CE 273
Markov Decision Processes

Lecture 6

Applications of Finite Horizon MDPs - Part II

Previously on Markov Decision Processes

Theorem (DP Algorithm)

The optimal cost $J^*(x_0)$ equals $J_0(x_0)$ which solves

$$J_N(x_N) = g_N(x_N)$$

$$J_k(x_k) = \min_{u_k \in U_k(x_k)} \mathbb{E}_{w_k} \left\{ g_k(x_k, u_k, w_k) + J_{k+1}(f_k(x_k, u_k, w_k)) \right\}$$

$$\forall\, k = N-1, \ldots, 1, 0$$

Further, if $u_k^* = \mu_k^*(x_k)$ minimizes the RHS of the above expression then $\pi^* = \{\mu_0^*, \mu_1^*, \ldots, \mu_{N-1}^*\}$ is optimal.

For certain classes of problems, the optimal policies can be shown to satisfy certain properties, called structural results. Unfortunately, there's no unified theory behind it and it is best understood from multiple examples.

The proof techniques in establishing these results almost always involves induction or recursion! In the following lecture(s), we will study few such problems.

Previously on Markov Decision Processes

In such cases, we can write the state as (x_k, y_k) where x_k is affected by u_k and y_k is not. Let p_i represent the pmf of y_k. In such cases, the DP algorithm can be simplified as

$$\hat{J}_k(x_k) = \sum_{i=1}^{m} p_i J_k(x_k, i)$$

$$\hat{J}_k(x_k) = \sum_{i=1}^{m} p_i \min_{u_k \in U_k(x_k)} \mathbb{E}_{w_k} \left\{ g_k(x_k, u_k, w_k) + \hat{J}_{k+1}(f_k(x_k, u_k, w_k)) | y_k = i \right\}$$

In the case of Tetris, x_k is the board configuration and y_k is the shape of the block. There is no exogenous disturbance and the action uniquely determines the new state. Hence, we can write

$$J_k(x_k) = \sum_{i=1}^{m} p_i \min_{u_k \in U_k(x_k)} \left\{ g_k(x_k, i, u_k) + J_{k+1}(f_k(x_k, i, u_k)) \right\}$$

f_k represents the new board position and g_k could be the number of rows cleared.

Previously on Markov Decision Processes

For the inventory control problem,

$$J_k(x_k) = -cx_k + \min_{u_k \geq 0} \left\{ c(x_k + u_k) + H_k(x_k + u_k) + \mathbb{E}J_{k+1}(x_k + u_k - w_k) \right\}$$

$$= -cx_k + \min_{y_x \geq x_k} \left\{ cy_k + H_k(y_k) + \mathbb{E}J_{k+1}(y_k - w_k) \right\}$$

Let $G_k(y_k) = cy_k + H_k(y_k) + \mathbb{E}J_{k+1}(y_k - w_k)$. $J_k(x_k)$ can be written as

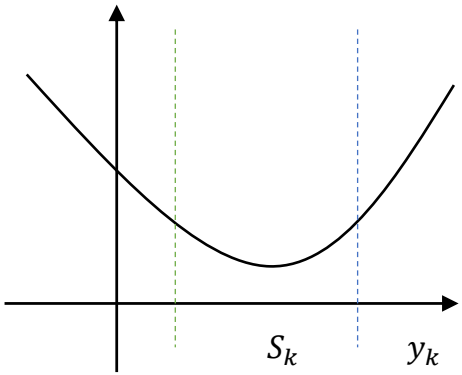
$$J_k(x_k) = -cx_k + \min_{y_k \geq x_k} G_k(y_k)$$

It turns out that $G_k(.)$ is convex for all k! (We will show this shortly.)

