Airline Yield Management with Overbooking, Cancellations, and No-Shows

JANAKIRAM SUBRAMANIAN

Integral Development Corporation, 301 University Avenue, Suite 200, Palo Alto, California 94301

SHALER STIDHAM JR.

Department of Operations Research, CB 3180, Smith Building, University of North Carolina, Chapel Hill, North Carolina 27599-3180

CONRAD J. LAUTENBACHER

NationsBank, 100 N. Tryon St., NC1-007-12-3, Charlotte, North Carolina 28255-0001

We formulate and analyze a Markov decision process (dynamic programming) model for airline seat allocation (yield management) on a single-leg flight with multiple fare classes. Unlike previous models, we allow cancellation, no-shows, and overbooking. Additionally, we make no assumptions on the arrival patterns for the various fare classes. Our model is also applicable to other problems of revenue management with perishable commodities, such as arise in the hotel and cruise industries. We show how to solve the problem exactly using dynamic programming. Under realistic conditions, we demonstrate that an optimal booking policy is characterized by state- and time-dependent booking limits for each fare class. Our approach exploits the equivalence to a problem in the optimal control of admission to a queueing system, which has been well studied in the queueing-control literature. Techniques for efficient implementation of the optimal policy and numerical examples are also given. In contrast to previous models, we show that 1) the booking limits need not be monotonic in the time remaining until departure; 2) it may be optimal to accept a lower-fare class and simultaneously reject a higher-fare class because of differing cancellation refunds, so that the optimal booking limits may not always be nested according to fare class; and 3) with the possibility of cancellations, an optimal policy depends on both the total capacity and the capacity remaining. Our numerical examples show that revenue gains of up to 9% are possible with our model, compared with an equivalent model omitting the effects of cancellations and no-shows. We also demonstrate the computational feasibility of our approach using data from a real-life airline application.

In the airline industry, the practice of selling identical seats for different prices to maximize revenues is commonly referred to as yield management (YM) (or seat inventory control). Yield management is an example of a more general practice known as revenue management (RM) or perishable inventory control, in which a commodity or service (such as the use of a hotel room on a particular date) is priced differently depending on various restrictions on booking (e.g., advance purchase requirements) or cancellation (e.g., nonrefundability or partial refundability). The common thread in all these examples is the perishability of the commodity: a seat on a particular flight is worthless after the flight departs, just as an unoccupied hotel room generates no revenue.

The YM problem studied in this paper can be described in its most general form as follows. Consider a single-leg flight on an airplane of known capacity C. Passengers can belong to one of m fare

Transportation Science
Vol. 33, No. 2, May 1999

0041-1655/99/3302-0147 $01.25

classes, class 1 corresponding to the highest fare and class m to the lowest. Booking requests in each fare class arrive according to a time-dependent process. Based on the number of seats already booked, we must decide whether to accept or reject each request. Passengers who have already booked may cancel (with known time- and class-dependent probability) at any time up to the departure of the flight. At the time of cancellation, the passenger is refunded an amount that may be time and class dependent. Passengers can also be no-shows at the time of departure. No-shows are refunded an amount that may be different from the amount refunded for cancellation. Overbooking is allowed, with corresponding penalties determined by an overbooking penalty-cost function.

Although, for concreteness, we use the terminology of airline yield management in this paper, many of our results are also applicable to other problems of RM. For example, in the hotel industry, rooms are sold at different rates, depending on the time until arrival and/or cancellation restrictions. Refunds for cancellation may also depend on the rate.

Our basic approach is to exploit the equivalence of this YM problem to a problem in the optimal control of arrivals to a queueing system, which has been the subject of an extensive literature over the past twenty-five years (see JOHANSEN and STIDHAM (1980), STIDHAM (1984, 1985, 1988), and STIDHAM and WEBER (1993) for surveys). In this way, we are able to solve a substantially more versatile and realistic model than those previously studied in the literature on airline seat allocation.

Among the references most relevant to this paper are (in chronological order) LITTLEWOOD (1972), BELOBABA (1987, 1989), CURRY (1990), WOLLMER (1992), BRUMELLE and MCGILL (1993), ROBINSON (1995), and LEE and HERSH (1993), all of whom consider the single-leg problem. Belobaba (1987) provides a comprehensive overview of the seat inventory control problem and the issues involved. See Belobaba (1989), Lee and Hersh (1993), and LAUTENBACHER and STIDHAM (1996) for surveys and additional references. SMITH, LEIMKUHLER, and DARROW (1992) review yield management practices as they are implemented at American Airlines. WEATHERFORD and BODILY (1992) (see also WEATHERFORD, BODILY, and PFEIFER (1993)) provide a useful taxonomy of perishable asset management (PARM) problems, which include YM and related problems. (In their terminology, our problem is (A1-B1-C1-D1-EI-F3-G3-H3-I1-J1-K1-L5-M2-N3).) WILLIAMSON (1992) and TALLURI and VAN RYZIN (1996) study the multi-leg problem (network RM), which we do not consider in this paper. Other related papers are ALSTRUP et al (1986), CHATWIN (1992), GALLEGO and VAN RYZIN (1994, 1997), and YOUNG and VAN SLYKE (1994).

A common assumption in many of the earlier papers (e.g., Belobaba (1987), Belobaba (1989), Wollmer (1992), and Brumelle and McGill (1993)) has been that all requests for a fare class arrive earlier than the requests for higher fare classes. This assumption is known to be unrealistic. Robinson (1995) relaxes this assumption, but still assumes that different fare classes book during nonoverlapping time intervals. Lee and Hersh (1993) consider the seat inventory control problem without making any assumption on the arrival pattern. They use dynamic programming methodology to solve for the optimal policy. They do not, however, permit cancellations, no-shows, or overbooking.

The rest of this paper is organized as follows. We begin in Section 1 with a simple extension of the model of Lee and Hersh (1993) that incorporates cancellations, no-shows, and overbooking (Model 1). The cancellation and no-show probabilities, although allowed to be time dependent, are assumed to be the same for all classes. There are no refunds for cancellations or no-shows, but the fare paid in each class may be time dependent. (Later, in Section 2, we show how to accommodate refunds by an equivalent-charging scheme borrowed from queueing-control theory, which results in a net fare that is time dependent, even if the gross fare is not.) As in Lee and Hersh (1993), the planning horizon is divided into discrete time periods, in each of which at most one event (reservation request or cancellation) can occur. We formulate the problem as a finite-horizon, discrete-time Markov decision process (MDP) in which the state variable is the total number of seats already booked. With booking requests viewed as customer arrivals and cancellations as service completions, the problem becomes equivalent to a problem in the optimal control of arrivals to a queueing system with an infinite number of servers. Exploiting this equivalence, we show that the YM problem satisfies well-known sufficient conditions for an optimal policy to be monotonic, which, in this context, translates to the existence of time-dependent booking limits for each class. Previously, the optimality of such a policy has been established in the YM literature only for the problem without cancellations or no-shows, and then by complicated arguments apparently dependent on the particular structure of the YM problem.

In Section 2, we extend the basic model to allow class-dependent cancellation and no-show probabilities and refund amounts, with refunds at the time of cancellation or no-show (Model 2). In this case, the

formulation as an MDP requires a multidimensional state variable, because now one must know the number of seats reserved in each fare class to predict future cancellations and no-shows (because the rates at which these occur are now allowed to be class dependent). The curse of dimensionality precludes solving problems of more than modest size, especially considering that an airline is faced with solving such problems for hundreds of legs on a daily basis. In the special case, in which cancellation and no-show probabilities are not class dependent, however, we show how to transform the problem into an equivalent MDP in which the expected lost revenue from cancellation or no-show refunds is subtracted from the fare paid at the time of booking. (This is the equivalent charging scheme referred to above.) This equivalent problem satisfies the conditions of Model 1, for which the total number of seats currently booked is a sufficient state description, thus reducing the problem to a one-dimensional MDP and avoiding the curse of dimensionality. The assumption that cancellation and no-show probabilities are class independent is not realistic, but still represents an improvement on previous models, which ignore the effects of cancellations and no-shows altogether. (In Appendix B, we propose a heuristic approximation for the case of class-dependent cancellation and no-show probabilities which retains the one-dimensional state variable and, thus, greatly enlarges the space of solvable problems.)

Section 3 contains our numerical results. First, we apply Model 1 to several single-leg YM problems with cancellations, no-shows, and overbooking, in which the (time-dependent) cancellation and no-show probabilities are the same for all classes. Our numerical results demonstrate that the optimal booking limits may not be nested by fare class, in contrast to the situation without cancellation or overbooking. More accurately, the order in which the classes are nested may be time dependent, based on the net fare: the gross fare minus the expected cancellation and/or no-show refund. We also compare our results to those for dynamic-programming models that do not allow for the possibility of cancellation and/or no-shows, such as the model of Lee and Hersh (1993). Even if one refines such models by subtracting the expected lost revenue caused by cancellations and no-shows from the fare, these models do not include the cancellation and no-show probabilities in their calculations of future state-transition probabilities. We show, by means of numerical examples, that our model can yield substantial increases in revenues—up to 9% when compared to an equivalent model omitting the effects of cancellations and no-shows. We also demonstrate the computational feasibility of our model, using data from a real-life airline application with six fare classes and a 331-day booking horizon for a plane with 100 seats. Next, we consider Model 2, with class-dependent cancellation and no-show probabilities. We solve a small but realistic model, using the multidimensional MDP formulation discussed in Section 2. Then, we compare this (exact) solution to one-dimensional approximations.

In Appendix A, we show how to apply our results to solve the continuous-time seat-allocation problem, in which requests for seats in each fare class arrive according to (time-dependent) Poisson processes (cf. Lee and Hersh, 1993). We give two (equivalent) approaches, one based on uniformization (LIPPMAN, 1975; SERFOZO, 1979) and the other on a time discretization (Lee and Hersh, 1993).

Appendix B presents our heuristic approximation for the case of class-dependent cancellation and no-show probabilities.

1. BASIC DISCRETE-TIME MODEL (MODEL 1)

IN THIS SECTION, we introduce and analyze our first discrete-time model for airline seat allocation (Model 1), which generalizes the single-seat model of Lee and Hersh (1993). There are m fare classes and N decision periods or stages, numbered in reverse chronological order, $n = N, N - 1, \ldots, 1, 0$, with stage N corresponding to the opening of the flight for reservations and stage 0 corresponding to its departure. At each stage, we assume that one (and only one) of the following events occurs: (1) an arrival of a customer (i.e., a request for a seat) in fare class i, $i = 1, \ldots, m$; (2) a cancellation by a customer currently holding a reservation; or (3) a null event (represented as the arrival of a customer of class 0). (In Appendix A, we show how this assumption is satisfied naturally when a continuous-time problem is approximated either by uniformization or time-discretization.)

In this model, cancellations and no-shows occur at class-independent rates, which allows us to use a one-dimensional state variable. Later (in Section 2) we shall relax this assumption.

Let p_{in} denote the probability of a request for a seat in fare class i in period n. Similarly, let $q_n(x)$ and $p_{0n}(x)$ denote the probability of a cancellation and a null event, respectively, in period n, given that the current number of reserved seats equals x. We assume that $q_n(x)$ is a nondecreasing and concave function of x, for each n. By our assumption that, at most, one request or cancellation can occur at each

stage, we have

$$\sum_{i=1}^{m} p_{in} + q_n(x) + p_{0n}(x) = 1, \qquad (1)$$

for all x and $n \geq 1$.

If the event occurring at stage n is the arrival of a seat request, the system controller (the booking agent) decides whether or not to accept the request, based on the fare class i of the customer and the current state x. If the event is a cancellation or null event, then no decision is made. If the booking agent accepts a request for a seat in fare class i at stage n, the airline earns revenue $r_{in} \geq 0$. (By allowing the revenue to depend on n, we shall be able, in subsequent analysis, to incorporate the effects of cancellation and no-shows by the equivalent charging scheme referred to in the Introduction.)

There are also no-shows. We assume that, at the time of departure, each customer holding a seat reservation is a no-show with probability β. Let $Y(x)$ denote the number of people who show up for the flight, given that the number of reserved seats is x just before departure, so that $x - Y(x)$ is the number of no-shows. Because each customer has a probability $1 - \beta$ of showing up for the flight, it is clear that $Y(x)$ has a binomial-$(x, 1 - \beta)$ distribution. Let C denote the capacity of the airplane. If, at the time of the departure, $Y(x) = y$, then we incur an overbooking penalty $\pi(y)$. We assume that $\pi(y)$ is non-negative, convex, and nondecreasing in $y \geq 0$, with $\pi(y) = 0$ for $y \leq C$.

