of monetary damage, this is often viewed as inadequate in the eyes of customers. A car rental customer who is planning to take a tour of the California coast would most likely find the prospect of getting a full refund plus 50% of the contracted rental rate as poor compensation for a ruined vacation. It is often more effective to offer customers a substitute service (such as an upgrade) plus ancillary services that may make the short-run disruption in their schedule more palatable.

To illustrate, the same car-rental customer may be much more satisfied with an offer to provide a ride to a competing rental company, a free upgrade to a luxury car, plus a voucher for future rentals. Compensation that is targeted to substitute for the denied service and perhaps enhance it somewhat is frequently less expensive for a firm and more effective in the eyes of the customer than pure monetary compensation.

4.1.2.2 Selection Criteria

Selecting which customers are to be denied service also can have a significant impact on both the firm's direct costs and customer goodwill. Prom a legal standpoint, such selection must not be discriminatory. For example, for airlines, current DOT regulations state that [523]

> Every carrier shall establish priority rules and criteria for determining which passengers holding confirmed reserved space shall be denied boarding on an oversold flight in the event that an insufficient number of volunteers come forward.
>
> Such rules and criteria shall not make, give, or cause any advantage to any particular person or subject any particular person to any unjust or unreasonable prejudice or disadvantage in any respect whatsoever.

The default option for allocating service to customers is usually to do it on a first-come, first-serve (FCFS) basis. While a FCFS allocation is perceived as fair and encourages customers to arrive on time, there are many business situations where this allocation is quite undesirable.

A good example is hotel overbooking. Using a FCFS allocation for a hotel means that the customers who are denied service are those who arrive very late in the evening. This creates two difficulties. First, it

is usually much more disruptive to relocate a customer who arrives late at night. These customers are often tired and irritable and simply want to go to bed as soon as possible. A customer who arrives in the late afternoon, in contrast, may be willing to sightsee around the town for several hours or go out for a meal while alternative accommodations are secured and their baggage is transported. Second, late-arriving hotel customers are typically business travelers who pay the highest rates, travel often, and therefore represent the most profitable segment for most hotels. In terms of the lifetime value, these customers are the most costly to lose.

Hotels therefore do not always allocate rooms on a FCFS basis. Rather, they monitor arrival rates and occupancy throughout the day to anticipate a potential oversold condition. If at some point managers expect an oversale, they may find alternative arrangements for early-arriving customers to avoid denying service to customers due to check in very late.

In other service situations, it is sometimes possible to select among a pool of customers when allocating service. For example, in airline boarding, customers usually gather to the gate before departure. This gives gate agents a chance to see which passengers have arrived for the flight and to selectively target specific passengers for denied-boarding offers. Indeed, we are aware of one Australian airline that trains its gate agents to solicit young, student travelers ("backpackers") as volunteers. The airline found that customers in this segment are eager to receive a nice hotel room and a good meal in exchange for taking a flight the following day.

4.1.2.3 Oversale Auctions

An alternative method of managing oversales is to conduct an auction to attract volunteers to give up their reservations in exchange for monetary or other compensation. While this practice is now widespread in airlines and familiar to most travelers, the idea was not well received initially.

In 1968, economist Julian Simon proposed what he called "an almost practical solution to airline overbooking, " in which airlines would conduct a sealed-bid "reverse auction" to find passengers willing to accept monetary compensation for being bumped. Simon predicted (rightly so, as initial responses to his letters to airline executives later indicated) that the airlines would object to the scheme

> because such an auction does not seem decorous; it smacks of the pushcart rather than the one-price store; it is "embarrassing" and "crass," i.e., frankly commercial, like 'being in trade' in Victorian England ([472]).

Simon cites this tongue-in-cheek reaction from airline executive Blaine Cooke:

> I greatly fear that your Overbooking Auction Plan suffers from a flawed premise and a fatal defect. The flawed premise is that you assume that airline management and regulation is a rational exercise. It is not. It is more accurately described as an exercise in applied insanity. The effect is your plan offers a market-sensitive and sensible solution to a real problem but a solution not conceived by an airline. Accordingly, the idea must be disallowed since it is well established in airline marketing that only ideas which originate within the airline industry are permissible.

Simon wrote many letters to executives, regulators, policy makers, and customer groups arguing for his "oversale-auction" idea. Despite these efforts, he failed to get even one airline to experiment with it on even a single flight. Even prominent fellow economists questioned the practicality of the idea. Simon [474] quotes a letter from Milton Friedman:

> If the plan is as good as you and I think it is, I am utterly baffled by the unwillingness of one or more of the airlines to experiment with it. I conclude that we must be overlooking something. I realize that you have tested this quite exhaustively, and I have no reason to question your results; yet I find it even harder to believe that opportunities for large increments of profit are being rejected for wholly irrational reasons.

The scheme continued to flounder until 1977 when Alfred Kahn, an economist, was appointed to head the CAB. Simon wrote to Kahn about his proposal and Kahn liked and largely adopted it under the heading of a "volunteer" denied-boarding plan, as mentioned above. At the same time, Kahn increased penalties for involuntary denied boardings.

Simon [474] quotes an American Airlines internal newsletter from April 27, 1979:

> The happiest result of the volunteer plan is that airlines now have a fair and efficient way to avoid denying seats to people who for business or personal reasons have a pressing need to make their flights as planned. VP, Passenger Services, Robert H. Phillips points out that the voluntary program has twin virtues: "It enables us to reduce costs while maintaining customer goodwill and thereby protecting future revenue".

Given the success of this volunteer denied-boarding plan, it appears that airline management has indeed inched closer to the notion of a "rational exercise."

4.1.3 Lessons Beyond the Airline Industry

There are several broader lessons to be learned from this history of airline overbooking. One is that it takes time for customers to get

used to and accept overbooking practices, and providers in turn have to learn how to develop strategies and operational practices that make overbooking as painless as possible for customers. In the airline industry, this process took decades to develop. A second lesson is that some seemingly fanciful techniques—in particular the oversale auction—can in fact prove to be surprisingly popular and effective in practice, which serves as a caution for those who are quick to criticize such innovations. Finally, while there is no denying that overbooking is a well-developed and refined practice in the airline and hotel industries, it nevertheless remains a primary source of dissatisfaction for customers. Overbooking is frequently cited in customer complaints, both to individual firms and to government regulators. So even at its best, overbooking is a somewhat awkward compromise between economic efficiency and service quality.

4.2 Static Overbooking Models

We next look at the methodology for making overbooking decisions. The simplest and most widely used methodology is based on *static overbooking models.* In static models, the dynamics of customer cancellations and new reservation requests over time are ignored. Rather, the models simply determine the maximum number of reservations to hold at the current time given estimates of cancellation rates from the current time until the day of service. This maximum number of reservations, or *overbooking limit,* is then recomputed periodically prior to service to reflect changing state and cancellation probabilities over time. While more sophisticated dynamic overbooking models have been developed and are discussed in Section 4.3, the simplicity, flexibility, and robustness of the simpler static models have made them more popular in practice.

Two types of events impact the overbooking decision—*cancellations* and *no-shows*—with the difference simply related to the timing of the events. (Again, a cancellation is a reservation that is withdrawn by a customer strictly prior to the time of service; a no-show is someone who does not cancel and does not show up at the time of service.) While both result in a situation where a reservation does not "survive" to the time of service, with a cancellation, the firm has an opportunity to possibly replace the cancelled reservation; in contrast, there is little recourse available to compensate for a no-show. Under a static model, the distinction between the two is unnecessary, since a static model assumes a static overbooking limit is set without recourse to adjust it. Thus, all that matters is the probability that a reservation survives to the time of service (the *show demand,* as it is called). In dynamic overbooking models, however, the distinction between no-shows and cancellations is important.

In airline and hotel practice, static models are used to compute overbooking limits—also called *virtual capacities* or *overbooking authorization levels* in the airline industry—which are, in turn, used as inputs to capacity-allocation models. These static overbooking models are typically re-solved periodically to account for changes in the cancellation and no-show probabilities over time, resulting in overbooking limits that vary (typically decline) over time. The current overbooking limit gives the maximum number of reservations one will accept at any time.

The situation is illustrated in Figure 4.2. The top, wide line is the overbooking limit over time. Solving a static model gives one point on this curve. Overbooking limits are set high initially because the probability of a reservation cancelling prior to the time of service or no-showing is usually higher the longer the time till service. As the time of service (T) approaches, the overbooking limits fall. At the same time, reservations are being accumulated in the system over time. The dark line in Figure 4.2 shows that with overbooking in place, the reservations in the system can exceed the capacity C, and we don't stop accepting reservations until the overbooking limit is reached. At that point reservations are rejected. The resulting show demand (demand that shows up finally) at time T is ideally close to the capacity C. The lower line shows the same trajectory of reservations without overbooking. In this case, the reservations in the system are truncated at the capacity C early on in the booking process. As a result, once reservations start to cancel and no-show, the show demand is significantly less than capacity.

4.2.1 The Binomial Model

The simplest static model is based on a *binomial model* of cancellations in which no-shows are lumped together with cancellations (that is, a no-show is treated simply as a cancellation that occurs at the day of service). The following assumptions are made:

- Customers cancel independently of one another.
- Each customer has the same probability of cancelling.
- The cancellation probability is Markovian; it depends only on the time remaining to service and is independent of the age of the reservation.

Let t denote the time remaining until service, C denote the physical capacity, y denote the number of reservations on hand, and q the probability that a reservation currently on hand shows up at the time of service ($1 - q$ is the probability that customers cancel prior to the time

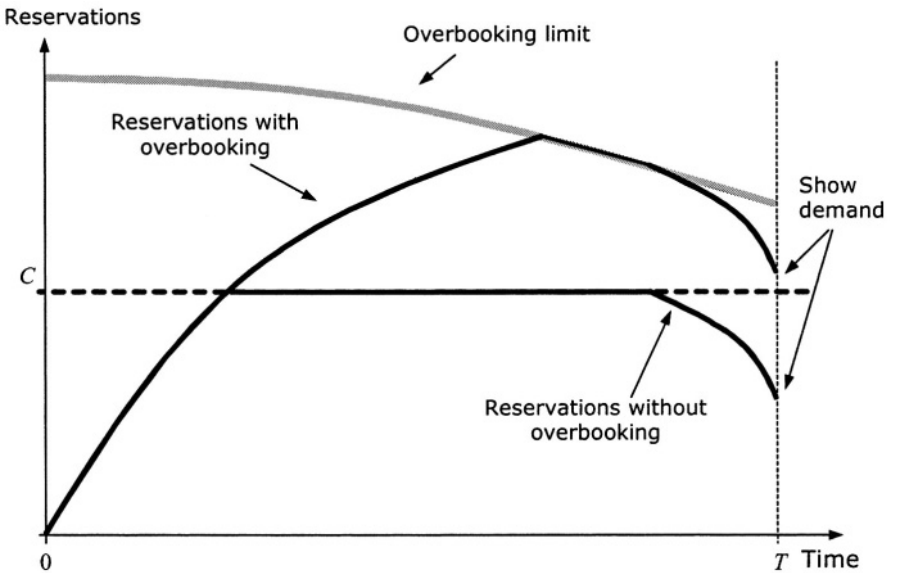

Figure 4.2. Illustration of overbooking limits and reservations over time.

of service). Note that q is really a function of the time remaining, since in general the more time remaining the more likely it is that customers cancel before the time of service. However, to keep the notation simple we suppress the dependence of q on t. Also, in practice estimates of q may be based on the ratio of show demand to reservations on hand (the net bookings) rather than on individual customer cancellation rates; this approach is discussed further in Section 4.3.2.