Suppose the unconstrained minimum of $G_k(y_k)$ occurs at S_k. What can we say about $\min_{y_k \geq x_k} G_k(y_k)$?

$$\min_{y_k \geq x_k} G_k(y_k) = \begin{cases} G_k(S_k) & \text{if } x_k \leq S_k \\ G_k(x_k) & \text{otherwise} \end{cases}$$

Previously on Markov Decision Processes



Thus, the optimal value function is

$$J_k(x_k) = \begin{cases} -cx_k + G_k(S_k) & \text{if } x_k \leq S_k \\ H_k(x_k) + \mathbb{E}J_{k+1}(x_k - w_k) & \text{otherwise} \end{cases}$$

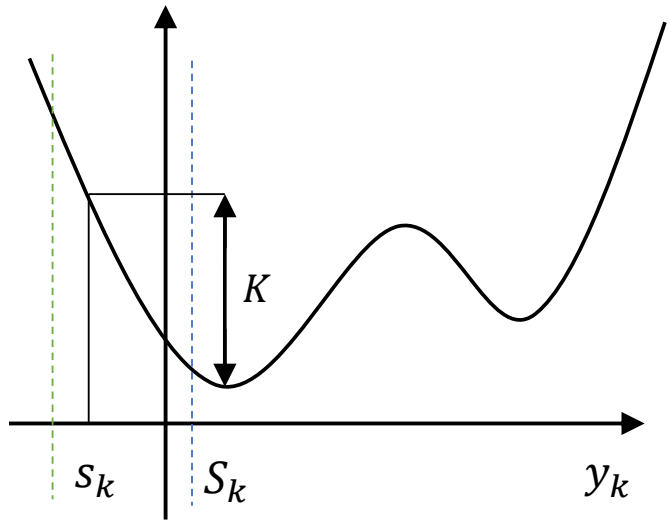
Note that now we've got rid of the min operator in the above expression.

The policy is constructed based on the control values at which the minimum occurs. Recall that $y_k = x_k + u_k$. Thus, when $x_k \leq S_k$, $y_k^* = S_k$ and when $x_k > S_k$, $y_k = x_k$,

$$\mu_k^*(x_k) = \begin{cases} S_k - x_k & \text{if } x_k \leq S_k \\ 0 & \text{otherwise} \end{cases}$$

Previously on Markov Decision Processes

When there are fixed costs, $G_k(.)$ is no longer convex because of the structure of the above equation. One can however identify two distinct regimes in the optimal value functions because it is K-convex.



Let s_k be the smallest y_k for which $G_k(y) = K + G_k(S_k)$.

Case I: When $x_k \leq s_k$,

$$\min\left\{G_k(x_k), \min_{y_k > x_k}\left\{K + G_k(y_k)\right\}\right\} = \min_{y_k > x_k}\left\{K + G_k(y_k)\right\} = K + G_k(S_k)$$

Case II: When $x_k > s_k$,

$$\min\left\{G_k(x_k), \min_{y_k > x_k}\left\{K + G_k(y_k)\right\}\right\} = G_k(x_k)$$

Previously on Markov Decision Processes

The optimal value functions can thus be written as

$$J_k(x_k) = \begin{cases} -cx_k + K + G_k(S_k) & \text{if } x_k \leq s_k \\ -cx_k + G_k(x_k) & \text{if } x_k > s_k \end{cases}$$

Now, let's look at the optimal policy. In Case I, minimum of the RHS occurs when $y_k = S_k$ and hence $u_k = S_k - x_k$. In Case II, the minimum occurs when $u_k = 0$. Thus, we have

$$\mu_k^*(x_k) = \begin{cases} S_k - x_k & \text{if } x_k \leq s_k \\ 0 & \text{if } x_k > s_k \end{cases}$$

Such policies are also called (s, S) policies.

Lecture Outline

1. Secretary Problem
2. Airline Revenue Management

Lecture Outline

Secretary Problem

Secretary Problem

Introduction

You are to choose a secretary from N potential candidates. The candidates have a **true ranking** but it is not known to you unless you interview everyone.

At any stage you can rank the candidates you have interviewed so far (**relative ranking**).