REMARK 1. Note that we are assuming that requests for seats are independent of the number of seats already booked (a realistic assumption), whereas cancellation and no-show probabilities depend on the total number of booked seats, but not on the number booked in each class. The assumption that $q_n(x)$ is nondecreasing in x expresses the plausible property that the higher the number of seats already booked, the higher the probability of a cancellation. The concavity assumption is needed for technical reasons (in the proof that a booking-limit policy is optimal). It is satisfied in all the applications discussed in this paper. It holds, for example, when $q_n(x) = q_n \cdot x$, as is the case when the discrete-time model is derived from a continuous-time model with a nonhomogeneous Poisson arrival process (see Appendix A).

REMARK 2. The assumption that $\pi(y)$ is convex and nondecreasing is realistic. For example, $\pi(y)$ could be piecewise linear,

$$\pi(y) = \sum_{j=1}^{k} a_j(y - b_j)^+,$$

where $C = b_1 < b_2 < \cdots < b_k$ and $a_j \geq 0$, $j = 1, \ldots, k$. In this case, the airline must pay a_1 to each of the first $b_2 - b_1$ passengers who volunteer to take a later flight, $a_1 + a_2$ to each of the next $b_3 - b_2$, and so forth.

Our objective is to maximize the expected total net benefit[1] of operating the system over the horizon from period N to period 0, starting from state $x = 0$, that is, with no seats booked, at the beginning of period N. The problem can be formulated as a discrete-time MDP with stages $n = N, N - 1, \ldots, 0$, in which the state x is the current number of booked seats. Because our model allows for the possibility of overbooking, cancellations, and no-shows, the state variable need not satisfy the constraint $x \leq C$, as we have noted above. However, because we start with no seats booked at stage N and, at most, one seat request can be accepted at each stage (because, at most, one arrives), it follows that at each stage n, $x \leq N - n$. (To reduce the computational burden in applications, one may wish to introduce an additional constraint, $x \leq C + v$, where v is the overbooking pad: the maximal amount by which the airline is willing to overbook. See Remark 4 below.)

As a function of the state x in period n, let $U_n(x)$ denote the maximal expected net benefit of operating the system over periods n to 0. The optimal value functions, U_n, are determined recursively by

$$U_n(x) = \sum_{i=1}^{m} p_{in}\max\{r_{in} + U_{n-1}(x+1),\ U_{n-1}(x)\} + q_n(x)U_{n-1}(x-1) + p_{0n}(x)U_{n-1}(x),$$

$$0 \leq x \leq N - n, \quad n \geq 1, \qquad (2)$$

$$U_0(x) = \mathrm{E}[-\pi(Y(x))], \quad 0 \leq x \leq N, \qquad (3)$$

where $Y(x) \sim \mathrm{Bin}(x, 1 - \beta)$.

Now we show how to write this optimality equation in an equivalent form that is characteristic of a problem in the optimal control of admission to a queueing system.

Let $p_n := \Sigma_{i=1}^{m} p_{in}$, and let R_n be a random variable with probability mass function $\mathrm{P}\{R_n = r_{in}\} =$

[1] An alternative objective function is the expected total discounted net benefit. The analysis of this problem is substantially the same. See JANAKIRAM, STIDHAM, and SHAYKEVICH (1994).

p_{in}/p_n, $i = 1, \ldots, m$, $n \geq 1$. Let

$$V_n(x, r) := \max\{r + U_{n-1}(x + 1), U_{n-1}(x)\},$$

for all $x = 0, 1, \ldots, N - n$ and real numbers r. (Here, r is a realization of the random variable R_n. In the present problem, the only values of r that have positive probability are of the form $r = r_{in}$ for some $i = 1, \ldots, m$. In more general queueing-control models, the distribution of R_n may be arbitrary, with support $[0, \infty)$.) Then, substituting in Eq. 2 and using Eq. 1, we obtain the equivalent optimality equations

$$U_n(x) = p_n \mathrm{E}[V_n(x, R_n)] + q_n(x)U_{n-1}(x - 1) + (1 - p_n - q_n(x))U_{n-1}(x), \quad 0 \leq x \leq N - n, \quad n \geq 1 \quad (4)$$

$$V_n(x, r) = \max\{r - (U_{n-1}(x) - U_{n-1}(x + 1)), 0\} + U_{n-1}(x), \quad 0 \leq x \leq N - n, \quad n \geq 1 \quad (5)$$

$$U_0(x) = \mathrm{E}[-\pi(Y(x))], \quad 0 \leq x \leq N. \quad (6)$$

In this form, the optimality equations can be seen to be formally equivalent to those for optimal control of admission to a queueing system over a finite horizon as studied, for example, in LIPPMAN and STIDHAM (1977) (see also STIDHAM, 1978; Johansen and Stidham, 1980; HELM and WALDMANN, 1984; Stidham, 1984, 1985, 1988). Lippman and Stidham (1977) study a queue with a Poisson arrival process and a state-dependent exponential service mechanism, which, after uniformization, reduces to a discrete-parameter MDP whose finite-horizon optimality equations (see Eqs. 2 and 3 in Lippman and Stidham, 1977) have exactly the same structure as our optimality Eqs. 4 and 5. (In Appendix A, we shall discuss how to use uniformization to apply our model to a continuous-time seat-allocation problem in which seat requests in each fare class arrive according to a Poisson process, and cancellations are governed by exponential distributions.) The model of Lippman and Stidham (1977) accommodates a holding cost, which is a convex, nondecreasing function of the state. In our model, the holding cost is identically zero.

Lippman and Stidham (1977) use induction on the remaining number of stages, n, to prove that the optimal value function, $U_n(\cdot)$, is concave and nonincreasing, from which it follows that an optimal admission policy is monotonic in the state. (Although their model assumes that rewards, arrival rates, and service rates do not depend on n, the inductive proof is not affected by permitting dependence on n, as we do in our seat-allocation model. See Johansen and Stidham (1980) and Helm and Waldmann (1984), for example, for related models allowing time dependence.) To apply their results to our model, we first need to verify that the function, $U_0(x) = \mathrm{E}[-\pi(Y(x))]$, is concave and nonincreasing in x to start the induction. We use the following result from the theory of stochastic ordering. (See Example 6.A.2 in SHAKED and SHANTHIKUMAR, 1994.)

LEMMA 1. *Let $f(y)$, $y \geq 0$, be a nondecreasing convex function. For each non-negative integer x, let $Y(x)$ be a binomial-(x, γ) random variable $(0 < \gamma < 1)$ and let $h(x) := \mathrm{E}[f(Y(x))]$. Then $h(x)$ is nondecreasing convex in $x \in \{0, 1, \ldots\}$.*

Because $\pi(\cdot)$ is convex and nondecreasing by assumption, it follows from Lemma 1 and Eq. 6 that $U_0(\cdot)$ is concave and nonincreasing. Then, by induction on n (cf. Theorem 1 in Lippman and Stidham, 1977), $U_n(\cdot)$ is concave and nonincreasing, so that $U_n(x) - U_n(x + 1)$ is nondecreasing in $x = 0, 1, \ldots, N - n - 1$. The difference, $U_n(x) - U_n(x + 1)$, is the opportunity cost of accepting a booking at stage $n + 1$, that is, the expected loss in future revenue that would result from accepting the booking. In our model, this opportunity cost plays the same role as the expected marginal seat revenue (EMSR) in the pioneering work of Belobaba (1989) (see also Brumelle and McGill, 1993; Wollmer, 1992). It is also the optimal bid price for our single-leg problem, in the sense of Williamson (1992) (see also Talluri and van Ryzin, 1996). For each stage n and each fare class i, define the booking limit b_{in} as

$$b_{in} := \min\{x: U_{n-1}(x) - U_{n-1}(x + 1) > r_{in}\}. \quad (7)$$

(If $U_{n-1}(x) - U_{n-1}(x + 1) \leq r_{in}$ for all x, set $b_{in} = \infty$.) Because $U_{n-1}(x) - U_{n-1}(x + 1)$ is nondecreasing in x, b_{in} is well defined, and an optimal policy will have the form

accept a fare-class-i request in state x at stage n if and only if $0 \leq x < b_{in}$.

In the notation of Lee and Hersh (1993), $b_{in} = C - \hat{s}_i(n) + 1$, where $\hat{s}_i(n)$ is the critical booking capacity for class i. An optimal policy will therefore accept a fare-class i request at stage n if and only if $s \geq \hat{s}_i(n)$, where $s = C - x$ is the number of seats still available (the booking capacity).

REMARK 3. In contrast to Lee and Hersh (1993), we allow the fares r_{in} to depend on n as well as i, and we do not assume that the ordering of fares by class

is the same for all stages n. So, there may be two classes i and j and stages n and k such that

$$r_{in} < U_{n-1}(x) - U_{n-1}(x+1) \leq r_{jn},$$

whereas

$$r_{ik} \geq U_{k-1}(x) - U_{k-1}(x+1) > r_{jk}.$$

In this case, it is optimal to reject a class i customer and accept a class j customer in state x at stage n, whereas the reverse is true at stage k. It follows that the ordering of the booking limits b_{in} in i may not be the same for all n. In this case, the nesting of the classes may be time dependent. As we shall see in the next section, this phenomenon can occur when the fare in each class is adjusted by subtracting the expected cost of cancellation and no-show refunds.

REMARK 4. As noted above, to reduce the computational burden, it may be advantageous to introduce a maximum overbooking level v (called the overbooking pad), resulting in an additional state constraint, $0 < x < C + v$, at each stage n. The qualitative properties of optimal policies, such as monotonicity of an optimal policy and the existence of booking limits, continue to hold in this setting. To see that this is the case, note that we can incorporate this constraint into the recursions as follows.

At stage $n-1$, given the value functions $U_{n-1}(x)$, for $0 \leq x \leq C + v$, define $U_{n-1}(x) := U_{n-1}(C+v) - M(x - C - v)$, for $x \geq C + v$, where M is a positive number such that $M \geq \max_{i,n}\{r_{in}\}$. Note that this definition ensures that booking requests will always be rejected in state $C + v$ at stage n, while preserving the properties of monotonicity and concavity of $U_{n-1}(\cdot)$ and hence the monotonicity properties of an optimal policy at stage n. It then follows by the inductive argument given previously that the function $U_n(x)$ will also be nonincreasing and concave for $0 \leq x \leq C + v$.

Now, define the opportunity cost functions $u_{n-1}(x) := U_{n-1}(x) - U_{n-1}(x+1)$, $0 \leq x \leq C + v$. Because an optimal policy now rejects all booking requests in state $C + v$, the optimality Eqs. 2 and 3 reduce to

$$U_n(x) = \sum_{i=1}^{m} p_{in}\max\{r_{in} - u_{n-1}(x),\ 0\} + xq_nU_{n-1}(x-1) + (1 - xq_n)U_{n-1}(x), \quad 0 \leq x \leq C + v - 1, \quad n \geq 1, \quad (8)$$

$$U_n(C+v) = (C+v)q_nU_{n-1}(C+v-1) + (1 - (C+v)q_n)U_{n-1}(C+v), \quad n \geq 1, \quad (9)$$

$$U_0(x) = \mathrm{E}[-\pi(Y(x))], \quad 0 \leq x \leq C + v. \quad (10)$$

We shall use the optimality equations in this form to solve for the optimal policy in the numerical examples in Section 3.

2. GENERAL MODEL WITH CLASS-DEPENDENT CANCELLATION AND NO-SHOW RATES (MODEL 2)

IN THIS SECTION, we introduce our most general model (Model 2), which extends Model 1 by allowing class-dependent cancellation and no-show probabilities, as well as refunds at the time of cancellation or no-show.