Under the assumptions stated above, the number of customers who show up at the time of service given there are y reservations on hand, denoted $Z(y)$ (the *show demand*), is binomially distributed with p.m.f.:

$$\begin{aligned} P_y(z) &= P(Z(y) = z) \\ &= \binom{y}{z} q^z (1-q)^{y-z}, \quad z = 0, 1, \ldots, y, \end{aligned} \tag{4.1}$$

and with c.d.f.:

$$\begin{aligned} F_y(z) &= P(Z(y) \leq z) \\ &= \sum_{k=0}^{z} \binom{y}{k} q^k (1-q)^{y-k}, \end{aligned} \tag{4.2}$$

with mean $E[Z(y)] = qy$ and variance $Var(Z(y)) = yq(1-q)$. It is convenient to work with the complement of the distribution F_y, denoted

by $\bar{F}_y$, which is defined by

$$\bar{F}_y(z) = P(Z(y) > z) = 1 - F_y(z).$$

Several airline industry studies have validated this binomial model of cancellations. For example, in one of the earliest investigations of overbooking, Thompson [508] considers data from 59 flights from Auckland to Sydney operated by Tasman Empire Airways. He eliminated groups of six or more since they exhibited much lower cancellation rates and although rare (11 total booking on the 59 flights), can significantly distort the cancellation rate on the flights involved. Parties of six or fewer constituted 99.6% of all bookings; 81% of the remaining were singles; 15% were paired and 4% were parties of three to six. (See Table 4.4.) While the results showed that group-cancellation behavior does invalidate the binomial model for certain cabins on certain flights, overall he concluded that the binomial model adequately fits the data. (Group-cancellation effects are discussed further in Section 4.2.4.)

4.2.1.1 Overbooking Based on Service-Level Criteria

One measure of service is the probability of oversale at the time of service, which we call the *Type 1 service level.* We assume the firm uses an *overbooking-limit policy* to control the number of reservations that are accepted. The overbooking limit is denoted x. In other words, we assume the firm continues to accept reservations as long as the number of reservations on hand y is less than x and stops accepting reservations once $y = x$.[3]

The Type 1 service level is denoted $s_1(x)$ and is given by

$$s_1(x) = \bar{F}_x(C) = P(Z(x) > C).$$

That is, if we assume that the number of reservations on hand y reaches our overbooking limit x, then $s_1(x)$ will be the probability that we have to deny service to one or more customers. Hence, setting an overbooking limit of x guarantees that the probability of oversale will not exceed $s_1(x)$.

An arguably more natural measure of service is the long-run fraction of customers who are denied service, which we call the *Type 2 service*

[3]Whether such a threshold policy is in any sense optimal for the dynamic decision-making problem is addressed in Section 4.3. Here we simply assume such an overbooking policy is used.

level denoted by $s_2(x)$. This fraction is given by[4]

$$\begin{aligned} s_2(x) &= \frac{E[(Z(x)-C)^+]}{E[Z(x)]} \\ &= \frac{\sum_{k=C+1}^{x}(k-C)P_x(k)}{x(1-q)}. \end{aligned}$$

Setting an overbooking limit of x ensures that, at most, a fraction $s_2(x)$ of customer will be denied service. Through some algebraic simplification, one can show that

$$s_2(x) = \bar{F}_{x-1}(C-1) - \frac{C}{qx}\bar{F}_x(C), \tag{4.3}$$

which is a more convenient formula for computations.

Table 4.2 shows the Type 1 and Type 2 service levels for an example with $C = 150$, $q = 0.85$, and varying overbooking limit x. In practice, we first specify a service level and then numerically search for the largest booking level x^* satisfying the specified service level. The resulting x^* is the overbooking limit. The quantity $x^* - C$ (the excess over capacity) is referred to as the *overbooking pad.*

Example 4.1 Suppose we want no more than 0.1% of customers to be denied service and our capacity is $C = 150$ and $q = 0.85$. From Table 4.2, we should accept at most 168 reservations ($x^* = 168$), since this is the largest value for which $s_2(x)$ less than .001, (though 169 has a service level only slightly over the standard and might be a candidate as well). Reservations would then be accepted as long as the number of bookings on hand is less than 168. The overbooking pad would be 168 – 150 = 18.

Note that if we do not receive at least x^* requests for reservations, the service levels will in fact be higher than $s_1(x^*)$ and $s_2(x^*)$. In other words, these measures predict the service level for instances in which demand exceeds x^*, but the service level will be higher if demand is strictly less than x^*. So effectively, we are considering a worst-case service level (demand exceeding the overbooking limit) rather than an average-case service level.

[4]As a technical aside, note that one may be tempted to define the Type 2 service level as

$$E\left[\frac{(Z(x)-C)^+}{Z(x)}\right],$$

the average fraction denied service, rather than by (4.3). This is wrong, however, because it does not account for the varying number of customers served. For example, if $C = 100$, then it would count a day in which $Z(x) = 1$ and a day in which $Z(x) = 100$ equally as two days with denied service fractions of zero, when in reality the second day represents 100 times as many customers. The renewal-reward theorem leading to (4.3) provides the correct measure of the long-run fraction of customers who are denied service.

Table 4.2. Binomial and normal approximation overbooking probabilities with $C = 150$, $q = 0.85$.

	Binomial Model			*Normal Approximation*	
x	$\bar{F}_{x-1}(C-1)$	$s_1(x)$	$s_2(x)$	$s_1(x)$	$s_2(x)$
160	0.00021	0.00019	0.00000	0.00097	0.00001
161	0.00053	0.00048	0.00001	0.00185	0.00002
162	0.00122	0.00111	0.00001	0.00340	0.00003
163	0.00262	0.00239	0.00003	0.00601	0.00006
164	0.00523	0.00480	0.00006	0.01022	0.00011
165	0.00979	0.00904	0.00012	0.01676	0.00020
166	0.01726	0.01603	0.00022	0.02652	0.00033
167	0.02883	0.02691	0.00039	0.04053	0.00053
168	0.04577	0.04294	0.00066	0.05989	0.00084
169	0.06935	0.06539	0.00107	0.08566	0.00127
170	0.10062	0.09533	0.00166	0.11873	0.00188
171	0.14025	0.13351	0.00247	0.15966	0.00270
172	0.18838	0.18015	0.00355	0.20855	0.00377
173	0.24449	0.23484	0.00494	0.26496	0.00514
174	0.30743	0.29654	0.00668	0.32785	0.00685
175	0.37549	0.36365	0.00879	0.39565	0.00891
176	0.44654	0.43411	0.01127	0.46635	0.01134
177	0.51828	0.50565	0.01414	0.53773	0.01415
178	0.58842	0.57600	0.01736	0.60753	0.01732
179	0.65493	0.64309	0.02093	0.67366	0.02084
180	0.71616	0.70520	0.02479	0.73442	0.02467
181	0.77096	0.76110	0.02891	0.78856	0.02877
182	0.81870	0.81006	0.03325	0.83539	0.03310
183	0.85919	0.85182	0.03776	0.87472	0.03761
184	0.89270	0.88657	0.04241	0.90681	0.04227
185	0.91975	0.91477	0.04715	0.93225	0.04703

The average service level is, however, easy to calculate for a given distribution of demand. To illustrate, consider the Type 2 service level. Let the random variable D denote the demand (unrestricted by capacity). Then by the renewal-reward theorem, the average Type 2 service for an overbooking level x, denoted $\bar{s}_2(x)$, is given by

$$\begin{aligned}\bar{s}_2(x) &= \frac{E[\min\{D,x\}s_2(\min\{D,x\})]}{E[\min\{D,x\}]} \\ &= \frac{\sum_{y=0}^{x-1} yP(D=y)s_2(y) + xP(D \geq x)s_2(x)}{\sum_{y=0}^{x-1} yP(D=y) + xP(D \geq x)}\end{aligned}$$

One then searches for the largest value of x that provides an average service level $\bar{s}_2(x)$ that is within a given limit.

The problem with using average service levels is that customers who make reservations only on congested days will experience service levels closer to $s_2(x)$ than to $\bar{s}_2(x)$. This is a form of the *inspection paradox* of probability theory, in which customers who book only on busy days experience worse-than-average service. Thus, a service standard based on $s_2(x)$ is often justified since it guarantees that all customers, regardless of their patterns of usage, will experience at least the target service standard.

4.2.1.2 Overbooking Based on Economic Criteria

An alternative to setting overbooking limits based on service levels is to use an economic criterion. This approach requires an estimate of the revenue loss from not accepting additional reservations and an estimate of the cost of denied service. We first develop the details of the economic model and then discuss some of the issues involved in estimating the revenue loss and cost inputs.

Model and Basic Results Suppose z customers show up on the day of service (the show demand), and let $c(z)$ denote the denied-service cost. We shall assume $c(z)$ is an increasing convex function of z. For example, a common assumption in practice is that each denied-service costs the firm a constant marginal amount h, in which case

$$c(z) = h(z - C)^+. \tag{4.4}$$

An arguably more realistic assumption is to assume strictly increasing marginal costs, reflecting the need to offer higher levels of compensation (or incur higher goodwill costs) as each additional customer is denied service.