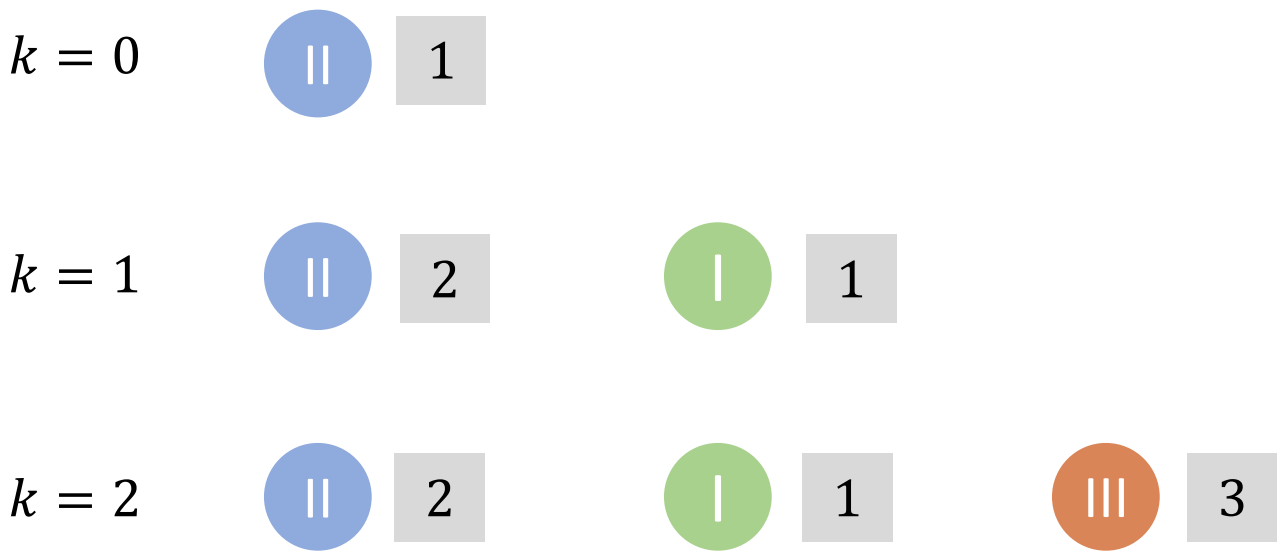
After interviewing a candidate you can either give them the job (in which case you do not interview more candidates) or continue interviewing (in which case you cannot hire a previously interviewed candidate).

The objective is to maximize the probability of picking the top true-ranked candidate. Formulate this as an MDP and solve it.

Secretary Problem

Introduction

For instance, if $N = 3$, and the three candidates are I, II, and III, with true ranking $1, 2$, and 3 respectively.



We will let $k \in \{1, 2, \ldots, N\}$ instead of starting from 0 since there are N candidates and we need to make only $N - 1$ decisions.

Secretary Problem

Observations

The MDP formulation is not very straightforward because of the objective. But first, note that

- For the above objective, it is not optimal to select a candidate who was just interviewed and whose relative rank is > 1. (Why?)
- So if we decide to stop, the recently interviewed candidate must have a relative rank 1.

http://www.randomservices.org/random/apps/SecretaryGame.html

Let the state variable x_k be 1 if the current candidate has a relative rank 1 and 0 otherwise.

Additionally, we need to account for the case in which the interview processes is terminated at some intermediate step. Therefore, we assume that x_k can equal T, which represents a stopped state.

Secretary Problem

MDP Formulation

Control:

Suppose in each time step, we can either stop (S) the interview process or continue (C) to interview one more candidate. For all time periods except the last,

$$U_k(x_k) = \begin{cases} \{\mathsf{S}, \mathsf{C}\} & \text{if } x_k \in \{0, 1\} \\ \emptyset & \text{if } x_k = \mathsf{T} \end{cases}$$

Although, it is not optimal to choose S when $x_k = 0$, we will allow this action and let the Bellman equations take care of optimality.