As with Model 1, there are m fare classes and N stages, numbered in reverse order, $n = N, N-1, \ldots, 1, 0$. Let x_i denote the number of seats currently reserved by class i customers, $i = 1, \ldots, m$. Our state variable is now $\mathbf{x} = (x_1, \ldots, x_m)$, the reservation vector. Let $p_{in}(\mathbf{x})$, $q_{in}(\mathbf{x})$, and $p_{0n}(\mathbf{x})$, respectively, denote the probabilities of a request for a seat in fare class i, a cancellation by a customer of class i, and a null event in period n, given the reservation vector $\mathbf{x}$. Once again, we assume that one and only one of these events occurs during each stage. Note that we are now allowing the arrival probability to depend on the current system state, as well as distinguishing between cancellations in different fare classes and allowing the probability of a cancellation in class i to depend upon the detailed state description, $\mathbf{x} = (x_1, \ldots, x_m)$, rather than just the total number of seats booked, $x = \Sigma_i x_i$. (In applications, as we shall see, the class i cancellation probability will typically depend on $\mathbf{x}$ through x_i, the number of seats currently reserved in class i.) By our assumption that, at most, one request or cancellation can occur at each stage, we have

$$\sum_{i=1}^{m} p_{in}(\mathbf{x}) + \sum_{i=1}^{m} q_{in}(\mathbf{x}) + p_{0n}(\mathbf{x}) = 1,$$

for all $\mathbf{x}$ and $n \geq 1$.

If the event occurring at stage n is the arrival of a seat request, the system controller (the booking agent) decides whether or not to accept the request, based on the fare-class i of the customer and the current state $\mathbf{x}$. If the event is a cancellation or null event, then no decision is made. A customer whose request for a seat in fare class i is accepted at stage n pays a fare $\hat{r}_{in}$. A customer in class i who cancels in decision period n receives a refund c_{in}.

With regard to no-shows, we now assume that, at the time of departure, each customer of class i has a probability β_i of being a no-show, dependent on the class i but independent of everything else. A class i

customer who is a no-show is refunded an amount d_i. Let $Y_i(x_i)$ denote the number of people of class i who show up for the flight, given that the number of reserved class i seats is x_i just before departure, so that $x_i - Y_i(x_i)$ is the number of no-shows in class i. Because each class i customer has a probability $1 - \beta_i$ of showing up for the flight, it is clear that $Y_i(x_i)$ has a binomial-$(x_i, 1 - \beta_i)$ distribution. Let $Y(x) := \Sigma_{i=1}^m Y_i(x_i)$ denote the total number of customers who show up for the flight. If, at the time of departure, $Y(\mathbf{x}) = y$, we incur an overbooking penalty $\pi(y)$. As before, we assume that $\pi(y)$ is non-negative, convex, and nondecreasing in $y \geqslant 0$, with $\pi(y) = 0$ for $y \leqslant C$.

Our objective is to maximize the expected total net benefit of operating the system over the horizon from period N to period 0, starting from state $\mathbf{x} = (0, \ldots, 0)$, that is, with no seats booked, at the beginning of period N. The problem can be formulated as an MDP with stages $n = N, N-1, \ldots, 0$, in which the state, $\mathbf{x} = (x_1, \ldots, x_m)$, is the current reservation vector. Once again, because our model allows for the possibility of overbooking, cancellations, and no-shows, the state variable need not satisfy the constraint $x = \Sigma_i x_i \leqslant C$. However, because we start with no seats booked at stage N and, at most, one seat request can be accepted at each stage (because, at most, one arrives), it follows that $x = \Sigma_i x_i \leqslant N - n$ at stage n. Let $X_n := \{\mathbf{x} = (x_1, \ldots, x_m): x_i \geqslant 0, i = 1, \ldots, m; \Sigma_i x_i \leqslant N - n\}$. Thus, at each stage n, there is an implicit constraint, $\mathbf{x} \in X_n$. (As with Model 1, one can also introduce the constraint, $\Sigma_i x_i \leqslant C + v$, where v is an overbooking pad.)

As a function of the state, $\mathbf{x} \in X_n$, in period n, let $\hat{U}_n(\mathbf{x})$ denote the maximal expected net benefit of operating the system over periods n to 0. The optimal value functions, $\hat{U}_n$, are determined recursively by

$$\hat{U}_n(\mathbf{x}) = \sum_{i=1}^m p_{in}(\mathbf{x}) \max\{\hat{r}_{in} + \hat{U}_{n-1}(\mathbf{x} + \mathbf{e}_i),\ \hat{U}_{n-1}(\mathbf{x})\} + \sum_{i=1}^m q_{in}(\mathbf{x})(-c_{in} + \hat{U}_{n-1}(\mathbf{x} - \mathbf{e}_i)) + p_{0n}(\mathbf{x})\hat{U}_{n-1}(\mathbf{x}), \quad \mathbf{x} \in X_n, \quad n \geqslant 1; \quad (11)$$

$$\hat{U}_0(\mathbf{x}) = \mathrm{E}\left[-\pi(Y(\mathbf{x})) - \sum_{i=1}^m (x_i - Y_i(x_i))d_i \right], \quad \mathbf{x} \in X_0, \quad (12)$$

where $\mathbf{e}_i$ is the ith unit m-vector.

The structure of this MDP reveals that the seat-allocation problem is again essentially equivalent to a problem of admission control to a queueing system, in which the customers are the requests for seats and the cancellation process plays the role of service mechanism. Now, however, because the different classes of customers have different cancellation (service) rates and refunds, it is necessary to keep track of the number of customers in each class, rather than just the total number, as a cancellation from a reservation vector heavily laden with passengers prone to cancel has a higher probability of occurrence than one from a plane filled with those unlikely to do so.

We now show how to transform the optimality Eqs. 11 into an equivalent set of equations in which all expected costs (caused by cancellations and no-shows) are assessed at the instant of admission (booking of a seat) along with the reward (payment of the fare). To do this, we borrow a technique (the equivalent charging scheme) from the queueing-control literature, first proposed in Lippman and Stidham (1977). This transformation will facilitate the reduction of the problem to a one-dimensional MDP in certain cases.

Let $H_n(\mathbf{x})$ denote the total expected loss of revenue over periods n to 0 caused by cancellations and no-shows. (Another interpretation for $H_n(\mathbf{x})$ is that it is the negative of the value function for the policy that rejects all arrivals, starting from state $\mathbf{x}$ at stage n.) Then, the functions H_n are given by the recursive equations

$$H_n(\mathbf{x}) = \sum_{i=0}^m p_{in}(\mathbf{x}) H_{n-1}(\mathbf{x}) + \sum_{i=1}^m q_{in}(\mathbf{x})(c_{in} + H_{n-1}(\mathbf{x} - \mathbf{e}_i)), \quad \mathbf{x} \in X_n, \quad n \geqslant 1 \quad (13)$$

$$H_0(\mathbf{x}) = \mathrm{E}\left[\sum_{i=1}^m (x_i - Y_i(x_i))d_i \right] = \sum_{i=1}^m \beta_i x_i d_i, \quad \mathbf{x} \in X_0. \quad (14)$$

Now, define $U_n(\mathbf{x}) := \hat{U}_n(\mathbf{x}) + H_n(\mathbf{x})$, $n \geqslant 0$. Then, U_n represents the maximal expected controllable net benefit of operating the system over periods n to 0, because $H_n(\mathbf{x})$ is an unavoidable loss of revenue. Using Eqs. 11, 12, 13, and 14, we obtain the follow-

ing recursive optimality equations satisfied by the functions U_n.

$$U_n(\mathbf{x}) = \sum_{i=1}^{m} p_{in}(\mathbf{x})\max\{\hat{r}_{in} - [H_{n-1}(\mathbf{x}+\mathbf{e}_i) - H_{n-1}(\mathbf{x})] + U_{n-1}(\mathbf{x}+\mathbf{e}_i), U_{n-1}(\mathbf{x})\} + \sum_{i=1}^{m} q_{in}(\mathbf{x})U_{n-1}(\mathbf{x}-\mathbf{e}_i) + p_{0n}(\mathbf{x})U_{n-1}(\mathbf{x}),$$

$$\mathbf{x} \in X_n, \quad n \geq 1; \quad (15)$$

$$U_0(\mathbf{x}) = \mathrm{E}[-\pi(Y(\mathbf{x}))], \quad \mathbf{x} \in X_0. \quad (16)$$

These equations are equivalent to the original optimality Eqs. 11 and 12 in the sense that they generate the same optimal booking policy. It follows from Eq. 15 that this policy takes the form

accept a class i request in stage n with reservation vector $\mathbf{x}$

$$\Leftrightarrow \hat{r}_{in} - [H_{n-1}(\mathbf{x}+\mathbf{e}_i) - H_{n-1}(\mathbf{x})] > U_{n-1}(\mathbf{x}) - U_{n-1}(\mathbf{x}+\mathbf{e}_i).$$

That is, in determining whether or not to accept a seat request in fare-class i, we should first calculate the expected net revenue from the booking by subtracting the marginal expected cancellation cost, $H_{n-1}(\mathbf{x}+\mathbf{e}_i) - H_{n-1}(\mathbf{x})$, from the gross fare received, $\hat{r}_{in}$. This quantity should then be compared to the opportunity cost of the booking, $U_{n-1}(\mathbf{x}) - U_{n-1}(\mathbf{x}+\mathbf{e}_i)$, and the request should be accepted if and only if the former exceeds the latter.

In principle, the recursive optimality Eqs. 15 and 16 can be used to calculate the optimal value functions, U_n, and the associated optimal booking policy, for each fare class i, reservation vector $\mathbf{x} \in X_n$, and stage $n = 0, 1, \ldots, N$. (This is just the standard backwards-recursive algorithm of dynamic programming.) However, for our model, in its present, very general form, this algorithm will often not be computationally feasible because of the curse of dimensionality—specifically, the combinatorial explosion in the number of possible states, $\mathbf{x} = (x_1, \ldots, x_m)$, which grows exponentially in m and N. (However, it is feasible to solve small problems exactly: see Example 5 in Section 3.) For this reason, we look for simplifying but realistic assumptions that will reduce the size of the state space.

To this end, we begin by making the following assumption.

ASSUMPTION 1. *$q_{in}(\mathbf{x}) = q_{in}(x_i)$, for all $\mathbf{x} = (x_1, \ldots, x_m)$, where $q_{in}(0) = 0$ and $q_{in}(x_i)$ is nondecreasing in $x_i \geq 1$, $i = 1, \ldots, m$, $n = N, N-1, \ldots, 1$.*

Assumption 1 says that the probability of a class i cancellation in a particular period depends only on the number of class i seats currently booked, and not on the reservations in other classes. This is a reasonable assumption in practice. The following lemma can be proved easily by induction on $n = 1, \ldots, N$.

LEMMA 2. *Under Assumption 1,*

$$H_n(\mathbf{x}) = \sum_{i=1}^{m} H_{in}(x_i), \quad n \geq 0, \quad (17)$$

where the functions H_{in} satisfy the recursive equations ($i = 1, \ldots, m$)

$$H_{in}(x_i) = (1 - q_{in}(x_i))H_{i,n-1}(x_i) + q_{in}(x_i)(c_{in} + H_{i,n-1}(x_i - 1)), \quad x_i \geq 0, \quad n \geq 1 \quad (18)$$

$$H_{i0}(x_i) = \beta_i x_i d_i, \quad x_i \geq 0. \quad (19)$$

Now, let $G_{in}(x_i) := H_{i,n-1}(x_i + 1) - H_{i,n-1}(x_i)$. It follows from Eq. 17 that $G_{in}(x_i) = H_{n-1}(\mathbf{x}+\mathbf{e}_i) - H_{n-1}(\mathbf{x})$, the marginal expected cancellation cost associated with a fare-class i booking in state $\mathbf{x}$ at stage n, which appears in Eq. 15 and was discussed above. Moreover, it follows from Eqs. 18 and 19 that the functions G_{in} satisfy the recursive equations ($i = 1, \ldots, m$)

$$G_{in}(x_i) = (q_{i,n-1}(x_i + 1) - q_{i,n-1}(x_i))c_{i,n-1} + (1 - q_{i,n-1}(x_i + 1))G_{i,n-1}(x_i) + q_{i,n-1}(x_i)G_{i,n-1}(x_i - 1), \quad x_i \geq 0, \quad n \geq 2 \quad (20)$$

$$G_{i1}(x_i) = \beta_i d_i, \quad x_i \geq 0. \quad (21)$$

Further simplifications occur if we strengthen Assumption 1, by requiring that the cancellation rate in each fare class in each period be proportional to the number of seats currently booked in that class:

ASSUMPTION 1′. *$q_{in}(\mathbf{x}) = x_i q_{in}$, for all $\mathbf{x}$, where $q_{in} > 0$, $i = 1, \ldots, m$, $n = N, N-1, \ldots, 1$.*

As we shall see (Appendix A), this assumption is reasonable in applications of our discrete-time model to a system operating in continuous time, in which booking requests arrive according to time-dependent Poisson processes (cf. Lee and Hersh,

1993). In that context, the assumption is equivalent to assuming that each customer cancels independently of all other customers, with a cancellation rate solely dependent on the customer's class.