Let p denote the marginal revenue generated by accepting an additional reservation. One could also allow this marginal revenue to vary, but it is a common simplification in practice to consider it fixed. (We discuss this issue further below.) Then the total expected profit from having y reservations on hand is given by

$$V(y) = py - E[c(Z(y))], \tag{4.5}$$

where, as before, the random variable $Z(y)$ denotes the number of customers who show up on the day of service out of y reservations. One can show for the binomial model that if $c(\cdot)$ is convex, then $V(\cdot)$ is concave, [5] in which case, since $V(y)$ is concave, it follows that it is optimal to accept the y^{th} reservation as long as the marginal profit $\Delta V(y) = V(y) - V(y-1)$

[5] This follows from stochastic convexity arguments; see Appendix B for a discussion.

is positive and to continue accepting reservations until this marginal profit turns negative. Thus, the optimal booking limit x^* is the largest value of x satisfying

$$\Delta V(x) = E[c(Z(x))] - E[c(Z(x-1))] \leq p.$$

For the binomial model with constant marginal costs and parameter q, this condition reduces to

$$hqP(Z(x-1) \geq C) \leq p. \tag{4.6}$$

This expression can be argued intuitively by noting that when we accept the x^{th} reservation, we incur a marginal denied-boarding penalty of h if and only if (1) the current reservations on hand consume all the capacity $(Z(x-1) \geq C)$, and (2) the x^{th} customer shows up. The left-hand side of (4.6) is simply the marginal penalty multiplied by the probability of this event or equivalently the expected marginal cost. Then x^* is the largest value of x for which the expected marginal cost is less than the marginal revenue.

We can express (4.6) as

$$\bar{F}_{x-1}(C-1) \leq \frac{p}{qh}. \tag{4.7}$$

Note that this is equivalent to setting a fixed Type 1 service level for a capacity of $C-1$. For large capacities C, $\bar{F}_x(C-1) \approx \bar{F}_x(C)$, so using economic criteria with constant marginal costs corresponds approximately to specifying a particular Type 1 service level. This fact provides one justification for using Type 1 service levels.

To illustrate (4.7), consider the following example:

Example 4.2 Suppose $C = 150$, $q = 0.85$, the overbooking cost is $h = \$500$, and the marginal revenue is $p = \$100$. Then $p/qh = 0.235$. From Table 4.2 and (4.7), we see that the optimal overbooking limit is then $x^* = 172$ since this is the largest value of x for which $\bar{F}_{x-1}(C-1) \leq 0.235$. The overbooking pad is then 172 – 150 = 22.

Cost and Revenue Parameters While overbooking based on economic criteria is conceptually appealing, it requires good estimates of the marginal revenues and costs. The marginal revenue is usually the easier of the two to determine. If there is only one class, the marginal revenue is simply the common price but with multiple classes determining the marginal revenue is more complex. A heuristic approach is to use the weighted-average revenue. However, as shown in Chapter 2, the marginal revenue produced by an additional unit of capacity is not, in general, equal to the weighted-average revenue. Moreover, the marginal

revenue is typically decreasing in the available capacity, so the linear marginal revenue assumption is violated. Both these factors complicate the estimation of marginal revenue in practice.

Estimating the denied-boarding cost involves other complications. Some elements of this cost are clear: in particular, any refund of the purchase price or monetary compensation or both is easy to quantify in most cases. But if auctions are used to determine compensation, then this compensation must be estimated. Vouchers for free service in the future require a more careful accounting of the actual cost of providing the service, as this is often less than the face value of the voucher.

Most difficult of all to quantify is the goodwill loss of upsetting a customer. In principle, this can be taken to be equal to the discounted potential revenue stream of future purchases from the customer (the so-called lifetime value of the customer). This is rather difficult quantify, but it is usually worth an attempt to make this calculation to at least get the correct order of magnitude of goodwill losses.

One useful idea to get around these estimation problems is to compute imputed costs based on subjective service-level criteria rather than specifying a denied-service cost a priori. To obtain an imputed cost h from a given overbooking limit x^* set according to Type 2 service levels, one simply rearranges (4.7) to obtain

$$h = \frac{p}{q\bar{F}_{x^*-1}(C-1)}.$$

The following example illustrates the use of this formula:

Example 4.3 We saw above that if the service standard is 0.1% $(s_2(y) \leq 0.001)$, we should accept at most $x^* = 168$ reservations. Since $\bar{F}_{x^*-1}(C-1) = 0.02883$, if the marginal revenue is $p = \$100$, this implies an imputed cost of denied service of

$$h = \frac{100}{0.85 \times 0.02883} = \$4,080.$$

The figure looks rather large relative to the \$100 revenue, so one might question if a 0.1% Type 2 service level is economically justified.

Often, these imputed cost numbers provide useful feedback, since they translate service levels, which are tangible and somewhat easier to specify, into economic penalties, which most people find harder to quantify. The economic costs, in turn, serve as a useful "sanity check" on the reasonableness of a given service level by giving the magnitude of the implied costs.

4.2.2 Static-Model Approximations

While the binomial model is quite simple, it is often desirable to have simpler, closed-form expressions for the overbooking limits. We next look briefly at such approximations.

4.2.2.1 Deterministic Approximation

The *deterministic approximation* simply sets the overbooking limit so that the average show demand is exactly equal to the capacity; namely,

$$x^* = \frac{C}{q}.$$

As simplistic as this approximation is, we have seen several RM implementations that use it. The approximation is not completely unjustified, however, as illustrated by the following example:

Example 4.4 Consider our continuing example where $C = 150$ and $q = 0.85$. The deterministic approximation yields an overbooking limit of $x^* = C/q = 176.5$. From Table 4.2, one can see that both service measures $s_1(y)$ and $s_2(y)$ begin to change rapidly in the range of $y = 170$ to $y = 180$, which is approximately centered around the deterministic level 176. So lacking detailed service standards or cost information, a value around the deterministic level is not an unreasonable heuristic to use.

4.2.2.2 Normal Approximation

In practical implementations, it is common to use a continuous approximation to the binomial model to simplify computations. One popular choice is the *normal approximation,* in which the distribution $F_x(\cdot)$ is replaced by the normal distribution with mean μ_x and variance σ_x^2 chosen to match the binomial, viz.,

$$\begin{aligned} \mu_x &= xq \\ \sigma_x^2 &= xq(1-q). \end{aligned}$$

The Type 1 service level is then approximated by

$$s_1(x) \approx 1 - \Phi(z_x),$$

where

$$z_x = \frac{C - \mu_x}{\sigma_x}$$

and $\Phi(\cdot)$ is the c.d.f. of the standard normal distribution.

The Type 2 service level is then approximated by[6]

$$s_2(x) \approx \frac{\sigma_x}{\mu_x}\left[\phi(z_x) - z_x(1 - \Phi(z_x))\right]. \tag{4.8}$$

Table 4.2 shows the estimates produced by the normal approximation for our example with $C = 150$ and $q = 0.85$. As can be seen, they are reasonably close to the values of the binomial model.

The economic overbooking limit (4.7) for the constant marginal-cost function (4.4) is approximated by choosing x^* to satisfy

$$\Phi_{x^*}(C) = 1 - \frac{p}{qh}. \tag{4.9}$$

4.2.2.3 Gram-Charlier Series Approximation

The *Gram-Charlier series* improves on the normal approximation of the binomial distribution by allowing for skewness of the distribution. The standardized density function for this distribution is

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} \left(1 + \frac{1}{6}\beta(z^3 - 3z)\right),$$

where

$$\beta^2 = \frac{E^2[(Z - E[Z])^3]}{E^3[(Z - E[Z])^2]}$$

is the squared coefficient of skewness. If $\beta = 0$, this reduces to the standard normal distribution. For the binomial model, the coefficient of skewness is given by

$$\beta = \frac{1 - q}{\sqrt{q(1 - q)x}}.$$

Letting $z_x = \frac{C - \mu_x}{\sigma_x}$ denote the standardized booking level as before, the fraction of overbooked passengers is approximated by

$$s_2(x) \approx \frac{\sigma_x}{\mu_x}\left[1 + \frac{1}{6}z_x\beta\phi(z_x) - z_x(1 - \Phi(z_x))\right], \tag{4.10}$$

where, as above, $\phi(z)$ and $\Phi(z)$ are the standard normal density and distribution, respectively.

---

[6]This follows from the fact that if Z *is* a normal random variable with mean μ and variance σ^2, then

$$E[(Z - C)^+] = \sigma\left(\phi(z) - z(1 - \Phi(z))\right),$$

where $z = (C - \mu)/\sigma$.

Table 4.3 shows some numerical comparisons of the normal and Gram-Charlier approximations of the binomial model. In general, the normal approximation tends to overestimate the fraction of denied boardings. The Gram-Charlier approximation also overestimates, but less so.

Table 4.3. Comparison of normal and Gram-Charlier (G-C) approximations.[a]

	Booking Limit	*Cancellation*	*Type-2 Service Level*		
Capacity	x	q	*Exact*	*Normal*	*G-C Series*
36	37	0.05	0.0047	0.0064	0.0051
		0.10	0.0006	0.0016	0.0010
	41	0.20	0.0030	0.0044	0.0034
		0.25	0.0006	0.0010	0.0007
	44	0.25	0.0058	0.0073	0.0052
		0.30	0.0013	0.0018	0.0014
		0.35	0.0002	0.0004	0.0003
12	13	0.25	0.0024	0.0059	0.0037
		0.30	0.0011	0.0026	0.0018
	14	0.35	0.0025	0.0046	0.0034
		0.40	0.0011	0.0018	0.0016
	15	0.40	0.0036	0.0049	0.0044

[a] *Note:* Data as reported by Taylor [504].

4.2.3 Customer Class Mix

One important practical issue that arises in overbooking is that different classes may have quite different cancellation rates. For example, in the airline case, full-coach customers often have no cancellation penalty, while discount-class customers typically incur a significant fee for cancelling a reservation. As a result, the two classes exhibit very different rates of cancellation. Thus, the cancellation rate observed in a collection of reservations may be highly dependent on the mix of classes.

Exact methods to handle this class-mix problem involve keeping track of the inventory of each class as a separate state variable and then making overbooking decisions based on this complete vector of state variables. Such an approach is described in detail for a multiclass model in Section 4.5 below.

The difficulty with such exact approaches, however, is that they result in significantly more complicated overbooking models and methodology. As a result, most often in practice, one of several heuristic approaches is used to account for customer class mix.

The most common practice is to use a cancellation probability that is empirically estimated for each resource separately. In this way, one

can capture at least the historical mix of customer segments booked on each resource. Another approach is to estimate the cancellation probability for each class and then use estimates of the class mix on each resource to construct a weighted average cancellation probability for each resource. Compared with straight estimation of the resource-level cancellation rate, this method has the advantage of reducing the variance in the estimates and allows for a more rapid adjustment of cancellation rates as the class mix changes over time.

4.2.4 Group Cancellations

The presence of groups also has an important effect on cancellation models in practice. If a group decides to cancel, then all reservations are cancelled simultaneously. The resulting positive correlation in cancellations increases the variance of the show demand. When dealing with large numbers of reservations, it is often possible to ignore the effect of groups, but with small numbers of reservations, group effects can result in significant deviations from the binomial model.

To gain some sense of the presence of groups in RM bookings, Table 4.4 provides an empirical distribution of group sizes over approximately half a million airline reservations. About half of the reservations are individual reservations, while the other half are from groups of two or more.