Disturbance:

We will define w_k as a Bernoulli random variable that is 1 if the subsequent candidate at time $k+1$ has a relative rank 1 among the first $k+1$ and is 0 otherwise. What is the pmf of w_k?

$$\mathbb{P}[w_k = 0] = k/(k+1)$$
$$\mathbb{P}[w_k = 1] = 1/(k+1)$$

Secretary Problem

MDP Formulation

Dynamics:

$$x_{k+1} = f_k(x_k, u_k, w_k) = \begin{cases} w_k & \text{if } u_k = \mathsf{C} \\ \mathsf{T} & \text{if } u_k = \mathsf{S} \text{ or } x_k = \mathsf{T} \end{cases}$$

Costs:
One-step costs are incurred only when we stop the interview process. They need to be carefully designed so that

$$\mathbb{E}\left\{ g_N(x_N) + \sum_{k=0}^{N-1} g_k(x_k, u_k, w_k) \right\}$$

corresponds to the probability of selecting the best candidate. Hence, we assume that costs are incurred *exactly once when the interview ends*.

- **Terminal Costs:**

$$g_N(x_N) = \begin{cases} 1 & \text{if } x_N = 1 \\ 0 & \text{if } x_N \in \{0, \mathsf{T}\} \end{cases}$$

- **One-step Costs:**

$$g_k(x_k, u_k, w_k) = g_k(x_k, u_k) = \begin{cases} k/N & \text{if } x_k = 1 \text{ and } u_k = \text{S} \\ 0 & \text{otherwise} \end{cases}$$

  Why k/N? The probability that the top true-ranked candidate is the kth candidate that you've interviewed is $\binom{N-1}{k-1}/\binom{N}{k}$.

Since we wish to maximize the probability of selecting the best candidate, the DP algorithm can be re-written as,

$$J_N(x_N) = g_N(x_N)$$

$$J_k(x_k) = \max_{u_k \in U_k(x_k)} \mathbb{E}_{w_k} \left\{ g_k(x_k, u_k, w_k) + J_{k+1}(f_k(x_k, u_k, w_k)) \right\}$$

$$\forall\, k = N-1, \ldots, 1$$

Secretary Problem

Bellman Equations

When $x_k \neq \mathrm{T}$,

$$J_k(1) = \max_{u_k \in \{\mathrm{S},\mathrm{C}\}} \mathbb{E}_{w_k} \left\{ g_k(1, u_k) + J_{k+1}(f_k(1, u_k, w_k)) \right\}$$
$$= \max \left\{ \frac{k}{N} + J_{k+1}(T), \frac{k}{k+1} J_{k+1}(0) + \frac{1}{k+1} J_{k+1}(1) \right\}$$

$$J_k(0) = \max_{u_k \in \{\mathrm{S},\mathrm{C}\}} \mathbb{E}_{w_k} \left\{ g_k(0, u_k) + J_{k+1}(f_k(0, u_k, w_k)) \right\}$$
$$= \max \left\{ 0 + J_{k+1}(T), \frac{k}{k+1} J_{k+1}(0) + \frac{1}{k+1} J_{k+1}(1) \right\}$$
$$= \frac{k}{k+1} J_{k+1}(0) + \frac{1}{k+1} J_{k+1}(1)$$

Thus, $J_k(1)$ can also be written as

$$J_k(1) = \max \left\{ \frac{k}{N}, J_k(0) \right\}$$

Secretary Problem

Bellman Equations

The Bellman equations are not very difficult to solve since there are only a handful of state-control pairs.

However, there are a neat structural result that can be uncovered using an inductive argument.

Proposition

The optimal policy is reject the first η candidates and then select the first top relative-ranked candidate.

In order to show this, we will prove a related proposition. The exact value of η is a function of N. Also, let $N \geq 3$ so that the trivial cases can be ignored.

What is the value of η when $N = 3$?

Secretary Problem

Bellman Equations

Proposition

If it is optimal to choose C at some time j, then it is optimal to choose C for all previous time periods.

Proof.