Using induction on n in Eqs. 20 and 21, we immediately obtain the following lemma.

LEMMA 3. *Under Assumption 1', $G_{in}(x_i) = G_n(i)$, independent of x_i, where the functions $G_n(i)$ satisfy the recursive equations* $(i = 1, \ldots, m)$

$$G_n(i) = q_{i,n-1}c_{i,n-1} + (1 - q_{i,n-1})G_{n-1}(i), \quad n \geqslant 2; \tag{22}$$

$$G_1(i) = \beta_i d_i. \tag{23}$$

This result makes sense intuitively. It states that the marginal expected cancellation cost associated with accepting a request for a seat in fare-class i in stage n is simply the expected cancellation cost attributable to that particular customer over the remaining horizon, which is independent of the number of seats already booked. (It is not difficult to solve Eqs. 22 and 23 for an explicit expression for G_i, but the expression is complicated and we shall not need it. The recursive Eqs. 22 and 23 are more suitable for numerical calculations.)

To recapitulate, we have shown how the calculation of the marginal expected cancellation cost, $H_n(\mathbf{x} + \mathbf{e}_i) - H_n(\mathbf{x})$, simplifies under Assumption 1′, a realistic assumption about cancellation rates. In this case, an optimal seat-allocation policy can be found by choosing, in state $\mathbf{x}$ at stage n, the maximizing action in the recursive optimality equations, which now take the form

$$\begin{aligned} U_n(\mathbf{x}) &= \sum_{i=1}^{m} p_{in}(\mathbf{x})\max\{\hat{r}_{in} - G_n(i) + U_{n-1}(\mathbf{x} + \mathbf{e}_i),\, U_{n-1}(\mathbf{x})\} \\ &\quad + \sum_{i=1}^{m} x_i q_{in} U_{n-1}(\mathbf{x} - \mathbf{e}_i) + p_{0n}(\mathbf{x})U_{n-1}(\mathbf{x}), \\ &\qquad\qquad \mathbf{x} \in X_n, \quad n \geqslant 1 \end{aligned} \tag{24}$$

$$U_0(\mathbf{x}) = \mathrm{E}[-\pi(Y(\mathbf{x}))], \quad \mathbf{x} \in X_0. \tag{25}$$

An optimal policy now takes the form

accept a class i request in stage n with reservation vector $\mathbf{x}$

$$\Leftrightarrow \hat{r}_{in} > G_n(i) + U_{n-1}(\mathbf{x}) - U_{n-1}(\mathbf{x} + \mathbf{e}_i).$$

REMARK 5. Using terminology from queueing control (borrowed in turn from welfare economics), the marginal expected cancellation cost, $G_n(i)$, is an internal effect of the acceptance of the customer (booking request). It is a cost associated directly with the accepted customer, namely, the expected amount that will be refunded to that customer resulting from cancellation or no-show. By contrast, the opportunity cost, $U_{n-1}(\mathbf{x}) - U_{n-1}(\mathbf{x} + \mathbf{e}_i)$, is an external effect, in that it represents the expected loss of revenue from future customers, caused by the admission of the customer in question. So, we see that an optimal policy admits the customer (accepts the booking request) if and only if the gross fare exceeds the sum of the internal and external effects.

We are still left with the necessity of evaluating the optimality Eqs. 24 and 25 for each state $\mathbf{x} = (x_1, \ldots, x_m) \in X_n$, for each stage n. In other words, the curse of dimensionality is still with us. Our next result shows that we can reduce the problem to an equivalent problem with a one-dimensional state variable (the total number of booked seats), provided that arrivals of booking requests are independent of the number of seats already booked, and the individual cancellation rates and no-show probabilities are independent of the fare class.

ASSUMPTION 1″. *$q_{in}(\mathbf{x}) = x_i q_n$, for all $\mathbf{x} \in X_n$, $i = 1, \ldots, m$, where $q_n > 0$, $n = N, N-1, \ldots, 1$.*

ASSUMPTION 2. *$p_{in}(\mathbf{x}) = p_{in}$, for all $\mathbf{x} \in X_n$, $i = 1, \ldots, m$, $n = N, N-1, \ldots, 1$.*

ASSUMPTION 3. *$\beta_i = \beta$, $i = 1, \ldots, m$.*

Let $r_{in} := \hat{r}_{in} - G_n(i)$.

THEOREM 1. *Under Assumptions 1″, 2, and 3, the optimal value functions, $U_n(\mathbf{x})$, depend on $\mathbf{x}$ only through $x = \Sigma_i x_i$, and are determined by the recursive optimality equations*

$$\begin{aligned} U_n(x) &= \sum_{i=1}^{m} p_{in}\max\{r_{in} + U_{n-1}(x+1),\, U_{n-1}(x)\} \\ &\quad + xq_n U_{n-1}(x-1) \\ &\quad + \left(1 - \sum_{i=1}^{m} p_{in} - xq_n\right) U_{n-1}(x), \\ &\qquad\qquad 0 \leqslant x \leqslant N - n, \quad n \geqslant 1 \end{aligned} \tag{26}$$

$$U_0(x) = \mathrm{E}[-\pi(Y(x))], \quad 0 \leqslant x \leqslant N, \tag{27}$$

where $Y(x) \sim \mathrm{Bin}(x, 1 - \beta)$.

Proof. The proof is by induction on n. By assumption, $U_0(\mathbf{x}) = U_0(x) = \mathrm{E}[-\pi(Y(x))]$ and so depends on $\mathbf{x}$ only through $x = \Sigma_i x_i$. Let $n \geqslant 1$ and suppose

$U_{n-1}(\mathbf{x}) = U_{n-1}(x)$ for all $\mathbf{x}$. Then, it follows from Eq. 26 and Assumptions 1, 2, and 3 that

$$U_n(\mathbf{x}) = \sum_{i=1}^{m} p_{in}\max\{r_{in} + U_{n-1}(x+1),\ U_{n-1}(x)\} + \sum_{i=1}^{m} x_i q_n U_{n-1}(x-1) + \left(1 - \sum_{i=1}^{m} p_{in} - xq_n\right) U_{n-1}(x), \quad n \geq 1, \tag{28}$$

where we have used the fact $\Sigma_i\, p_{in} + \Sigma_i\, x_i q_n + p_{0n}(\mathbf{x}) = 1$. It follows from Eq. 28 that $U_n(\mathbf{x}) = U_n(x)$ and Eq. 26 holds. This completes the induction and the proof of the theorem. □

The problem represented by the optimality Eqs. 26 and 27 now has exactly the form of the one-dimensional problem (Model 1) discussed in Section 1, with $r_{in} = \hat{r}_{in} - G_n(i)$, and $q_n(x) = xq_n$ (a concave function of x). So, all the monotonicity results for Model 1 apply, in particular, the optimality of a booking-limit policy.

REMARK 6. Note that we were able to reformulate the problem with the one-dimensional state variable x even though the cancellation and no-show costs are still allowed to be class dependent. This would not have been possible had we continued to charge cancellation and no-show costs in the periods in which they occur, rather than charging the expected cancellation and no-show cost during the period in which the seat is booked.

REMARK 7. In most YM applications, $\hat{r}_{in} = r_i$. That is, the gross fare for class i is independent of n. Assume (without loss of generality) that the classes are ordered so that

$$r_1 \geq r_2 \geq \cdots \geq r_m.$$

Without cancellations, it is well known that the optimal booking limits are nested in the same order as the classes,

$$b_{1n} \geq b_{2n} \geq \cdots \geq b_{mn},$$

and this nesting is the same for all stages n. As we saw in Section 1, however, when the fares r_{in} are time dependent, the ordering of the booking limits may be different for different stages. In particular, this may happen in the present case, in which $r_{in} = r_i - G_n(i)$. Consider two fare classes, i and j, with $r_i > r_j$ (so that $i < j$). If $r_i - G_n(i) < U_n(x) - U_n(x+1) < r_j - G_n(j)$ (which can occur if $G_n(i) - G_n(j) > r_i - r_j$), then it will be optimal to reject fare class i and simultaneously accept fare class j in period n, even though class i has the higher (gross) fare. This could happen if $c_{ik} > c_{jk}$, for all $n > k \geq 1$, (e.g., if fare class i is fully refundable and fare class j is nonrefundable).

REMARK 8. We recognize that it is not realistic to assume that cancellation and no-show rates do not depend on class. We believe, however, that our model is a significant improvement over previous models, in that it explicitly models the effects of cancellations on future seat availability, as well as class-dependent refunds. The class-dependence of refunds permits accurate modeling of the internal effect, $G_n(i)$, of accepting a request (cf. Remark 5 above), whereas the dependence of cancellation and no-show rates on class influences the solution only through the calculation of the external effect, $U_n(x) - U_n(x+1)$. Typically, the internal effect is a first-order effect, whereas the external effect is a second-order effect. Therefore, we have reason to hope that assuming class-independent cancellation and no-show rates (as we do, for example, with our proposed heuristic approximation in Appendix B) will still give us a good approximation to the optimal solution, while retaining the computational advantage of a one-dimensional state variable. Our numerical results (see Example 6 in the next section) tend to support this hope, while emphasizing the need for care in choosing the parameters in the heuristic approximation.

3. NUMERICAL EXAMPLES

IN THIS SECTION, we report the results of numerical solution of several examples. We divide the results into four subsections, each highlighting a different aspect of the work in this paper. Section 3.1 provides counterexamples to commonly held notions in the YM literature, such as the monotonicity of the booking limits and bid prices in the number of booking periods remaining. In Section 3.2, we provide an example of our algorithm performed on actual airline data. We demonstrate that the MDP method can handle real-world-size problems. Section 3.3 includes a model of modest size in which the cancellation rates are class-dependent. This demonstrates the computational feasibility of the multidimensional MDP model of Section 2 for small problems. Finally, in Section 3.4, we demonstrate the power of modeling cancellations through an example that compares various approaches to solving YM problems in the presence of class-dependent cancellation

and no-show rates. We show that significant increases in revenue can result from fully taking into account the effects of cancellations and no-shows.

For all the examples, we work with the special case in which $r_{in} = r_i - G_n(i)$, where r_i is the fare in class i (invariant with n) and $G_n(i)$ is the expected loss of revenue resulting from cancellation or no-show (defined recursively by Eqs. 22 and 23)—in other words, the version that results from applying the equivalent-charging transformation to Model 2 under Assumption 1′. For computational purposes, we introduce an overbooking pad v. When cancellation and no-show rates are independent of the class (Assumption 1″), the result is a version of Model 1 (as we saw in Section 2). The recursive optimality equations in this case take the form

$$U_n(x) = \sum_{i=1}^{m} p_{in}\max\{r_i - G_n(i) - u_{n-1}(x),\ 0\} + xq_nU_{n-1}(x-1) + (1 - xq_n)U_{n-1}(x), \quad 0 \leq x \leq C + v - 1, \quad n \geq 1,$$

$$U_n(C+v) = (C+v)q_nU_{n-1}(C+v-1) + (1-(C+v)q_n)U_{n-1}(C+v), \quad n \geq 1,$$

$$U_0(x) = \mathrm{E}[-\pi(Y(x))], \quad 0 \leq x \leq C + v.$$

(cf. Remark 4). These are the equations that we use for the numerical solution of the one-dimensional examples (Sections 3.1 and 3.2). For the multidimensional examples (Sections 3.3 and 3.4) we use the corresponding special case of Eqs. 24 and 25, with the addition of an overbooking pad v.

3.1 Counterintuitive Results

EXAMPLE 1. Example 1 shows that the functions $u_n(x)$ behave counterintuitively in that they are not always monotone in n. We consider two booking classes with fares $r_1 = 100$ and $r_2 = 10$. The refund amounts are $c_{in} = r_i$, $i = 1, 2$. The maximum overbooking level is $v = 3$. The overbooking cost is given as 55 per seat overbooked. There are no no-shows. The values of the parameters p_{in}, $i = 1, 2$, and q_n are shown in Table I.

Figure 1 plots the bid prices, $u_n(x)$ with respect to n for two different capacities, $C = 5$ and $C = 10$, and various fixed values of x. Note first that the functions $u_n(x)$ are not monotonically increasing in n, in contrast to the examples in Lee and Hersh (1993). Second, if we observe carefully, we note that the capacity remaining, $s = C - x$, is identical ($s = -2$) for the plots corresponding to $x = 7$ and $x = 12$, and that the former plot is always above the latter. The capacity remaining is also identical ($s = 2$) for the plots corresponding to $x = 3$ and $x = 8$, and the former plot is always above the latter. This indicates two things: (1) when we have the possibility of cancellations, the functions u_n and the optimal policy depend on the total capacity, C, and the capacity remaining, $s = C - x$; and (2) for a given s, the u_n functions are monotone (nonincreasing) in C. It is clear from Figure 2 that the booking limits for class 2 are not monotonic in n.