Table 4.4. Empirical distribution of group sizes.[a]

Number in Party	*Count of Passengers*	*Percentage*	*Cumulative Percentage*
1	198,056	45.1	45.1
2	114,418	26.0	71.1
3	28,641	6.5	77.6
4	25,688	5.8	83.5
5	17,930	4.1	87.6
6	7,134	1.6	89.2
7	2,135	0.5	89.7
8	2,896	0.7	90.3
9	1,125	0.3	90.6
10	4,960	1.1	91.7
>10	36,375	8.3	100.0
Total	439,358	100.0	

[a] *Note:* Data reported by Rothstein and Stone [446].

One simple technique used in practice to adjust for group size is to simply inflate the variance of the show demand by a factor that accounts

for group size. For example, if we are using the normal approximation to the binomial model as described in Section 4.2.2.2, then the estimate of the mean show demand, μ_x, is unchanged but the variance estimate, σ_x^2, is modified as follows:

$$\begin{aligned} \mu_x &= xq \\ \sigma_x^2 &= kxq(1-q), \end{aligned}$$

where k is a factor to account for group cancellations—for example, the average group size.[7]

A more refined technique for adjusting for groups is based on moment-generating functions. Recall that the moment-generating function of a random variable Z is $\psi_Z(t) = E[e^{tZ}]$. We can then find the moments of Z using the fact that

$$E[Z^n] = \frac{d^n}{dt}\psi_Z(t)\bigg|_{t=0}.$$

Let x denote the overbooking limit and x_k denote the number of groups of size k. We will assume that

$$x_k = \frac{\alpha_k x}{k},$$

where α_k is the historical fraction of reservations that are from groups of size k. As an approximation, we allow x_k to be nonintegral. Let q_k denote the probability that a group of size k survives to the time of service (called the *utilization ratio* in the airline industry). Then the moment-generating function for $Z(x)$, the number of survivals from x total reservations, is

$$\psi_{Z(x)}(t) = \prod_{k=1}^{n} (1 - q_k + q_k e^t)^{x_k},$$

from which one can find the first three central moments of show demand,

$$\begin{aligned} E[Z(x)] &= \sum_{k=1}^{n} x_k k q_k \\ E[(Z(x) - E[Z(x)])^2] &= \sum_{k=1}^{n} x_k k^2 q_k (1 - q_k) \\ E[(Z(x) - E[Z(x)])^3] &= \sum_{k=1}^{n} x_k k^3 q_k (1 - q_k)(1 - 2q_k). \end{aligned}$$

[7] Setting k equal to the average group size is obtained by assuming that all reservations are in groups of exactly size k, in which case with x reservations on hand, there are x/k groups of size k, so the variance of the show demand is $k^2(x/k)q(1-q) = kxq(1-q)$.

These three moments can then be used in the Gram-Charlier series approximation (4.10); alternatively, the first two moments can be used in the normal approximation (4.8).

4.3 Dynamic Overbooking Models

The static models do not explicitly account for the dynamics of arrivals, cancellations, and decision making over time. Here we look at models of overbooking that account for such intertemporal effects. We first look at an exact *dynamic overbooking model* and then discuss heuristic approaches.

4.3.1 Exact Approaches

The model presented here is a simplification of one due to Chatwin [109]. The state variables are time, $t = 1, \ldots, T$, and the number of reservations on hand, y. Let the value function be denoted $V_t(y)$. The terminal costs are

$$V_{T+1}(y) = \begin{cases} 0 & y \leq C \\ -c(y - C) & y > C, \end{cases} \tag{4.11}$$

where C is the fixed capacity and $c(\cdot)$ is a convex cost function penalizing denied service. The revenue received from accepting a new reservation in period t is denoted $p(t) \geq 0$. If a reservation is cancelled in period t, the firm pays a refund of $r(t) \geq 0$. (Note that in this model the refund depends only on the time the reservation is cancelled and not on the time-period in which the reservation was made; this (somewhat unrealistic) assumption is necessary to simplify the state space.)

Let $Z_t(y)$ denote the number of surviving reservations at the end of period t, so that given y reservations are on hand at the end of period t, $Z_t(y)$ is the random number surviving to the start of period $t + 1$. We assume that $Z_t(y)$ has a binomial distribution with survival probability q_t. Let D_t denote the random number of new reservation requests in period t (the demand in period t). D_t is assumed independent across time and independent of $Z_t(y)$.

The order of events in a period is as follows: (1) there are y reservations on hand, and D_t new reservation requests arrive; (2) booking decisions are made for the new reservation requests, raising the booking level to x where $y \leq x \leq y + D_t$; then (3) cancellations are observed at the end of the period. The dynamic programming recursion is then

$$\begin{aligned} v_{t+1}(x) &= E[V_{t+1}(Z_t(x)) - (x - Z_t(x))r(t)] \\ V_t(y) &= E[\max_{y \leq x \leq y + D_t} \{v_{t+1}(x) + (x - y)p(t)\}]. \end{aligned}$$

We then have the following result:

PROPOSITION 4.1 *If the denied-service cost* $c(\cdot)$ *is convex, then an overbooking-limit policy is optimal. That is, in each period* t *there exists a critical value* $x^*(t)$ *such that it is optimal to continue accepting new reservations until the total number of reservations on hand reaches* $x^*(t)$.

This result provides some theoretical support for the use of booking limit policies. The following proposition in turn provides sufficient conditions for optimal booking limits to be monotone in time:

PROPOSITION 4.2 *Suppose the denied-service cost* $c(\cdot)$ *is convex and the survival probabilities* q_t, *the revenues* $p(t)$, *and the refunds* $r(t)$ *satisfy*

$$q_t(p(t) - p(t+1)) + (1 - q_t)(p(t) - r(t)) \geq 0.$$

Then the optimal overbooking limit $x^*(t)$ *(or greatest optimal booking limits if more than one optimal limit exists) decline with time. That is,* $x^*(1) \geq x^*(2) \geq \cdots \geq x^*(T)$.

This declining-booking-limit situation corresponds to the overbooking curve shown in Figure 4.2. Note that this condition is satisfied whenever the revenues are decreasing over time t $(p(t) \geq p(t+1))$ and refunds paid in period t do not exceed the price in period t $(p(t) \geq r(t))$.[8]

Another important monotonicity result concerns how overbooking limits are affected by the magnitude of future demand. In particular, let θ be a parameter of the distribution of arrivals so that $D_t = D_t(\theta)$. Then we have the following

PROPOSITION 4.3 *Suppose the denied-service cost* $c(\cdot)$ *is convex and* $D_t(\theta)$ *is stochastically increasing in* θ. *Then the optimal booking limits* $x^*(t)$ *(or greatest optimal booking limits if more than one optimal limit exists) are nonincreasing in* θ.

---

[8]To gain some intuition for this condition, note it can be written as

$$p(t) \geq q_t p(t+1) + (1 - q_t)r(t).$$

Roughly, this can be interpreted as follows. Suppose we are willing to accept a booking in period $t+1$ in state y. Then $p(t+1)$ must exceed the opportunity cost of an additional reservation in state y in period $t+1$. Now consider state y in period t. If we accept an additional booking, we collect revenue $p(t)$. If this reservation cancels at the end of period t, we pay a refund $r(t)$; this occurs with probability $1-q_t$, and the state (after the cancellation) returns to y. If the reservation survives to period $t+1$, it creates an opportunity cost analogous to accepting a request in period $t+1$, which by the argument above is at most $p(t+1)$ (since we accept a request in state y in period $t+1$); this occurs with probability q_t. So if the revenue $p(t)$ exceeds the "average cost" of these outcomes, then we should be willing accept an arrival in period t in state y as well. Hence, the booking limit in period t is at least as large as that in period $t+1$.

This result says that as demand to come increases (stochastically), it is better to be less aggressive in overbooking at any given point in time. The intuition is that if we have more opportunities to book seats in the future, we do not need to take as great an overbooking risk in the current period. The result also highlights the fact that the optimal overbooking limits in general do depend on future demand, which is something that the static overbooking models ignore. In particular, note that the calculation of costs in the static overbooking model effectively assumes there are no opportunities to replace cancelled reservations with new reservations.[9] Since the degenerate case of no future demand is always stochastically smaller than any nontrivial distribution of future demand, Proposition 4.3 implies that static overbooking models will produce overbooking limits that are higher than optimal.

4.3.2 Heuristic Approaches Based on Net Bookings

While dynamic overbooking models provide some nice insights, they are not used very often in practice. This is due partly to their added complexity and partly because to their similarity to the more general combined capacity control and overbooking models that we look at in Section 4.4 below.

In RM practice, the dynamics of cancellations and new reservations and arrivals are more commonly accounted for by using relative changes in bookings on hand (so-called *net bookings*) rather than customer-level cancellation probabilities as a basis for estimating cancellation rates in a static overbooking model. The idea of net bookings can be best illustrated by going back to Figure 4.2, which shows a sample of the level of booking on hand over time. The change in bookings on hand from one time-period to the next depends on both the number of cancellations and the number of new reservations that are accepted. Looking at changes in the bookings on hand gives us a measure of the net bookings. Quite often in practice, net-bookings data is in fact the only data available for use in estimating overbooking parameters.

Since net bookings reflect both cancellations and new reservations, they can be used to provide an alternative estimate of the cancellation

[9]More precisely, the assumption in the static model is that the show demand when the booking limit x is reached is $Z(x)$, a binomial random variable representing the number of surviving reservations out of x total reservations. Hence, show demand consists *only* of those reservations that survive from the current time until the time of service. If new reservations are accepted to replace cancelled reservations, then the show demand will be larger than $Z(x)$, which is what the dynamic model accounts for.

rate, which one can interpret as an approximation to an exact dynamic model. More precisely, one can estimate the survival "rate" q_t as the average ratio of show demand to the number of bookings on hand in time t (or to the number of peak bookings on hand if t is before the peak; see Figure 4.2.). This net-bookings approach to estimating cancellation rates is again quite prevalent in RM practice and seems to lead to better approximations of real world service levels and costs.

4.4 Combined Capacity-Control and Overbooking Models

Thus far, we have analyzed the overbooking problem in isolation without considering the interaction of overbooking decisions with capacity controls. We next look at both exact and approximate methods to model cancellations and no-shows together with the class allocations of quantity-based RM.

Incorporating no-shows or cancellations in either the static or dynamic single-resource model is not too difficult theoretically, provided one makes the following (not entirely satisfying, but analytically useful) set of assumptions:

ASSUMPTION 4.1
(i) The cancellation and no-show probabilities are the same for all customers.
(ii) Cancellations and no-shows are mutually independent across customers.
(iii) Cancellations and no-shows in any period are independent of the time the reservations on hand were accepted.
(iv) The refunds and denied-service costs are the same for all customers.