Since it is optimal to choose C at j,

$$J_j(1) = \max\left\{\frac{j}{N}, J_j(0)\right\} = J_j(0)$$

Thus, either $J_j(0) = J_j(1) > j/N$ or $J_j(0) = J_j(1) = j/N$. Using the Bellman equation, for time step $j-1$,

$$J_{j-1}(0) = \frac{j-1}{j}J_j(0) + \frac{1}{j}J_j(1) = J_j(0) \geq \frac{j}{N} > \frac{j-1}{N}$$

$$J_{j-1}(1) = \max\left\{\frac{j-1}{N}, J_{j-1}(0)\right\} = J_{j-1}(0) > \frac{j-1}{N}$$

Thus, it is optimal to choose C at time $j-1$. Going backwards, recursively, we can conclude that C is optimal for all $i < j$. ■

Secretary Problem

Optimal Value Functions

Therefore, we will never have a policy that looks like

k	1	...	k'	...	k''	...
x_k	1	...	1	...	1	...
$\mu_k^*(x_k)$	C	...	S	...	C	...

Note that the above policy is still admissible even though it prescribes to choose S since the policy does not keep track of history of actions. It just tells what to do at different states.

The optimal action when x_k is 0 is always C and hence has been omitted from the above table.

In other words, if it is optimal to choose S at some stage k', for all future periods when the state is 1, the optimal action must be S.

Secretary Problem

Optimal Value Functions

We can analytically solve the optimal value functions. Suppose, we reject the first η candidates and select the first relative-ranked candidate,

Case I: $k \leq \eta$

$$J_k(1) = J_k(0)$$

Further, when $k < j$, $J_k(1) = J_{k+1}(1)$. (Why?) Use the expression for value function of $J_k(0)$. Hence,

$$J_1(0) = J_1(1) = \ldots = J_\eta(0) = J_\eta(1)$$

Case II: $k > \eta$

$$J_k(1) = \frac{k}{N}$$

To compute $J_k(0)$, we use backward induction. For $k = N$, $J_N(0) = 0; J_N(0) = 1$. For $k = N - 1$,

$$\begin{aligned} J_{N-1}(0) &= \frac{N-1}{N} J_N(0) + \frac{1}{N} J_N(1) \\ &= \frac{1}{N} = \frac{N-1}{N} \frac{1}{N-1} \end{aligned}$$

For $k = N - 2$,

$$\begin{aligned} J_{N-2}(0) &= \frac{N-2}{N-1} J_{N-1}(0) + \frac{1}{N-1} J_{N-1}(1) \\ &= \frac{N-2}{N-1} \frac{1}{N} + \frac{1}{N} = \frac{N-2}{N} \left\{ \frac{1}{N-1} + \frac{1}{N-2} \right\} \end{aligned}$$

For $k > \eta$,

$$J_k(0) = \frac{k}{N} \left\{ \frac{1}{N-1} + \frac{1}{N-2} + \ldots + \frac{1}{k} \right\}$$

Secretary Problem

Optimal Value Functions

We now know the analytical expressions for $J_k(0)$ and $J_k(1)$ for all k. Recall that

$$J_k(1) = \max\left\{\frac{k}{N}, J_k(0)\right\}$$

Since we stop when the state is 1 after rejecting η candidates, when $k = \eta + 1$,

$$J_{\eta+1}(1) = \frac{\eta+1}{N} \geq J_{\eta+1}(0)$$

Since, $\eta + 1 > \eta$, using the results from Case II,

$$J_{\eta+1}(0) = \frac{\eta+1}{N}\left\{\frac{1}{N-1} + \frac{1}{N-2} + \ldots + \frac{1}{\eta+1}\right\}$$

Substituting this in the above inequality,

$$\frac{\eta+1}{N} \geq \frac{\eta+1}{N}\left\{\frac{1}{N-1} + \frac{1}{N-2} + \ldots + \frac{1}{\eta+1}\right\}$$

Secretary Problem

Optimal Value Functions

Thus, the number of candidates to reject before selecting the candidate with relative rank 1 is the smallest integer for which