TABLE I
Parameters for Example 1

Parameters	Period n: 10–5	Period n: 4–1
p_{1n}	0.0714	0.0
p_{2n}	0.0	1.0
q_n	0.0714	0.0

EXAMPLE 2. This example is taken from Lee and Hersh (1993). We consider four booking classes with fares $r_1 = 200$, $r_2 = 150$, $r_3 = 120$, and $r_4 = 80$. The capacity of the airplane is $C = 10$. There are no cancellations, no-shows, and no overbooking. The values of the parameters p_{in}, $i = 1, 2$, are shown in Table II. (Because Lee and Hersh do not allow cancellations, $q_n = 0$, for all n.) Figure 3 plots $u_n(x)$ versus n for different x values. In this example, the u_n functions are monotone in n because of the absence of cancellations.

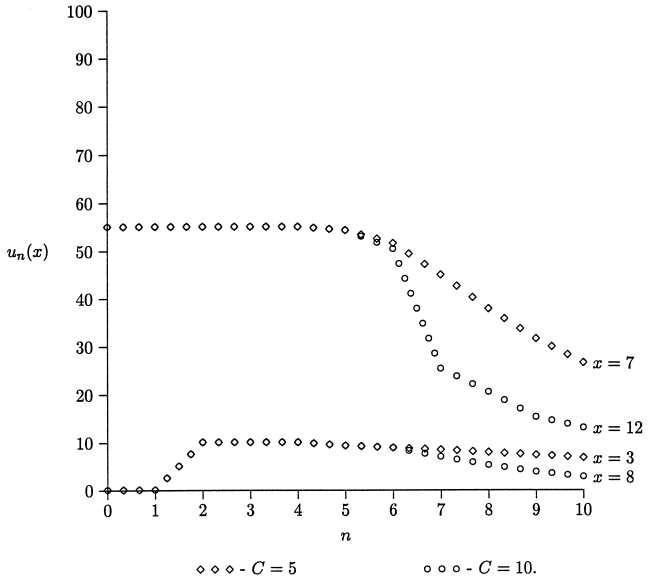

Fig. 1. $u_n(x)$ versus n for different x values for Example 1.

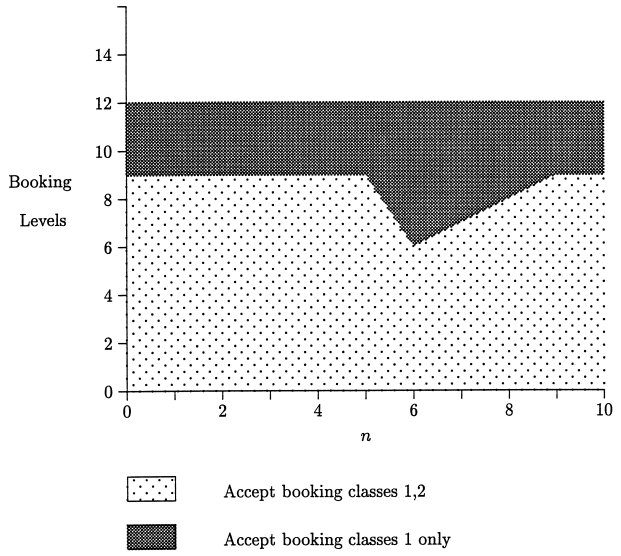

Fig. 2. Partition of the state space by critical values for Example 1.

EXAMPLE 3. We consider four booking classes with fares $r_1 = 200$, $r_2 = 150$, $r_3 = 120$ and $r_4 = 80$. The capacity of the airplane is $C = 20$. The overbooking cost is a piecewise linear and concave function of the overbooking level. The maximum overbooking level is $v = 10$. Cancellations take place at rate q_n. On cancellation, each fare class is refunded a different amount c_{in}, $c_{1n} = 200$, $c_{2n} = 120$, $c_{3n} = 60$, and $c_{4n} = 0$. Class 4 is non-refundable. There are no no-shows. The values of the parameters p_{in}, $i = 1, 2$, and q_n are shown in Table III.

TABLE II

Parameters for Example 2

	Period n				
Parameters	30–26	25–19	18–12	11–5	4–1
p_{1n}	0.08	0.06	0.10	0.14	0.15
p_{2n}	0.08	0.06	0.10	0.14	0.15
p_{3n}	0.14	0.14	0.10	0.16	0.00
p_{4n}	0.14	0.14	0.10	0.16	0.00

TABLE III

Parameters for Example 3

	Period n				
Parameters	30–26	25–19	18–12	11–5	4–1
p_{1n}	0.182	0.086	0.143	0.241	0.25
p_{2n}	0.182	0.086	0.143	0.241	0.25
p_{3n}	0.318	0.2	0.143	0.0	0.0
p_{4n}	0.318	0.2	0.143	0.0	0.0
q_n	0.0	0.014	0.014	0.017	0.0167

Figure 4 shows a plot of $u_n(x)$ versus n for different x values, whereas Figure 5 plots $u_n(x)$ versus x for different values of n. Note that $u_n(x)$ is nondecreasing in x. Figure 6 partitions the state space into the various acceptance regions.

3.2 Real Airline Data

EXAMPLE 4. (This example was previously reported in SUBRAMANIAN, CAMPBELL, and PHILLIPS (1996).) We obtained airline demand and capacity data from a major U.S. airline. To protect the interests of this airline and to guard against release of sensitive data, we have held back (or modified) certain details. The following example uses this data to construct a realistic airline YM problem for a single leg with an airplane of capacity $C = 100$, six fare classes, and a 331-day booking horizon.

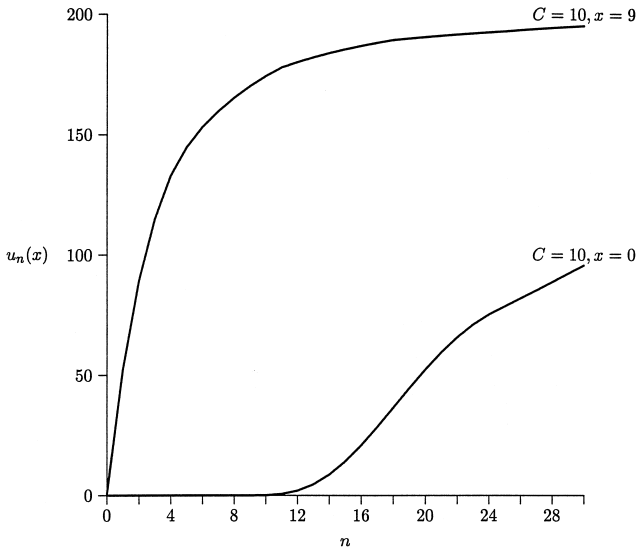

Fig. 3. $u_n(x)$ versus n for different x values for Example 2.

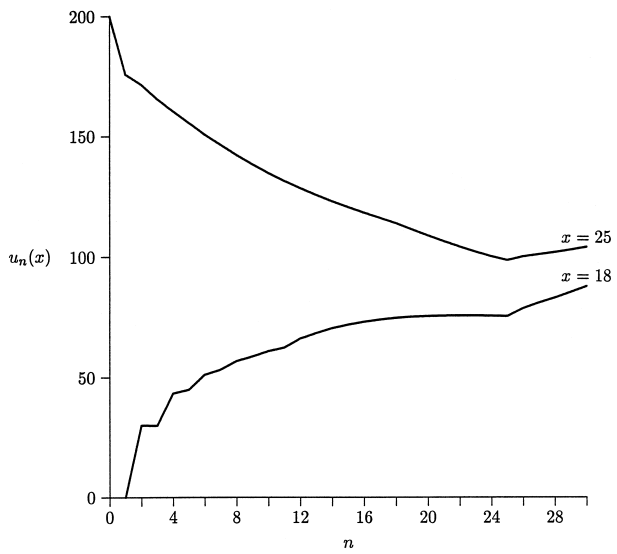

Fig. 4. $u_n(x)$ versus n for different x values for Example 3.

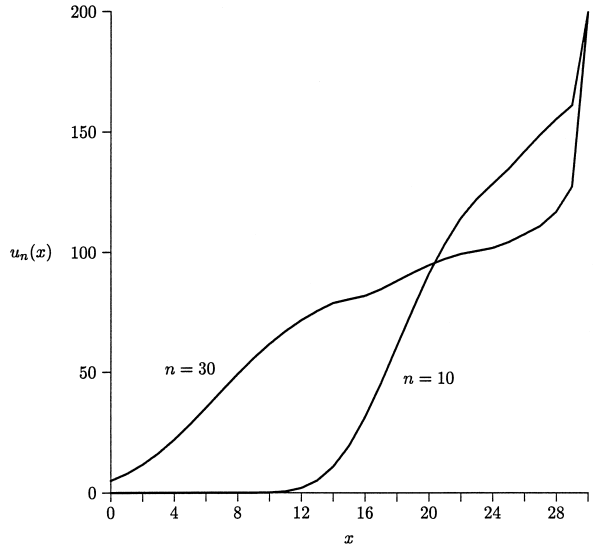

Fig. 5. $u_n(x)$ versus x for different n values for Example 3.

3.2.1 Calculation of Parameters

We received historical airline demand data for all traffic on one flight leg for several departure dates (including both local and multileg traffic). The demand data are given by booking-class, snapshot, point of sale, and passenger type (i.e., individual or group). For a given date, we selected only the local traffic for individual bookings, and restricted point of sale to the United States. We calculated the arrival and cancellation parameters as follows.

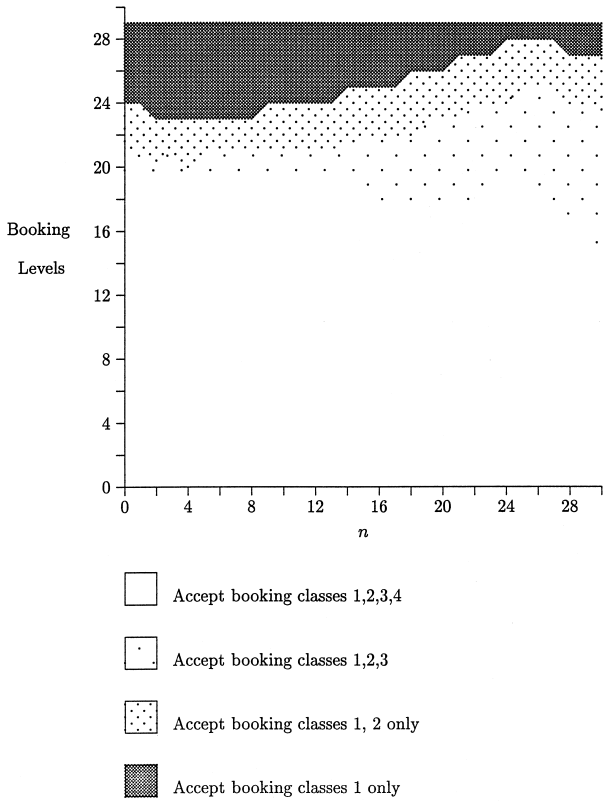

Fig. 6. Partition of the state space by critical values for Example 3.

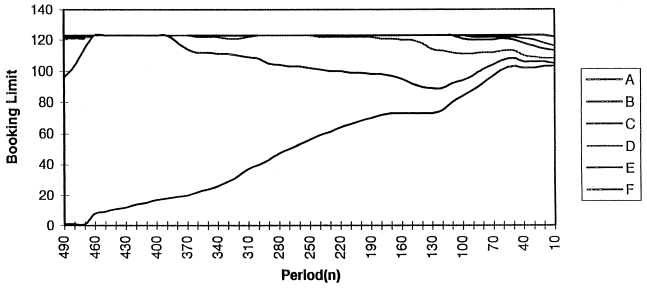

Fig. 7. Booking limits versus time for Example 4: real airline data.