The assumptions imply that the number of no-shows and the costs incurred are only a function of the total number of reservations on hand. As a result, we need only to retain a single state variable, and the resulting dynamic programs are only slightly more complex than those presented in Chapter 2.

The most restrictive of these assumptions in practice are (i) and (iv): cancellation options and penalties are often linked directly to a class, so cancellation and no-show rates and costs can vary significantly from one class to the next. Ideally, these differences should be accounted for when making allocation decisions. However, this significantly complicates the problem, as we show below. As already mentioned, Assumption 4.1(ii)

is often unrealistic because reservations from people in groups typically cancel at the same time. Assumption 4.1(iii) is less of a problem in practice and has some empirical support [508].

In most implementations, the overbooking problem is separated from the capacity-allocation problem. Often, an approximate static overbooking model can be solved that is able to relax (at least heuristically) some or all parts of Assumption 4.1. However, given Assumption 4.1, the overbooking and capacity-allocation problems can be combined exactly, as we show next.

4.4.1 Exact Methods for No-Shows Under Assumption 4.1

We first consider only no-shows and assume that there are no cancellations prior to the time of service. Let q_0 denote the probability that a customer with a reservation shows up for service ($1 - q_0$ is the no-show probability). Assumption 4.1(i) says this probability is assumed to be the same for all customers, and Assumption 4.1 (iii) that it is independent of when the reservation was made.

Let $Z_i = 1$ if customer i shows up for service and $Z_i = 0$ otherwise. Given there are y reservations on hand just prior to the time of service, the number of customers who show up at time zero (the show demand), denoted $Z(y)$, is then

$$Z(y) = \sum_{i=1}^{y} Z_i,$$

and by Assumption 4.1(iii) $Z_0(y)$ is a binomial(q_0, y) random variable, with

$$P(Z_0(y) = z) = \left(\begin{array}{c} y \\ z \end{array} \right) q_0^z (1 - q_0)^{y-z}, z = 0, 1, \ldots, y.$$

By Assumption 4.1(iv), the total cost of denied service is only a function of the show demand z. Let $c(z)$ denote the overbooking cost given z. We will require that $c(z)$ be increasing and convex with $c(0) = 0$. Convexity in cost is quite natural since the marginal cost of denying service to customers tends to increase with the number denied. For example, we could have a simple linear cost h per denied customer in which case $c(z) = h(z - C)^+$ where, as before, C is the capacity.

Given this no-show model, $V_0(y)$, the expected cost of service given that there are y reservations on hand at the time of service, is given by

$$V_0(y) = E[-c(Z(y)], y \geq 0. \tag{4.12}$$

Stochastic convexity arguments (see Appendix B) show that $V_0(y)$ is concave in y if $c(\cdot)$ is convex. The above expression then replaces the

boundary conditions of the dynamic program for the static and dynamic models.

4.4.1.1 Static Model

Consider the static model of Section 2.2, where the classes are ordered by prices $p_1 > p_2 > \cdots > p_n$, and we assumed classes arrive in the order of lowest to highest revenue. Classes and stages are indexed by j. The state variable is now defined to be the number of reservations on hand y rather than the remaining capacity x as in Section 2.2.

The Bellman equation (2.3) for the static model is then modified to account for no-shows as follows

$$V_j(y) = E\left[\max_{0\le u\le D_j}\{p_j u + V_{j-1}(y+u)\}\right], \tag{4.13}$$

with boundary conditions (4.12) and $V_j(0) = 0$ for all j, where $V_j(y)$ is now interpreted as the expected *net benefit* (expected revenue minus the expected terminal cost) of operating the system from stage j onward given that there are y reservations on hand.[10]

Given the concavity of $V_0(y)$, the same argument as in Proposition 2.1 from Chapter 2 shows that the value function $V_j(y)$ in (4.13) is concave in y for all j and y. Since there is no hard capacity constraint in this case, it is more meaningful to express the optimal policy in terms of booking limits. The optimal nested booking limits are given by

$$b_j^* = \min\{y \ge 0 : p_j < \Delta V_{j-1}(y)\}, \quad j = 1, \ldots, n-1,$$

where $\Delta V_{j-1}(y) \doteq V_{j-1}(y) - V_{j-1}(y+1)$ now has the interpretation as the marginal cost of holding another reservation in stage $j-1$ and is increasing in y. It is then optimal to accept class j if and only if the number of reservations on hand y is strictly less than b_j^*.

4.4.1.2 Dynamic Model

Similarly, the optimality equations (2.17) for the dynamic model of Section 2.5 are modified to account for no-shows as follows:

$$V_t(y) = E\left[\max_{u\in\{0,1\}}\{R(t)u + V_{t+1}(y+u)\}\right], \tag{4.14}$$

[10]Note that in this case $V_j(y)$ is a decreasing function of y, since the more reservations we have on hand now, the fewer the future opportunities to collect revenue or the higher the expected future terminal costs.

where, recall, $R(t)$ is the random revenue in period t, equal to p_j with probability $\lambda_j(t)$. The boundary conditions are

$$V_{T+1}(y) = E[-c(Z(y)], \quad y \geq 0 \tag{4.15}$$

and $V_t(0) = 0$ for all t. It is optimal to accept an arrival of class j if and only if

$$p_j \geq \Delta V_{t+1}(y),$$

where again $\Delta V_{t+1}(y) = V_{t+1}(y) - V_{t+1}(y+1)$ is interpreted as the marginal cost of accepting another reservation.

Note that under this model, one can always justify accepting a sufficiently high revenue p_j, provided the marginal cost $\Delta V_{t+1}(y)$ is finite. This makes perfect economic sense since we should in principle be willing to accept an almost certain denied-service cost if some customer is willing to pay enough to compensate us for this cost. For example, if the overbooking cost is linear of the form $c(z) = h(z - C)^+$, then the marginal cost is never more than h, so any request with revenue greater than h will always be accepted.

This property of not having an explicit limit on the number of reservations (rather, just an economic limit) has been called *infinite overbooking* by some in the airline industry, since it is in sharp contrast to the usual practice of setting a hard overbooking limit. Also, it highlights the potential suboptimality of using fixed overbooking limits.

4.4.2 Class-Dependent No-Show Refunds

If one relaxes one or more parts of Assumption 4.1, then the problem becomes considerably more difficult. The difficulty stems from the fact that if no-show rates or costs depend on customer class or the time of purchase or both, then one must retain a state variable for each class or each time-period or both. The resulting increase in dimensionality of the dynamic program makes it essentially intractable. However, it turns out that class-dependent refunds can be readily incorporated through an appropriate change of variable.

Suppose customers of class j who no-show in period zero are given a refund h_{j0} that is strictly less than the revenue we receive from them, $h_{j0} < p_j$. However, all other assumptions in Assumption 4.1 hold. A naive formulation of this class-dependent refund feature would require keeping track of each class separately so that refunds can be properly awarded at the time of service. However, whether a given customer no-shows is completely independent of all other decisions and events in the system. Thus, one can in fact charge for the expected refund at the time the reservation is accepted rather than at the time of service, with no

resulting difference in total expected revenues and costs. (This is merely a bookkeeping change.)

More precisely, if we accept a reservation from a customer of class j, it will yield a reduced revenue of

$$\hat{p}_j = p_j - (1 - q_0)h_{j0}$$

independent of everything else in the system. Therefore, we simply use $\hat{p}_j$ in place of p_j in either (4.13) or (4.14) to modify the problem formulation. Note, however, that depending on the refund, the ordering of $\hat{p}_j$ may be different from the ordering of p_j. For example, customers in the high revenue class who receive a full refund if they no-show may yield a lower net revenue $\hat{p}_j$ than customers of a lower class who get no refund if they no-show. Since the nested protection levels are now based on the net revenue $\hat{p}_j$ rather than the gross revenue p_j, the optimal policy may reject the high gross-revenue customer in favor of the high net-revenue one.

4.4.3 Exact Methods for Cancellations Under Assumption 4.1

Cancellations complicate the dynamic program a little more than no-shows, but they are still quite manageable under Assumption 4.1. Again, we look at the static and dynamic models in turn.

4.4.3.1 Static Model

Let q_j denote the probability that a reservation in the system at the start of stage j survives to stage $j-1$ (recall that in the single-resource static model of Section 2.2 stages go from N to 0). So $1 - q_j$ is the probability that a reservation cancels in stage j. By Assumption 4.1 (i), (ii), and (iii), these probabilities are the same and independent for all customers as well as the age of their reservations. Let $Z_j(y)$ denote the number of reservations that survive stage j given that there are y reservations on hand in stage j (so $y - Z_j(y)$ are the number of cancellations in stage j).

The Bellman equation (2.3) for the static model is then modified to account for cancellations as follows

$$V_j(y) = E\left[\max_{0 \le u \le D_j} \{p_j u + H_{j-1}(y + u)\}\right], \tag{4.16}$$

with boundary conditions (4.12), where

$$H_{j-1}(y) = E[V_{j-1}(Z_j(y))] = \sum_{z=0}^{y} \binom{y}{z} q_j^z (1 - q_j)^{y-z} V_{j-1}(z)$$

is the expected value function after cancellations in stage j. Again, stochastic convexity arguments show that if $V_{j-1}(z)$ is concave in z, then $H_{j-1}(y)$ is concave in y, and hence a modification of the argument in Proposition 2.1 shows that the value function $V_j(y)$ defined by (4.16) is concave in y.

Nested booking limits are optimal with the optimal booking limits given by

$$b_j^* = \min\{y \geq 0 : p_j < H_{j-1}(y) - H_{j-1}(y+1)\}, \quad j = 1, \ldots, n-1,$$

where we accept class j if and only if the number of reservations on hand y is strictly less than b_j^*.

4.4.3.2 Dynamic Model

Let q_t denote the probability that a reservation in the system at the start of period t survives to period $t+1$, so by Assumption 4.1 (i), (ii), and (iii) the number of surviving reservations $Z_t(y)$ is again binomial. The optimality equations for the dynamic model with cancellations become

$$V_t(y) = E\left[\max_{u \in \{0,1\}} \{R(t)u + H_{t+1}(y+u)\}\right], \tag{4.17}$$

where

$$H_{t+1}(y) = E[V_{t+1}(Z_t(y))] = \sum_{z=0}^{y} \binom{y}{z} q_t^z (1-q_t)^{1-z} V_{t+1}(z)$$

is the expected value function after cancellations in period t. The boundary conditions are given by (4.15).

As a result, it is optimal to accept an arrival of class j if and only if

$$p_j \geq H_{t+1}(y) - H_{t+1}(y+1).$$

4.4.4 Class-Dependent Cancellation Refunds

Again, relaxing the fact that cancellation rates or costs depend on the class or the time of purchase (or both) requires expanding the state space and is not practical if one has more than two classes. However, as with no-shows, a change of accounting can be used to allow for class-dependent refunds. We illustrate the idea for the dynamic model only, but a similar idea applies for the static model.

Suppose a customer of class j who cancels in period t is given a refund h_{jt} which is strictly less than the revenue we receive, $h_{jt} < p_j$. All other assumptions in Assumption 4.1 hold. As in the no-show case, one can charge for the expected refund at the time the reservation is accepted

rather than at the time of service, with no resulting difference in total expected revenues and costs.

This is accomplished as follows. Let $G_t(j)$ denote the expected refund given to a class j reservation from period t through to the time of service. We can solve for $G_t(j)$ recursively using

$$G_t(j) = (1 - q_t)h_{jt} + q_t G_{t-1}(j), \quad t = 1, 2, \ldots, T$$

with boundary condition

$$G_0(j) = (1 - q_0)h_{j0}.$$

We then form the reduced revenue

$$\hat{p}_{jt} = p_j - G_t(j)$$

and simply use $\hat{p}_{jt}$ in place of p_j in (4.17) to modify the problem formulation. Note, as in the case of no-show refunds, that the ordering of $\hat{p}_{jt}$ may be different from the ordering of p_j.

Again, the key practical insight here is that the reduced-revenue $\hat{p}_{jt}$ should be used in evaluating the economic benefit of accepting a class j customer—not the gross-revenue p_j. This is because even if a class gives a higher current revenue, much of that revenue may be forfeited on average, so the net benefit of accepting it can be quite different from the gross revenue.

4.5 Substitutable Capacity

We next look at an overbooking problem with multiple classes and multiple resources (types of capacity). Here, we assume classes correspond to different products a customer can purchase, while resources are physically different, albeit related, types of capacity. The multiple capacity types may be used to satisfy the demand of a given class—or multiple classes may use a single capacity type. A prominent example is overbooking jointly in multiple cabins of an aircraft (coach and business class), where the first-class cabin serves as a substitute capacity if the coach cabin is oversold. Another example is overbooking on back-to-back scheduled flights between a pair of cities, where customers booked on an early flight can be served (perhaps at a cost) on a later flight. Hotels with multiple room types and car rental fleets with multiple car types are further examples. The case of a single resource with multiple classes can be applied to a traditional single-resource problem to control overbooking when cancellation rates differ across classes (for example, to determine separate overbooking limits for each class based on the joint vector of reservations on hand for each class). All these problems share

the feature that capacity of a different resource (such as a later flight, an alternative room type, or a vehicle type) can serve as a substitute in the case of oversales.

In the presence of such substitution effects, the overbooking decisions for the resources are related. For example, we might tolerate a higher level of overbooking in the coach cabin of an aircraft if we know that the number of bookings in the first-class cabin is low, and conversely we would be more conservative about overbooking the coach cabin if the first-class cabin was fully booked. Therefore, the key question in such situations is how to jointly determine optimal overbooking levels.

4.5.1 Model and Formulation

One approach to joint overbooking across resources is to approximate this problem as a two-period optimization problem. In the first period (the *reservation period*), we assume reservations are accepted given only probabilistic knowledge of cancellations. In the second period (the *service period*), cancellations are realized, and surviving customers are assigned to the various resources to maximize the net benefit of assignments (for example, minimize downgrading penalties). This gives us essentially a multiclass version of the traditional static overbooking model.

Let n denoted the number of classes and m denote the number of resources. In the reservation period, assume that for each class j we currently have y_j reservations on hand. (This is the current "state.") The decision variables are the maximum number of reservations we are willing to hold after the reservation period is over, denoted by x_j, $j = 1, \ldots, n$. These decision variables have to satisfy $x_j \geq y_j$ for all $j = 1, \ldots, n$, since the maximum number of reservations after the reservation period must be at least as large as the number at the start of the reservation period. (There are no cancellations *during* the reservation period.)

In the service period, cancellations and no-shows are realized, and all remaining customers are either assigned to one of m resources, indexed by i, or they are denied service. This assignment of customers to resources is modeled as a deterministic network-flow problem. The following notation is used:

h_{ji}—The net benefit of assigning a customer of class j to resource i during service period (objective function coefficients).

C_i— The capacity of resource i.

z_j—The number of customers of class j that show up at the service period (number of survivals).

z_{ji}—The number of customers among the z_j who showed up assigned to resource i during the service period (decision variables).

One can add a virtual resource, type $i = 0$, to account for denied service. This resource has finite but very high capacity, and assigning a customer to it means that the customer is denied service. The assignment variables corresponding to the virtual resource are z_{j0}, and the objective function coefficients h_{j0} take into account the loss-of-goodwill cost incurred by denying service to customers of reservation class j, as well as any other direct compensation costs.

Let $\mathbf{z}$ denote the n-vector of show demand and $\mathbf{C}$ denote the $(m+1)$-vector of resource capacities (including the denied-service, virtual-resource capacity, C_0). The maximum value obtained during the service period is denoted by $V(\mathbf{z}, \mathbf{C})$. The allocation problem can be represented as

$$
\begin{aligned}
(TP) \qquad V(\mathbf{z}, \mathbf{C}) = \quad \max \quad & \sum_{j=1}^{n} \sum_{i=0}^{m} h_{ji} z_{ji} \\
\text{s.t.} \quad & \sum_{i=0}^{m} z_{ji} = z_j \quad j = 1, \ldots, n \qquad (4.18) \\
& \sum_{j=1}^{n} z_{ji} \leq C_i \quad i = 0, 1, \ldots, m \qquad (4.19) \\
& z_{ji} \geq 0; \quad j = 1, \ldots, n; \quad i = 0, 1, \ldots, m.
\end{aligned}
$$

(TP) is a transportation problem in which the supplies are customers requesting service and demands are the available capacities. Let the dual variables associated with constraints (4.18) and (4.19) in *(TP)* be $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ and $\boldsymbol{\lambda} = (\lambda_0, \ldots, \lambda_m)$, respectively.

To formulate the reservation-period problem, let Z_j be the show demand for customers from class j. This show demand is, of course, a function of the number of accepted reservations, so $Z_j = Z_j(x_j)$. We let q_j denote the probability that a class j reservation shows up in the service period. Several models can be used for this show demand, most naturally the binomial model discussed in Section 4.2.1. But it is useful theoretically and computationally to approximate the binomial with a Poisson distribution, in which case, booking limits can be treated as continuous variables.

Let $\mathbf{y} = (y_1, \ldots, y_n)$, $\mathbf{x} = (x_1, \ldots, x_n)$, and $\mathbf{Z}(\mathbf{x}) = (Z_1(x_1), \ldots, Z_n(x_n))$. Let the price and refund (on cancellation) vectors for the classes be denoted by $\mathbf{p}$ and $\mathbf{s}$, where we assume $\mathbf{p} \geq \mathbf{s}$. Finally, let $G(\mathbf{x})$ be the expected value of future revenues and costs (*net revenue*) as a function of the final overbooking level, $\mathbf{x}$.

The reservation period problem is, then,

$$\max_{\mathbf{x} \geq \mathbf{y}} G(\mathbf{x}), \tag{4.20}$$

where

$$G(\mathbf{x}) = \mathbf{p}^\top(\mathbf{x} - \mathbf{y}) - E[\mathbf{s}^\top(\mathbf{x} - \mathbf{Z}(\mathbf{x}))] - E[V(Z(\mathbf{x}), \mathbf{C})] \tag{4.21}$$

and the expectation above is with respect to the random vector of survivals $Z(\mathbf{x})$.

4.5.2 Joint Optimal Overbooking Levels

The following proposition shows how the overbooking levels for the classes are related if show demand is modeled as a Poisson random variable:

PROPOSITION 4.4 *If for each* $j = 1, \ldots, n$, $Z_j(x_j)$ *is a Poisson distributed random variable with mean* $q_j x_j$, *then the function* $G(\mathbf{x})$ *defined by (4.21) is component-wise concave in each* x_j, $j = 1, \ldots, n$, *and submodular in* $\mathbf{x}$. *That is, letting* $\mathbf{e}_j$ *denote the* j^{th} *unit vector, for all* j, *the first differences*

$$G(\mathbf{x} + \mathbf{e}_j) - G(\mathbf{x}),$$

are decreasing in x_j, $j = 1, \ldots, n$.

The component-wise concavity of the expected net revenue function implies that there are critical booking levels for each class j beyond which the expected value does not increase, provided booking levels of other classes are kept constant. The submodularity property implies that the optimal booking limit for class j is nonincreasing in the booking limit for any other class $i \neq j$. These are natural and intuitive properties. They simply reflect the fact that low reservation levels in one class mean that capacity will be less constrained in the service period, and this in turn reduces the potential costs of overbooking in other classes because more (or at least less costly) substitution options will be available.

In Appendix 4.A we give a stochastic gradient method for computing the optimal joint overbooking limits. It solves (4.20) in the case of the Poisson cancellation model using a simulation-based, stochastic gradient algorithm. The following example illustrates how this method compares with the independent binomial model:

Example 4.5 There are four consecutive flights between the same city pair. For simplicity, assume that all four flights serve one class each and each flight has the same capacity of 100. Flights are ordered in time (the earliest flight is flight 1). Overbooking leads to substitution forward in time, so customers denied boarding on

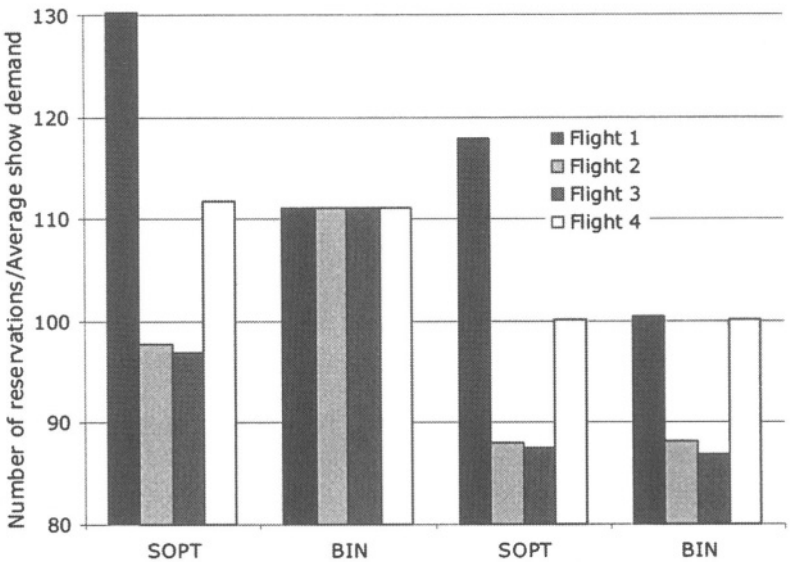

Figure 4.3. **Overbooking limits and show demand for the multiclass and binomial models.**

an oversold flight can take later flights with some loss of goodwill. Denying service completely to a customer results in a higher cost compared with the cost of goodwill due to delays. Delaying a customer by one flight costs \$300; delaying by two flights costs \$400; delaying by three flights is \$500. The cost of denying service to a customer on any flight is \$1,000. The unit revenue for reservations is \$500, which is fully refundable on cancellation. There are 10 reservation periods in the planning horizon, and the survival probabilities are 0.81, 0.82, . . . , 0.90 from the first period to the last. Flights 1 and 4 receive 30% of reservation demand, while flights 2 and 3 receive only 20% each.