1. Each snapshot was divided into subperiods.
2. The booking probability for fare class i in subperiod n (in snapshot k) was calculated as

$$p_{in} = \frac{\text{number of bkgs in fare class } i \text{ and } k\text{th snapshot}}{N_k}.$$

3. The cancellation rate in subperiod n was calculated as

$$q_n = \frac{\text{total cancellations in } k\text{th snapshot}}{N_k \text{ (load booked at the beginning of } k\text{th snapshot)}}.$$

The number of subperiods N_k was calculated so that

$$\sum_{i=1}^{m} p_{in} + (C + v)q_n < 1.$$

3.2.2 Results

Figure 7 shows the plot of the booking limits versus the time for each of the booking classes. To understand clearly the dynamics of the bid price with respect to time, it will be helpful to refer to Table IV, which shows a mapping of the booking periods to weeks-before-departure for this example. Note that the booking periods grow shorter as departure approaches. This is natural because the event density (expected number of events per unit time) typically increases as takeoff nears. In Fig. 7, the six booking classes are lettered in decreasing order of fare, with class A representing the highest fare class and class F the lowest.

TABLE IV
Mapping of Periods to Weeks before Departure

Snapshot	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
wbd	1	2	3	4	5	6	7	8	9	11	13	15	17	19	21	23	25	33	39	47
period	35	54	94	125	173	243	288	306	339	370	462	465	467	470	489	490	491	494	495	496

wbd, weeks before departure.

3.2.3 Computation-Time Analysis

Solving a single-leg problem with 100 seats and 500 periods on a PC with a 100-MHz Intel 486 processor takes approximately ten seconds. This time increases roughly as the product of the number of seats and number of periods in the model. We are currently investigating performance on a high-end UNIX server and are pursuing other strategies to improve run time. The desired target is to solve a single-leg problem of this magnitude in one second. This would translate to recalculation of 500 legs for 100 departure dates on a 6-processor UNIX server in roughly one hour.

3.3 Class-Dependent Cancellations

EXAMPLE 5. We now solve an example in which the cancellation rates are class dependent. Although the size of the model is modest, this does not entirely diminish the usefulness of the results. In larger problems, it is typically the end effects—when both the time remaining until departure and the remaining capacity are small—that are most crucial to the performance of a booking policy.

We consider three booking classes, with fares $r_1 = 5$, $r_2 = 3$, and $r_3 = 1$. The airplane capacity is 30, with an overbooking pad of 2. The booking horizon is 40 periods. Class 1 is fully refundable, $c_{1n} = d_1 = 5$, whereas class 2 receives only a partial refund, $c_{2n} = d_2 = 1$, and class 3 is nonrefundable, $c_3 = d_3 = 0$. The penalties for overbooking by one and two seats, respectively, are 4 and 10. The values of the parameters p_{in} and q_{in}, $i = 1, 2, 3$ are given in Table V.

In this example, class 1 is intended to be a fully-refundable full-fare ticket, class 2 a mildly discounted ticket with some cancellation restrictions, and class 3 a discount fare with both purchase and cancellation restrictions.

Finding the optimal booking policy for this example took approximately 30 seconds of CPU time on a machine with a 200-MHz Pentium processor running the Linux operating system. The details of the optimal value functions, bid prices, and optimal booking limits are available from the authors upon request. We have omitted them here because of the difficulty in displaying the output from a multidimensional model in a compact yet meaningful way. (In Example 6 below, we will examine the optimal value functions and optimal booking limits for a smaller multidimensional example.) Here we do show, however, that class-dependent cancellation rates can cause the nesting of the net fare to be time dependent (cf. Remarks 3 and 7 above). Table VI shows the net fare of the three fare classes for the final periods in the booking horizon. Note that the nesting order of classes 1 and 2 reverses toward the end of the horizon.

3.4 Power of Cancellations

EXAMPLE 6. The cornerstone of the approach demonstrated in this paper is the inclusion of customer cancellations. To highlight the effects of allowing cancellations, we consider a small example with class-dependent cancellation and no-show rates. Available capacity is $C = 4$ with an overbooking pad of $v = 2$. There are 2 classes, with $r_1 = 3$ and $r_2 = 1$. Class 1 is fully refundable, $c_{1n} = d_1 = 3$, whereas class 2 is nonrefundable, $c_{2n} = d_2 = 0$. Penalties of 2 and 6 correspond to overbooking levels of 1 and 2, respectively. The remaining parameters are summarized in Table VII.

We compare four methods for solving this example.

1. Completely ignore cancellations.

TABLE V
Parameters for Example 5

	Period n				
Parameters	40–21	20–11	10–6	5–1	0
p_{1n}	0.05	0.1	0.2	0.3	0.0
p_{2n}	0.2	0.2	0.2	0.2	0.0
p_{3n}	0.2	0.3	0.0	0.0	0.0
q_{1n}	0.013	0.013	0.015	0.015	0.03
q_{2n}	0.05	0.05	0.0	0.0	0.0
q_{3n}	0.0	0.0	0.0	0.0	0.0

TABLE VI
Net Fare versus n for Example 5

	Period n				
Class	36	37	38	39	40
1	2.967000	2.929000	2.891000	2.853000	2.816000
2	2.878000	2.873000	2.869000	2.865000	2.860000
3	1.0	1.0	1.0	1.0	1.0

TABLE VII

Parameters for Example 6

Parameters	Period n				
	16–13	12–9	8–5	4–1	0
p_{1n}	0.0	0.1	0.3	0.4	0.0
p_{2n}	0.3	0.5	0.0	0.0	0.0
q_{1n}	0.0	0.0	0.05	0.1	0.2
q_{2n}	0.0	0.0	0.0	0.0	0.0

TABLE VIII

Comparison of Class-Independent Cancellation Rates for Example 6

Method	Period n			
	16–9	8–5	4–1	0
3a	0.0	0.025	0.05	0.1
3b	0.0	0.05	0.1	0.2
3c	0.0	0.02	0.04	0.08

2. Adjust fares by expected lost revenue due to cancellations, but ignore effects of cancellations on future seat availability.
3. Approximate the class-dependent cancellation and no-show rates with a single class-independent rate.
4. Solve the problem as stated.

Method 1 is the easiest method and reflects the general approach in the YM literature. The effects of cancellations and overbooking on both the net fare received and on the number of passengers booked are ignored. [Following the previous literature, however, we do add a pad to the physical capacity and then allocate this limit—without cancellations or overbooking—as if it were the true capacity. In effect, this approach decomposes the problem, first setting an overbooking limit (the pad), perhaps based on a simplified cancellation and overbooking model, then applying a seat allocation model that assumes no cancellations or overbooking.] Method 2 is a first step toward incorporating cancellations explicitly. Fares are adjusted by subtracting the (readily calculated) expected lost revenue. The effects of cancellations on the state variable are still omitted from the MDP optimality equations. Method 3 incorporates both the reduction in net fare and the fluctuations of total passengers booked produced by cancellations, but does so by introducing a single cancellation and no-show rate that is assumed to hold across all classes. Note that Methods 1–3 use one-dimensional MDP models (Model 1). Method 4 solves the problem optimally as stated, using the multidimensional MDP model (Model 2).

When using Method 3, for the purposes of comparison, we solved the problem with three different rates, summarized in Table VIII. For Method 3a, we averaged the cancellation and no-show rates of the two classes. In Method 3b, we simply used the rates corresponding to class 1. We chose the class-independent rates in Method 3c in such a way as to approximate as closely as possible the solution obtained in Method 4. In this particular example, that consisted of using 40% of the class-1 cancellation and no-show rates.

We compare the four methods by the following procedure. For Methods 1–3, we solve the simplified, one-dimensional problem optimally. The optimal booking limits from each of these methods are then evaluated in the context of the actual (multidimensional) problem, and the results compared to those of the optimal solution as obtained through Method 4.

The expected revenues obtained by each of the methods are summarized in Table IX, where % Sacrificed is the additional revenue that could be gained by solving the problem optimally (Method 4), expressed as a percentage of the revenue from the given method. Note that Method 2 actually exhibits worse performance than does Method 1, even though it appears to be a more accurate model in that cancellation and no-show refunds are subtracted from the gross fare. Evidently, adjusting the fare in this way without also taking into account the effects of cancellations on future seat availabilities sends the wrong signal to the algorithm for this example. To understand why this might be the case, consider a class $i > 1$ with a large cancellation/no-show refund and a high cancellation rate. After subtracting the expected cancellation and no-show refunds from the fare, the net fare for class i may be so low that it ends up being rejected to save seats for later arriving higher-fare requests. But, because Model 2 does not take into account the fact that a class i customer, if accepted, has a high probability of canceling later (thus making the seat available anyway), it tends to penalize this class more than it should.

This result indicates the importance of accurately modeling both the internal and the external effects of cancellations, as we do in Models 3c and 4, for

TABLE IX

Summary of Methods 1–4

Method	$U_{16}(0, 0)$	% Sacrificed
1	5.86	9.39
2	5.74	11.67
3a	6.22	3.05
3b	5.05	26.93
3c	6.38	0.47
4	6.41	

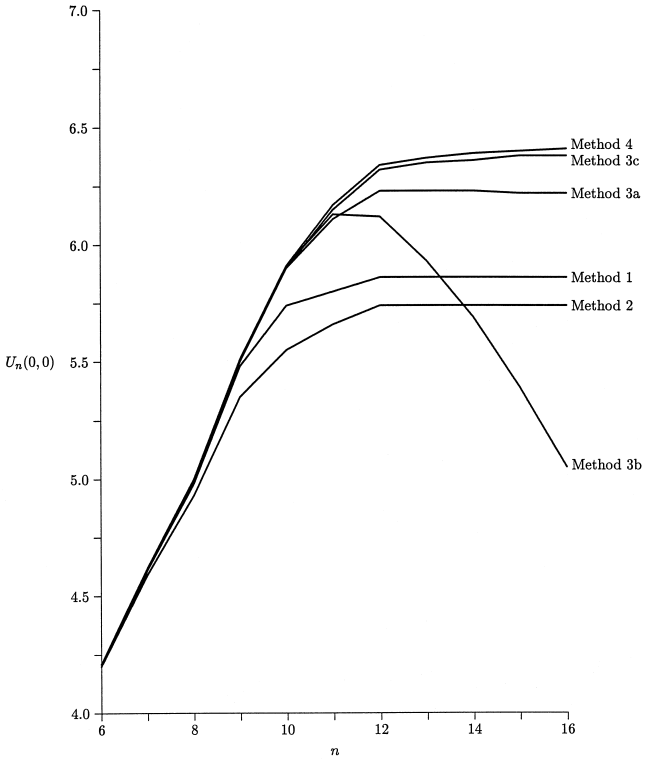

Fig. 8. $U_n(0)$ versus n for the different methods for Example 6.

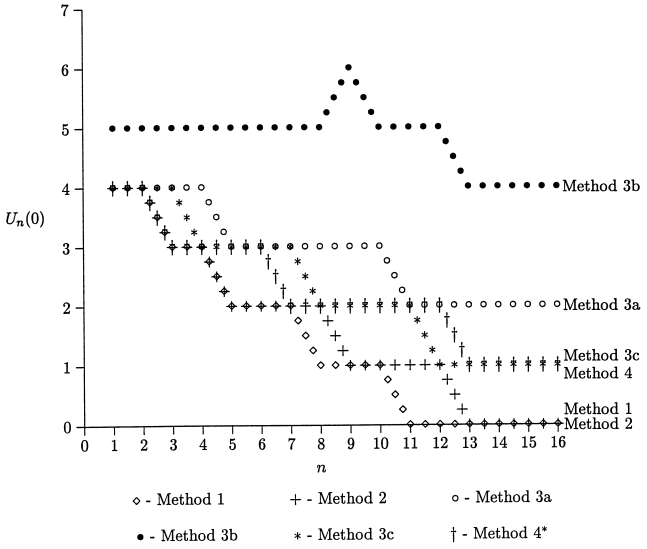

Fig. 9. Class-2 booking limits for the different methods for Example 6.

example. In particular, the YM practitioner should be wary of simplistic approaches, such as simply calculating the net fare by subtracting the expected cancellation and no-show refund from the gross fare and then using one of the standard YM approaches, such as EMSR or Lee and Hersh's model. Such an approach, as we have seen in this example, could actually lead to lower revenue.

Figure 8 compares the values of $U_n(0)$ for each of the four methods. Note how important it is to choose the single-leg cancellation rate properly. (Periods 0–5 are omitted from the graph because all 4 methods produce identical values of $U_n(0, 0)$ during those periods.)