Figure 4.3 shows the overbooking limits and final average show demand for the multiclass, stochastic gradient method (SOPT) and binomial model (BIN) for this example. Note that the SOPT procedure is much more aggressive in overbooking flight 1 than flight 4 even though they have the same demand. This is natural, since overselling flight 1 is less costly because passengers can be put onto later flights; oversold passengers on flight 4 must be denied service. Indeed, SOPT in a sense deliberately "plans" oversales on flight 1, since delayed customers on these flights generate more revenue than penalties. This results in nearly 5% of passengers being delayed for one or more flights, while with the BIN procedure only 0.5% of passengers are delayed. Nevertheless, the multiclass model produces a 1.4% increase in revenues (net of penalty costs) over the independent binomial model, as the increased revenues more than compensate for the increase in delay penalties.

While the parameters of this example are not the most realistic, the example illustrates how coordinated overbooking policies for related resources may differ from those computed using independent models.

4.6 Network Overbooking

We next consider how to set overbooking levels on a network. The capacities of network resources are key inputs to capacity-control problems. Using overbooking, these capacities may be inflated—defining virtual capacities for each resource that exceed the physical capacity. This increase in capacity, in turn, affects the accept or reject decisions of the capacity-control method. On the other hand, capacity-control decisions clearly influence the opportunity cost of capacity, which is a key input to economic overbooking models. Hence, the total revenue for a network (net of penalties) is affected both by overbooking and seat inventory-control practices. Despite the strong interdependence of these decisions, the two problems are typically separated in practice.

In this section, we look at one model for coordinating network-capacity controls and overbooking decisions. The method combines the deterministic linear programming model of Section 3.3.1 with a single-period overbooking model, though it can be adapted to other network approximations as well (such as PNLP and RLP).

As in Chapter 3, consider a network with n products and m resources. We divide the time horizon into two periods: a reservation period, and a service period. The reservation period spans $(0, T]$ and is the period the reservations can be made for any of the n products. The reservation period is followed by the service period, during which the customers with reservations show up or become no-shows. During the service period, the firm may deny service to customers who show up in case of insufficient capacity, in which case it pays a penalty.

The demand or reservation requests arrive according to a stochastic process during $(0,T]$. As before, let $\mathbf{p} = (p_1, \ldots, p_n)$ denote the vector of prices and $\mathbf{C} = (C_1, \ldots, C_m)$, the vector of resource capacities. There is a denied-service cost on each resource given by the vector $\mathbf{h} = (h_1, \ldots, h_m)$. The denied-service cost may differ from one resource to another, but it does not vary with time or product type. The matrix $\mathbf{A} = [a_{ij}]$ is the usual network incidence matrix with $a_{ij} = 1$ if resource i is used by product j, and $a_{ij} = 0$, otherwise. Recall $\mathbf{A}^i$ denotes the i^{th} row and $\mathbf{A}_j$ the j^{th} column of matrix $\mathbf{A}$. For simplicity, we ignore refund for cancellations or no-shows, but this can be included easily in this model.

One way to formulate this overbooking problem is as a two-stage, static model that combines the DLP model and the cost-based overbooking models. The same formulation applies to a variety of network bid-price methods, though we focus on the DLP method for simplicity.

The decision variables are $\mathbf{x}$, the vector of overbooking levels (virtual capacities), and $\mathbf{y}$, the vector of primal allocations. (Note here we are changing our running definition of $\mathbf{y}$ to conform with Chapter 3; that is, $\mathbf{y}$ is now a vector of capacity allocations not a vector of reservations on hand.) The formulation is as follows:

$$\begin{aligned} \max_{\mathbf{x},\mathbf{y}} \quad & \mathbf{p}^{\top}\mathbf{y} - E[\mathbf{h}^{\top}(\mathbf{Z}(\mathbf{x}) - \mathbf{C})^{+}] \\ \text{s.t.} \quad & \mathbf{A}\mathbf{y} \leq \mathbf{x} \\ & 0 \leq \mathbf{y} \leq E[\mathbf{D}] \\ & \mathbf{x} \geq \mathbf{C}. \end{aligned} \tag{4.22}$$

The problem parameters are $E[\mathbf{D}]$, the vector of expected demand to come for the n classes. The objective function is the total revenue-to-come, net of denied-service costs.

Note that the show demand for resource i in this formulation is approximated by the random variable $Z_i(x_i)$. The actual show demand, however, will be less, since the show demand for i is $Z_i(x_i)$ *only* if the overbooking limit x_i is reached. (Recall the discussion after Section 4.2.1.1.) Otherwise, the number of reservation on resource i will be less than x_i, and so the show demand will be less than $Z_i(x_i)$. However, this approximation greatly simplifies the model and is a good approximation in the important case where demand is high.

We let $H(\mathbf{x}) = E[\mathbf{h}^{\top}(\mathbf{Z}(\mathbf{x}) - \mathbf{C})^{+}]$ denote the overbooking-cost function and $F(\mathbf{y}) = \mathbf{p}^{\top}\mathbf{y}$ denote the revenue function in (4.22). The overbooking-cost function H is a nondecreasing and convex function of the overbooking limit $\mathbf{x}$ if the random variable associated with the number of survivors for leg i, $Z_i(x_i)$ is assumed to follow the binomial or Poisson model with survival probability q_i. Thus, the objective function of problem (4.22) is jointly concave in $\mathbf{y}$ and $\mathbf{x}$ under these two models of cancellation. One can use a general-purpose nonlinear programming method to solve (4.22), but Appendix 4.B provides an algorithm specialized to this problem's structure. The following numerical example from [292] shows the performance of this method:

Example 4.6 The example here is based on the same network of Williamson [566] as shown in Figure 3.3 of Chapter 3. The itinerary revenue values and base-case mean demand values are show in Table 3.5 of Chapter 3 as well. The cancellation rate is assumed to be 15%, and the denied service penalty is assumed to be \$1,000 on all legs. Different load factors, proportions of local versus through traffic, and arrival order were simulated to create 4 variations of the problem from the base case. A version of the network overbooking model using binomial, rather than Poisson, assumptions of cancellations (denoted BIN) was computed to find joint overbooking levels and corresponding DLP solutions. The resulting overbooking limits and dual prices were then tested by simulation.

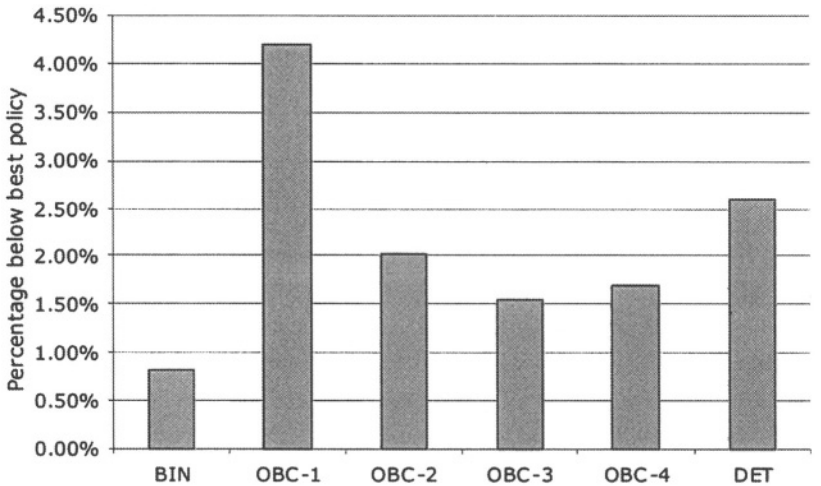

Figure 4.4. Network overbooking example: numerical comparison of policies.

This BIN policy is compared with several versions of ad-hoc overbooking rules. Four of these are cost-based overbooking models, denoted OBC-1 to OBC-4. They differ only in terms of the revenue values used to compute the overbooking limits. Finally, a deterministic overbooking limit (DET) based on the approximation described in Section 4.2.2.1 was also tested. Thus overall five methods are compared with BIN. Once overbooking levels were determined, a DLP model was solved, and the resulting bid prices were used to allocate capacity.

Since no exact methods are known for this problem, the deviation from the best of the six methods was used as the performance metric. That is, the maximum expected revenue (net of penalties) from all the policies is computed and for each individual policy, and the percentage deviation from this maximum is recorded.

Figure 4.4 from [292] shows the average percentage deviation of the six methods. Note that BIN is not always the best method (its average percentage deviation is slightly positive), but it is better than all the other methods.

Similar behavior is observed in other examples in [292], where BIN is not uniformly better than the ad-hoc overbooking mechanisms but is never very far from the best policy and moreover is significantly more robust than any of the ad-hoc methods. These and other tests of the method show the importance of network overbooking; the deviations between the best and worst policy can be quite large—several percentage points of difference in net revenues.

4.7 Notes and Sources

Much of the material in Section 4.1.1 comes from the carefully documented work of Rothstein [447–449] on the development of overbooking

in the airline industry. For papers on overbooking from a policy perspective, see Falkson [180] and Ruppenthal [450]. Overview articles on overbooking include Bodily and Pfeifer [80], Dunleavy [165] and Rothstein [447, 449].

There was also much lively debate surrounding the oversale-auction idea, captured in a series of articles by Simon [472–475]. Vickery [534] proposed using his second-price auction mechanism for the problem as well.

The static overbooking problem of Section 4.2 first appeared in a pair of papers by Beckmann [31, 32]. Other early treatments of the static problem are Taylor [504], Thompson [508], and Rothstein and Stone [446]. See also Bierman and Thomas [66] and Shlifer and Vardi [465]. Martinez and Sanchez [362] test the memoryless property of the binomial model empirically.

The Gram-Charlier approximation in Section 4.2.2.3 is due to Taylor [504] as is the moment-generating-function method presented in Section 4.2.4.

The material in Section 4.3 on dynamic overbooking is from Chatwin's thesis ([107]) and subsequent published articles [108, 109].

The material on combined allocation and overbooking problem of Section 4.4 is from Subramanian et al. [494], who also developed the cost transformation technique of Sections 4.4.2 and 4.4.4.

The multiclass overbooking model with substitution (and associated optimization algorithm) in Section 4.5 are from Karaesmen and van Ryzin [290]. The network overbooking model and algorithm presented in Section 4.6 is from Karaesmen and van Ryzin [291]; see also Karaesmen's thesis [292]. Ladany [320] analyzes a two-class version of this problem for hotels using dynamic programming.

Some papers on practical considerations in hotel overbooking include Lambert et al. [326] and Lefever [337]. The latter discusses handling oversales situations in hotels. For models of hotel overbooking, see Ladany [320, 321] and Liberman and Yechiali [343, 344]. Bitran and Gilbert [71] analyze the problem of sequentially determining when to deny service to arriving customers based on the relative costs of denying service early and late in the evening.

APPENDIX 4.A: Computations for the Substitutable Capacity Model

The optimization problem (4.20) with Poisson cancellations and continuous booking limits can be solved numerically using a simulation-based optimization (stochastic gradient) method. To do so we need an estimator of the gradient of the objective func-

tion $G(\mathbf{x})$. Let the vector $\mathbf{H}^{(k)}$ denote a gradient estimator at the k^{th} iteration of the algorithm. (How this estimator is constructed is discussed below.) The algorithm requires a sequence of step sizes, $\{\gamma_k\}$ satisfying $\sum_{k=1}^{\infty} \gamma_k = +\infty$ and $\sum_{k=1}^{\infty} \gamma_k^2 < +\infty$; for example $\gamma_k = 1/k$. Then, the algorithm proceeds as follows:

STEP 0: Initialize: $k = 1$ and $\mathbf{x}^{(k)} := \mathbf{y}$.

STEP 1: Get the next stochastic gradient:

- Randomly generate a new vector $\mathbf{Z}(\mathbf{x}^{(k)})$.
- Compute the gradient estimate $\mathbf{H}^{(k)}$ (discussed below).

STEP 2: Compute

$$\mathbf{x}^{(k+1)} = \Pi(\mathbf{x}^{(k)} + \gamma_k \mathbf{H}^{(k)}) \tag{4.A.1}$$

where $\Pi(.)$ projects $\mathbf{x}^{(k)} + \gamma_k \mathbf{H}^{(k)}$ onto $\{\mathbf{x} : \mathbf{x} \geq \mathbf{y}\}$

STEP 3: Set $k := k + 1$ and GOTO STEP 1.

An estimator for $\mathbf{H}^{(k)}$ can be constructed using the random function

$$\Delta V_j(\mathbf{Z}(\mathbf{x})) = q_j(V_0(\mathbf{Z}(\mathbf{x}) + \mathbf{e}_j) - V_0(\mathbf{Z}(\mathbf{x}))). \tag{4.A.2}$$

Letting

$$\Delta V(\mathbf{Z}(\mathbf{x})) = (\Delta V_1(\mathbf{Z}(\mathbf{x})), \ldots, \Delta V_n(\mathbf{Z}(\mathbf{x}))),$$

one can show that

$$\nabla_{\mathbf{x}} E[V(\mathbf{Z}(\mathbf{x}))] = E[\Delta V(\mathbf{Z}(\mathbf{x}))],$$

so $\Delta V(\mathbf{Z}(\mathbf{x}))$ is an unbiased estimator of the gradient of $E[V(\mathbf{Z}(\mathbf{x}))]$. The estimator $\Delta V(\mathbf{Z}(\mathbf{x}))$ can be obtained easily by simulating $\mathbf{Z}(\mathbf{x})$ and solving a network-flow problem to obtain $V(\mathbf{Z}(\mathbf{x}))$. Then each estimate $V(\mathbf{Z}(\mathbf{x}) + e_j)$, $j = 1, \ldots, n$ can be determined by perturbing $\mathbf{Z}(\mathrm{x})$ and re-solving the network problem.

Let $\mathbf{z}^{(k)}$ be a realization of show demand when the number of reservations on hand is $\mathbf{x}^{(k)}$ at the k^{th} iteration of the stochastic gradient algorithm. Then the gradient of the objective function at that time is given by the vector $\mathbf{H}^{(k)} = (H_1^{(k)}, \ldots, H_n^{(k)})$, where

$$\begin{aligned} H_j^{(k)} &= p_j - s_j(1 - q_j) + \Delta V_j(\mathbf{Z}(\mathbf{x}^{(k)})) \\ &= p_j - s_j(1 - q_j) + q_j(V(\mathbf{z}^{(k)} + \mathbf{e}_j) - V(\mathbf{z}^{(k)})) \end{aligned} \tag{4.A.3}$$

for $j = 1, \ldots, n$.

APPENDIX 4.B: Alternating-Direction Method for Network Overbooking

To determine the optimal solution $(\mathbf{y}^*, \mathbf{x}^*)$ for the model (4.22), one can use an alternating-direction method for the function. This method efficiently exploits the structure of the problem.

Define the set $\mathbf{\Omega} = \{\mathbf{y} : 0 \leq \mathbf{y} \leq E[\mathbf{D}]\}$ and $\mathbf{\Upsilon} = \{\mathbf{x} : \mathbf{x} \geq \mathbf{C}\}$. The augmented Lagrangian function is

$$L(\mathbf{y}, \mathbf{x}, \pi) = F(\mathbf{y}) - H(\mathbf{x}) - \pi^T(\mathbf{A}\mathbf{y} - \mathbf{x}) - \frac{\gamma}{2}\|\mathbf{A}\mathbf{y} - \mathbf{x}\|^2,$$

where γ is a positive (scalar) parameter. An alternating-direction method can be used to find the maximizers of the augmented Lagrangian. The method proceeds at iteration $k+1$ as follows:

$$\mathbf{y}^{(k+1)} = \arg\max_{\mathbf{y}\in\mathbf{\Omega}}\{F(\mathbf{y}) - (\pi^{(k)})^\top\mathbf{A}\mathbf{y} - \frac{\gamma}{2}\|\mathbf{A}\mathbf{y} - \mathbf{x}^{(k)}\|^2\} \quad (4.B.1)$$

$$\mathbf{x}^{(k+1)} = \arg\max_{\mathbf{x}\in\mathbf{\Upsilon}}\{-H(\mathbf{x}) + (\pi^{(k)})^\top\mathbf{x} - \frac{\gamma}{2}\|\mathbf{A}\mathbf{y}^{(k+1)} - \mathbf{x}\|^2\} \quad (4.B.2)$$

$$\pi^{(k+1)} = \pi^{(k)} + \gamma(\mathbf{A}\mathbf{y}^{(k+1)} - \mathbf{x}^{(k+1)}). \quad (4.B.3)$$

The parameter $\gamma > 0$, initial vectors $\mathbf{x}^{(0)} \geq \mathbf{C}$ and $\pi^{(0)} \geq \mathbf{0}$ are arbitrary. Let $\gamma = 1$. One can show that a sequence $\{\mathbf{y}^{(k)}, \mathbf{x}^{(k)}, \pi^{(k)}\}$ generated by the algorithm (4.B.1), (4.B.2), and (4.B.3) is bounded and every limit point of $\{\mathbf{y}^{(k)}, \mathbf{x}^{(k)}\}$ is an optimal solution to the original problem (4.22). Furthermore, $\{\pi^{(k)}\}$ converges to the optimal dual variable associated with the virtual capacity constraints. A proof of this fact and more details on the method are provided in Bertsekas and Tsitsiklis [56].

To apply this algorithm, we have to solve two different nonlinear programming problems.

Finding $\mathbf{y}^{(k+1)}$, requires solving the following problem:

$$\max F(\mathbf{y}) - (\pi^{(k)})^\top\mathbf{A}\mathbf{y} - \frac{1}{2}\|\mathbf{A}\mathbf{y} - \mathbf{x}^{(k)}\|^2 \quad (4.B.4)$$

for $\mathbf{y} \in \mathbf{\Omega}$. For the DLP model, this is equivalent to the following quadratic program:

$$(QP) \quad \max \quad (\mathbf{p} - \mathbf{A}^\top\pi^{(k)} + \mathbf{A}^\top\mathbf{x}^{(k)})^\top\mathbf{y} - \frac{1}{2}\mathbf{y}^\top\mathbf{A}^\top\mathbf{A}\mathbf{y} \quad (4.B.5)$$

$$\text{s.t.} \quad 0 \leq \mathbf{y} \leq E[\mathbf{D}]. \quad (4.B.6)$$

Problem (QP) can be solved by any standard nonlinear programming method specialized to quadratic programming.

Finding $\mathbf{x}^{(k+1)}$ requires solving

$$(SP) \quad \min_{\mathbf{x}\geq\mathbf{C}} H(\mathbf{x}) - (\pi^{(k)})^\top\mathbf{x} + \frac{1}{2}\|\mathbf{A}\mathbf{y}^{(k+1)} - \mathbf{x}\|^2 \quad (4.B.7)$$

$$= h^\top E[(\mathbf{Z}(\mathbf{x}) - \mathbf{C})^+] - (\pi^{(k)} + \mathbf{A}\mathbf{y}^{(k+1)})^\top\mathbf{x} + \frac{1}{2}\mathbf{x}^\top\mathbf{x} + \frac{1}{2}(\mathbf{y}^{(k+1)})^\top\mathbf{A}^\top\mathbf{A}\mathbf{y}^{(k+1)}.$$

The function in (4.B.7) is separable, convex, and differentiable under the Poisson model of cancellation, and can therefore be solved with a simple line-search method.

We summarize the steps of the algorithm:

STEP 0: Initialize: $\gamma = 1,\ \mathbf{x}^{(0)} = 0,\ \boldsymbol{\pi}^{(0)} = 0,\ k = 1.$

STEP 1: Solve problem *(QP)* and get $\mathbf{y}^{(k)}$.

STEP 2: Solve problem *(SP)* and get $\mathbf{x}^{(k)}$.

STEP 3: Compute $\boldsymbol{\pi}^{(k)}$ using (4.B.3).

STEP 4: Set $k \leftarrow k+1$ and GOTO STEP 1 if $\mathbf{y}^{(k)}, \mathbf{x}^{(k)}, \boldsymbol{\pi}^{(k)}$ do not meet a stopping criterion.

There are several options for the stopping criteria: (1) check that $\mathbf{y}^{(k)}, \mathbf{x}^{(k)}, \boldsymbol{\pi}^{(k)}$ satisfy the KKT conditions, (2) check that $\mathbf{y}^{(k)}, \mathbf{x}^{(k)}, \boldsymbol{\pi}^{(k)}$ are not significantly different from the values of $\mathbf{y}^{(k-1)}, \mathbf{x}^{(k-1)}, \boldsymbol{\pi}^{(k-1)}$, or (3) reach a preset number of iterations; this can be done if one has prior experience with the algorithm and the problems. Karaesmen and van Ryzin [291] show that the algorithm is quite fast and stable on many examples.