Figure 9 plots the optimal booking limits for class 2 for Methods 1–4. In this figure, we use booking limit to mean the maximum number of reservations to accept in a given period. Thus, a booking limit of 4 implies that a passenger would be accepted provided 3 or fewer passengers had been accepted previously. Because Method 4 uses class-dependent cancellation and no-show rates, the optimal booking limits depend on the number of passengers booked in each class. In Figure 9, we have rounded down the policy for Method 4. For example, if we put down the booking limit as three, that means that we always accept whenever there are two or fewer customers in the system, regardless of how they are distributed. It is possible that we might also accept requests when there are more than two customers, but only if they are distributed in a certain way. Table X displays the optimal booking limits for class 2 customers. Note the slight deviations from Figure 9.

4. CONCLUSIONS AND SUGGESTIONS FOR FUTURE WORK

WE HAVE CONSIDERED a single-leg airline YM problem with multiple fare classes and time-dependent arrival probabilities. Currently booked customers may cancel at any time or become no-shows at the time of departure, with probabilities and refunds that, in general, may be both class and time dependent. Overbooking is permitted.

We have formulated the problem as a discrete-time MDP and used dynamic programming (backward induction) to analyze it. The analysis has two components: (1) characterization of the structure of an optimal policy; and (2) numerical calculation of the parameters of an optimal policy. Our analysis exploits the equivalence of the YM problem to the well-studied problem of optimal control of admission to a queueing system. In its most general form (Model 2), the MDP has a multidimensional state space, because it is necessary to keep track of the number of seats booked in each fare class. When cancellation and no-show probabilities are independent of the fare class, we have used a transformation (the equivalent-charging scheme) to convert the problem to an equivalent problem (Model 1) with a one-dimensional state variable: the total number of seats booked. In this case, we have used well-known results from queueing-con-

TABLE X

Booking Limits for Class 2 from Method 4

	Period n															
(x_1, x_2)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
(0, 0)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
(1, 0)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
(0, 1)	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
(2, 0)	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
(1, 1)	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
(0, 2)	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
(3, 0)	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
(2, 1)	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
(1, 2)	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
(0, 3)	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(4, 0)	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(3, 1)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(2, 2)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(1, 3)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(0, 4)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(5, 0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(4, 1)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(3, 2)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(2, 3)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(1, 4)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(0, 5)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(6, 0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(5, 1)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(4, 2)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(3, 3)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(2, 4)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(1, 5)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(0, 6)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

1, accept; 0, reject.

trol theory to show that an optimal policy is monotonic in the state variable and thus is characterized by booking levels for each fare class. These booking levels are determined by the optimal bid prices, which are easily characterized in terms of the dynamic-programming optimal value functions.

In our numerical examples, we have demonstrated that an optimal booking policy can have counterintuitive properties when cancellations and no-shows are included. For example, booking limits need not be monotonic in the number of booking periods remaining. Also, the opportunity cost of a booking now may depend on both the remaining capacity and the number of seats already booked, rather than just the remaining capacity. Using data from a real airline application, we have shown that our one-dimensional model is computationally feasible for realistic-size problems, in terms of both the availability of data and the running times of the algorithm. We have also demonstrated computational feasibility for the multidimensional model on modest-size problems, which suggests that it might be used to explore the end effects as both the remaining horizon length and the remaining capacity become small. Our experience and that of other researchers suggest that this is where the action is, in the sense that revenues are most sensitive to model accuracy in this region.

Finally, we have considered a small problem with class-dependent cancellation and no-show rates and compared the performance of four methods. In order of increasing accuracy (and complexity), these are: (1) Model 1, ignoring cancellations and no-shows (essentially the model of Lee and Hersh, 1993); (2) Model 1, using equivalent charging—subtracting the expected cancellation and no-show refunds from the gross fare—but ignoring the (probabilistic) effects of cancellations on future seat availabilities; (3) Model 1, incorporating both cancellation and no-show refunds and probabilities, but using approximate probabilities independent of class; (4) Model 2, with both refunds and probabilities dependent on class (as given). For this example, revenue actually decreases when going from Method 1 to Method 2. An increase of nearly 9% can be achieved by incorporating both refunds and cancellation and no-show

probabilities, with careful use of the one-dimensional heuristic approximation (Method 3). The revenue increment in going from Method 3 to Method 4 is less than 1%, which suggests that the full multidimensional model may not always be needed. Additional numerical work is needed to see if these qualitative conclusions are valid over a wider range of parameter values.

In principle, the methodology of this paper can be applied to the multi-leg YM problem. The hurdle is the combinatorial explosion of the state space in the number of legs. We are currently working on solving the two-leg problem exactly and efficiently. For the general multi-leg problem, we hope to develop efficient heuristics to approximate the optimal policy.

APPENDIX A: APPLICATION TO SYSTEM WITH TIME-DEPENDENT POISSON ARRIVAL PROCESS

IN THIS APPENDIX, we show how to apply the MDP model of Section 1 to a continuous-time seat-allocation problem. In the continuous-time problem, the decision horizon is the time interval $[0, T]$. Seat requests for fare class i ($i = 1, \ldots, m$) arrive according to a time-dependent Poisson process with rate $\lambda_i(t)$, $0 \leq t \leq T$. So, the probability that a class i request occurs in the interval $(t, t + dt)$ is $\lambda_i(t)dt + o(dt)$, independent of the previous arrivals and everything else that occurred in $[0, t)$. Cancellations occur in a memoryless, possibly time-dependent, manner. Specifically, each customer in the system at time t has a probability $\mu(t)dt + o(dt)$ of canceling in $(t, t + dt)$, independent of everything that occurred in $[0, t)$. A customer whose request for a seat in fare class i is accepted pays the fare r_i (assumed independent of t, to keep the exposition simple). A customer in class i who cancels at time t receives a refund $c_i(t)$, $0 \leq t \leq T$. Each customer who holds a seat reservation in class i just before departure has a probability β_i of being a no-show, in which case the customer receives a refund d_i, as in the discrete-time model.

To approximate this model by a discrete-time model, we divide the time interval $[0, T]$ into N periods, where N is a sufficiently large integer. There are two ways of doing this, both of which lead to an MDP model of the form presented in Section 2. The first method is an extension to time-dependent processes of the approach introduced into queueing control by Lippman (1975) and commonly called uniformization. This approach results in periods of random length, with an exponential distribution that is independent of the state but may be time dependent. The second approach is simply to divide the interval $[0, T]$ into N deterministic subintervals, possibly of varying lengths, each small enough so that the probability of more than one event (arrival or cancellation) in a subinterval is negligible.

Approximation using Uniformization

In the uniformization approach the continuous-time system is observed at discrete, randomly spaced points in time, each point corresponding to the arrival of a seat request, a cancellation, or a null event. Null events, which do not change the state of the system, are introduced from a Poisson process with a state- and time-dependent rate in such a way that the total rate at which the observation points occur is independent of the state. In addition, this total event rate and the total number of periods N are chosen large enough so that the expected horizon length is close to the actual horizon length T (by the law of large numbers).

To construct the time-dependent uniformization and the equivalent discrete-time MDP, we must specify the probabilities p_{in}, $i = 1, \ldots, m$, and q_n, for each decision period n and state $\mathbf{x}$. Beginning with period N, let $\lambda_{iN} := \lambda_i(0)$ ($i = 1, \ldots, m$), $\lambda_N := \sum_{i=1}^m \lambda_{iN}$, and $\mu_N := \mu(0)$. Choose $\Lambda_N \geq \lambda_N$, and set

$$p_{iN} = \frac{\lambda_{iN}}{\Lambda_N}, \quad i = 1, \ldots, m, \qquad q_N = \frac{\mu_N}{\Lambda_N}.$$

For each period n, $N > n \geq 1$, assume Λ_N, Λ_{N-1}, $\ldots$, Λ_{n+1} have been chosen and let

$$t = \frac{1}{\Lambda_N} + \frac{1}{\Lambda_{N-1}} + \cdots + \frac{1}{\Lambda_{n+1}}.$$

Let $c_{in} := c_i(t)$, $\lambda_{in} := \lambda_i(t)$ ($i = 1, \ldots, m$), $\lambda_n := \sum_{i=1}^m \lambda_{in}$, and $\mu_n := \mu(t)$. Choose $\Lambda_n \geq \lambda_n + (N - n)\mu_n$, and set

$$\gamma_n = \frac{\Lambda_n}{\Lambda_n},$$

$$p_{in} = \frac{\lambda_{in}}{\Lambda_n}, \quad i = 1, \ldots, m,$$

$$q_n = \frac{\mu_n}{\Lambda_n}.$$

Upon substitution of these expressions in Eqs. 4 and 5, the optimality equations for this problem take the form

$$U_n(x) = \frac{1}{\Lambda_n}[\lambda_n E[V_n(x, R_n)] + x\mu_n U_{n-1}(x - 1) + (\Lambda_n - \lambda_n - x\mu_n)U_{n-1}(x)], \quad 0 \leq x \leq N - n, \quad n \geq 1, \quad \text{(A1)}$$

$$V_n(x, r) = \max\{r - (U_{n-1}(x) - U_{n-1}(x+1)), 0\} + U_{n-1}(x),$$
$$0 \leq x \leq N - n, \quad n \geq 1, \quad \text{(A2)}$$

$$U_0(x) = \mathrm{E}[-\pi(Y(x))], \quad 0 \leq x \leq N. \quad \text{(A3)}$$

Comparing Eqs. A1 and A2 to the optimality Eqs. 2 and 3 in Lippman and Stidham (1977), we see once again that our model takes the same form as the finite-horizon arrival-control model in Lippman and Stidham (1977), with the arrival and service rates allowed to depend on the number of periods remaining. Specifically, with n periods remaining and x customers in the system, the time Z_n until the next event occurs is exponentially distributed with mean rate Λ_n. The next event is an arrival (booking request) with probability λ_n/Λ_n, a departure (cancellation) with probability $x\mu_n/\Lambda_n$, and a null event with probability $(\Lambda_n - \lambda_n - x\mu_n)/\Lambda_n$. (The requirement that $\Lambda_n \geq \lambda_n + (N - n)\mu_n$ ensures that $0 \leq (\Lambda_n - \lambda_n - x\mu_n)/\Lambda_n \leq 1$.

For the case of time-homogeneous parameters, LIPPMAN (1976) has observed that the MDP that results from uniformization can be used to solve approximately a continuous-time Markovian decision process with a finite horizon of fixed, deterministic length T. More precisely, by choosing the uniformization parameter Λ and the number of periods, N, so that $N/\Lambda = T$ and then letting $\Lambda \to \infty$, $N \to \infty$, one can show that optimal policies for the discrete-parameter MDP converge to optimal policies for the original continuous-time MDP. The condition that $N/\Lambda = T$ ensures that the discrete-parameter MDP has an expected horizon length T; moreover, as N and Λ approach ∞, the actual horizon length converges to T, by the law of large numbers. In addition, for any t, $0 \leq t \leq T$, an approximately optimal policy to follow at time t may be found by calculating the optimal policy for the discrete-parameter MDP with n periods remaining, where $(N - n)/\Lambda = t$, when both N and Λ are sufficiently large.

Our continuous-time seat-allocation problem differs from that considered by Lippman (1976) in that it has time-dependent parameters in addition to a finite horizon. As a result, the discrete-parameter MDP that we have constructed involves an additional level of approximation. First, we approximate the continuously varying parameters, $\lambda_i(t)$ and $\mu(t)$, by parameters λ_{in} and μ_n, that are constant over random (exponentially distributed) intervals. Then we choose the number of periods, N, and the total-event rate, Λ_n, for each period $n = N, \ldots, 1$ so large that the length of each such random interval approaches zero. We expect that an approximately optimal policy to follow at time t may be found by calculating the optimal policy for the discrete-parameter MDP with n periods remaining, where $\Lambda_N^{-1} + \Lambda_{N-1}^{-1} + \cdots + \Lambda_{N-n+1}^{-1} = t$, when N and $\Lambda_N, \ldots, \Lambda_1$, are sufficiently large.

Although a theoretical proof that this approach provides approximately optimal policies for the continuous-time problem with time-dependent parameters does not seem to be available, we have proposed this approach as an intuitively plausible and computationally feasible heuristic. In any case, in the scenario proposed by Lee and Hersh (1993), in which it is assumed that the time-dependent Poisson arrival processes for each fare class have constant arrival rates over fixed time intervals (called booking periods), the validity of the approximation follows by a straightforward extension of the arguments in Lippman (1976).

Approximation using Decision Periods of Deterministic Length

Here, we divide the horizon into N deterministic intervals, which can, however, be of unequal length. The length of these intervals is chosen to be small enough so that the probability of more than one event (arrival/cancellation) occurring in an interval is small, and the probability of each type of event can be approximated by the mean rate for that event times the length of the interval. Lee and Hersh (1993) follow essentially this approach, although they do not allow cancellations and no-shows.

Suppose the length of the nth period is Δ_n, $n = N, \ldots, 1$, where, as usual, we number the periods in reverse chronological order. Beginning with period N, let $\lambda_{iN} := \lambda_i(0)$ $(i = 1, \ldots, m)$, $\lambda_n := \Sigma_{i=1}^m \lambda_{iN}$, $\mu_N := \mu(0)$, and set

$$p_{iN} = \lambda_{iN}\Delta_N, \quad i = 1, \ldots, m, \quad q_N = \mu_N\Delta_N.$$

For each period n, $N > n \geq 1$, let $t = \Delta_N + \Delta_{N-1} + \cdots + \Delta_{n+1}$. Let $\lambda_{in} := \lambda_i(t)$ $(i = 1, \ldots, m)$, $\lambda_n := \Sigma_{i=1}^m \lambda_{in}$, $\mu_n := \mu(t)$, and set

$$p_{in} = \lambda_{in}\Delta_n, \quad i = 1, \ldots, m, \quad q_n = \mu_n\Delta_n.$$

For sufficiently small Δ_n, it follows from the properties of the exponential distribution that these expressions for the discount factor γ_n and the probabilities, p_{in} and q_n, are accurate within an error that is $o(\Delta_n)$. Upon substitution of these expressions in Eqs. 4 and 5, the optimality equations for this problem take the form

$$U_n(x) = \lambda_n\Delta_n E[V_n(x, R_n)] + x\mu_n\Delta_n U_{n-1}(x-1) + (1 - (\lambda_n + x\mu_n)\Delta_n)U_{n-1}(x),$$
$$0 \leq x \leq N - n, \quad n \geq 1, \quad \text{(A4)}$$

$$V_n(x, r) = \max\{r - (U_{n-1}(x) - U_{n-1}(x+1)), 0\} + U_{n-1}(x), \quad 0 \leq x \leq N-n, \quad n \geq 1, \quad \text{(A5)}$$

$$U_0(x) = \mathrm{E}[-\pi(Y(x))], \quad 0 \leq x \leq N. \quad \text{(A6)}$$

Note that these equations are in the same form as the optimality Eqs. A1, A2, and A3 for the uniformization approximation, as can be seen by replacing Δ_n by $(\Lambda_n)^{-1}$ in Eq. A4.

APPENDIX B: HEURISTIC MODEL FOR CLASS-DEPENDENT CANCELLATIONS AND NO-SHOWS

RECALL THAT WE WERE able to transform our MDP model to an equivalent model with a one-dimensional state variable, provided that Assumption 1″ holds. For convenience, we reproduce Assumption 1″ below.

ASSUMPTION 1″. $q_{in}(\mathbf{x}) = x_i q_n$, *for all* $\mathbf{x}$, $i = 1, \ldots, m$, *where* $q_n > 0$, $n = N, N-1, \ldots, 1$.

This assumption may not be realistic in applications. As an extreme example, in the case of nonrefundable fares, the cancellation or no-show probability is close to zero, whereas cancellation and no-show rates tend to be positive and are often non-negligible among passengers who pay full fare. Assumption 1′ (reproduced below) is more realistic, inasmuch as it allows cancellation rates to be class dependent.

ASSUMPTION 1′. $q_{in}(\mathbf{x}) = x_i q_{in}$, *for all* $\mathbf{x}$, *where* $q_{in} > 0$, $i = 1, \ldots, m$, $n = N, N-1, \ldots, 1$.

Assumption 1′ may be more realistic, but then, we are faced with the problem of keeping track of an enlarged state $\mathbf{x}$. To get around this problem, we propose a heuristic approximation scheme, in which we estimate the number of seats, x_i, booked in each fare class i, based on the observed total number of seats x. From these estimates, we can compute $q_n(x)$, the estimated total cancellation rate, given that x seats are currently booked in period n. We suggest below two intuitive ways to estimate $q_n(x)$, given x and q_{in}, $i = 1, \ldots, m$.

METHOD 1. Define

$$q_n(x) := \sum_{i=1}^{m} f_{in} x q_{in}, \quad \text{(A7)}$$

where f_{in} is the expected fraction of class i customers with current reservations in period n. If we simplify the above equation, we obtain $q_n(x) = x \sum_i f_{in} q_{in} = x\hat{q}_n$, where $\hat{q}_n := \sum_i f_{in} q_{in}$ can be interpreted as the expected average cancellation rate per customer in period n.

METHOD 2. Define

$$q_n(x) := \sum_{i=1}^{m} f_{in}(x) q_{in}, \quad \text{(A8)}$$

where $f_{in}(x)$ is the expected number of class-i reservations given that the total number of reservations is x.

With the above setup, the optimality Eqs. 4, 5, and 6 hold, and the problem is seen to be equivalent to control of admission to a queueing system. Note that, in Method 2, the service rate $q_n(x)$ is a possibly nonlinear function of x, the number of customers present. As long as $q_n(x)$ is concave and nondecreasing in the state x, however, the inductive argument (Theorem 1 of Lippman and Stidham, 1977) continues to be applicable and implies that the optimal value function $U_n(\cdot)$ is concave and nonincreasing, so that the optimal admission policy is monotonic in the state. When $q_n(x) = \sum_i f_{in} x q_{in} = \hat{q}_n x$ (Method 1), then $q_n(x)$ is a linear function of x and, hence, trivially concave. But when we use Method 2 to estimate $q_n(x) = \sum_i f_{in}(x) q_{in}$, then we have to choose the functions $f_{in}(x)$ carefully so that $q_n(x)$ is concave. We plan to explore this approach in more depth in a future paper.

ACKNOWLEDGMENTS

THE AUTHORS WOULD LIKE to acknowledge the contribution of Anatoly Shaykevich to an earlier version of this paper (Janakiram, Stidham, and Shaykevich, 1994). Anatoly's economic insights inspired us to develop the equivalent charging scheme described in Section 2. We also thank R.L. Phillips (President and CEO, DFI.Aeronomics) and G.C. Campbell (Vice President, DFI.Aeronomics) for valuable assistance in developing the ideas in Section 3.2. We are grateful to G. Colville (currently at Delta Airlines) for giving us permission to use real airline demand and capacity data in our analysis. Finally, we thank Jeff Rutter (DFI.Aeronomics) for help with transferring and manipulating large volumes of airline demand data.

REFERENCES

J. ALSTRUP, S. BOAS, O. MADSEN, AND R. VIDAL, "Booking Policy for Flights with Two Types of Passengers," *Eur. J. Oper. Res.* **27,** 274–288 (1986).

P. P. BELOBABA, "Airline Yield Management: An Overview of Seat Inventory Control," *Transp. Sci.* **21,** 63–73 (1987).

P. P. BELOBABA, "Application of a Probabilistic Decision Model to Airline Seat Inventory Control," *Opns. Res.* **37,** 183–197 (1989).

S. BRUMELLE AND J. MCGILL, "Airline Seat Allocation with Multiple Nested Fare Classes," *Opns. Res.* **41,** 127–137 (1993).

R. CHATWIN, Optimal Airline Overbooking. Ph.D. Dissertation, Department of Operations Research, Stanford University, Stanford, CA, 1992.

R. CURRY, "Optimal Seat Allocation with Fare Classes Nested by Origins and Destinations," *Transp. Sci.* **24,** 193–204 (1990).

G. GALLEGO AND G. VAN RYZIN, "Optimal Dynamic Pricing of Inventories with Stochastic Demand over Finite Horizons," *Management Sci.* **40,** 999–1020 (1994).

G. GALLEGO AND G. VAN RYZIN, "A Multiproduct Dynamic Pricing Problem and its Applications to Network Yield Management," *Operations Research* **45,** 24–41 (1997).

W. E. HELM AND K.-H. WALDMANN, "Optimal Control of Multiserver Queues in a Random Environment," *J. Appl. Prob.* **21,** 602–615 (1984).

S. JANAKIRAM, S. STIDHAM, AND A. SHAYKEVICH, Airline Yield Management via Arrival Control to a Stochastic Input-Output System. Paper presented at ORSA/TIMS National Meeting, Detroit, October, 1994.

S. G. JOHANSEN AND S. STIDHAM, "Control of Arrivals to a Stochastic Input-Output System," *Adv. Appl. Probab.* **12,** 972–999 (1980).

C. LAUTENBACHER AND S. STIDHAM, "The Underlying Markov Decision Process in the Single-Leg Airline Yield-Management Problem," *Transp. Sci.* **33,** 136–146 (1999).

T. LEE AND M. HERSH, "A Model for Dynamic Airline Seat Inventory Control with Multiple Seat Bookings," *Transp. Sci.* **27,** 252–265 (1993).

S. A. LIPPMAN, "Applying a New Device in the Optimization of Exponential Queuing Systems," *Opns. Res.* **23,** 687–710 (1975).

S. A. LIPPMAN, "Countable-State, Continuous-Time Dynamic Programming with Structure," *Opns. Res.* **24,** 477–489 (1976).

S. A. LIPPMAN AND S. STIDHAM, "Individual versus Social Optimization in Exponential Congestion Systems," *Opns. Res.* **25,** 233–247 (1977).

K. LITTLEWOOD, "Forecasting and Control of Passenger Bookings," *AGIFORS Symp. Proc.* **12,** 95–117 (1972).

L. ROBINSON, "Optimal and Approximate Control Policies for Airline Booking with Sequential Nonmonotonic Fare Classes," *Opns. Res.* **43,** 252–263 (1995).

R. SERFOZO, "An Equivalence between Continuous and Discrete-Time Markov Decision Processes," *Opns. Res.* **27,** 616–620 (1979).

M. SHAKED AND J. G. SHANTHIKUMAR, *Stochastic Orders and Their Applications.* Academic Press, San Diego, California, 1994.

B. SMITH, J. LEIMKUHLER, AND R. DARROW, "Yield Management at American Airlines," *Interfaces* **22,** 8–31 (1992).

S. STIDHAM, "Socially and Individually Optimal Control of Arrivals to a GI/M/1 Queue," *Management Sci.* **24,** 1598–1610 (1978).

S. STIDHAM, "Optimal Control of Admission, Routing, and Service in Queues and Networks of Queues: A Tutorial Review," in *Proceedings ARO Workshop: Analytic and Computational Issues in Logistics R and D,* George Washington University, Washington, DC, 330–377, 1984.

S. STIDHAM, "Optimal Control of Admission to a Queuing System," *IEEE Trans. Automat. Control,* **AC-30,** 705–713 (1985).

S. STIDHAM, "Scheduling, Routing, and Flow Control in Stochastic Networks," in *Stochastic Differential Systems, Stochastic Control Theory and Applications*, volume IMA-10, W. Fleming and P. L. Lions (eds) IMA Volumes in Mathematics and its Applications, Springer-Verlag, New York, 529–561, 1988.

S. STIDHAM AND R. WEBER, "A Survey of Markov Decision Models for Control of Network of Queues," *Queueing Syst. Theory Appl.* **13,** 291–314 (1993).

J. SUBRAMANIAN, G. CAMPBELL, AND R. PHILLIPS, Research on Bid-Price Dynamics. Paper presented at INFORMS National Meeting, Atlanta, Georgia, 1996.

K. TALLURI AND G. VAN RYZIN, An Analysis of Bid-Price Controls for Network Revenue Management, Technical Report, Graduate School of Business, Columbia University, 1996.

L. WEATHERFORD AND S. BODILY, "A Taxonomy and Research Overview of Perishable-Asset Revenue Management," *Opns. Res.* **40,** 831–844 (1992).

L. WEATHERFORD, S. BODILY, AND P. PFEIFER, "Modeling the customer arrival process and comparing decision rules in perishable asset revenue management situations," *Transp. Sci.* **27,** 239–251 (1993).

E. WILLIAMSON, Airline Network Seat Control, Ph.D. Dissertation, Flight Transportation Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 1992.

R. WOLLMER, "An Airline Seat Management Model for a Single-Leg Route when Lower Fare Classes Book First," *Opns. Res.* **40,** 26–37 (1992).

Y. YOUNG AND R. VAN SLYKE, Stochastic Knapsack Models of Yield Management, Technical Report 94-76, Polytechnic University, Brooklyn, NY, 1994.

(Received: August 1997; revisions received: April 1998, August 1998; accepted: December 1998)