$$\frac{1}{N-1}+\frac{1}{N-2}+\ldots+\frac{1}{\eta+1}\leq 1$$

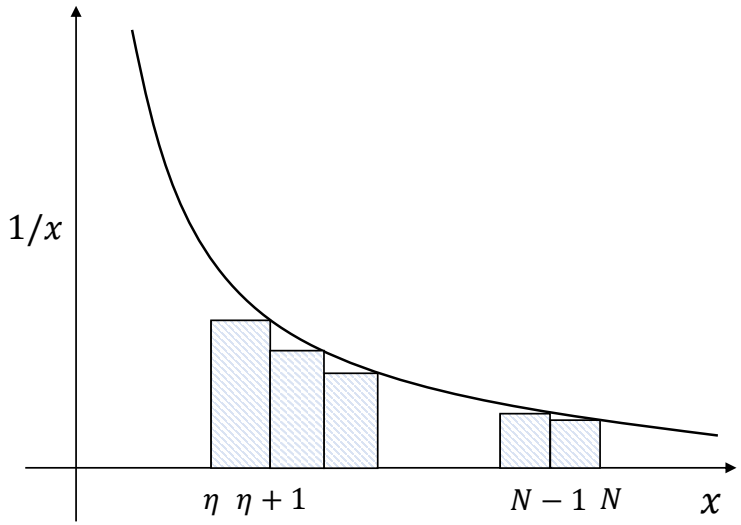
What is the value of η for $N=5$? What if $N\to\infty$? We can approximate the above inequality as

$$\frac{1}{N}+\frac{1}{N-1}+\frac{1}{N-2}+\ldots+\frac{1}{\eta+1}\approx 1$$

which can be written as

$$\int_{\eta}^{N}\frac{1}{x}\approx 1\Rightarrow ln(N/\eta)\approx 1$$

$$\Rightarrow \eta/N\approx e^{-1}=0.3679$$



Thus, when N is large, it is optimal to reject 36.79% of the candidates and then select the top relative-ranked one!

Secretary Problem

Historical Notes

The Secretary Problem is nearly 150 years old. A variant of it was first posed by Arthur Cayley in 1875!

The objective of maximizing the probability of a certain event is very common to several other MDPs, especially those that involve game playing.

The model we saw is also an example of what is called as an **optimal stopping problem** in which the problem ends after a certain action is chosen and the goal is to determine when to stop.

Additional Reading:

Ferguson, T. S. (1989). Who solved the secretary problem?. Statistical science, 4(3), 282-289.

Lecture Outline

Airline Revenue Management

Airline Revenue Management

Introduction

We will now explore structural results for a simple revenue management problem with a name-your-own-price (NYOP) feature (e.g., Priceline).

Consider a single BLR-DEL flight with a seat capacity C. Suppose

- There are n fare classes with prices $p_1, p_2, \ldots, p_n$ such that $p_1 \geq p_2 \geq \ldots \geq p_n$.
- The time period of interest is divided as $0, 1, \ldots, N-1, N$, where N is the time at which the flight takes off.
- At each time-step, at most one customer arrives and makes an offer from the above denominations to purchase one seat.
- No overbooking is allowed.

The airlines must decide whether to accept the offer or not and has an objective of maximizing expected revenue.

The different price classes can also reflect options like refundable/non-refundable/partially-refundable segments in a non-NYOP model.

Similar models can be used for markets in which purchases are to be made before a deadline such as hotels.

The probability with which a customer makes an offer p_j is assumed to be known q_j. These are estimated from historic data and are called booking curves.

It is not necessary that a customer arrives in a particular time period. Hence, we assume that $\sum_{j=1}^{n} q_j \leq 1$.

Airline Revenue Management

Introduction

Airlines can reserve a certain number of seats for different classes of passengers. These are called **protection levels**.

They can also calculate **booking limits**, which is the number of tickets that the airlines is willing to sell for various classes of customers.

Say an Airbus A320 has 180 seats. Suppose there are 4 classes with $p_1 \geq p_2 \geq p_3 \geq p_4$. Then, a sample protection levels and booking limits could be

Class	Protection Level	Booking Limit
1	40	180
2	60	140
3	60	80
4	-	20

Protection levels are sometimes defined by including the reserved capacity of higher classes. For e.g., the protection level for Class 2 is 100. (We'll use this version.)

One could use these types of measures to determine whether to accept or reject offers. What are the disadvantages?

Airline Revenue Management

MDP Formulation

State:
Let (x_k, y_k) represent the remaining capacity and the offer price made by a customer at time k. If no customer arrives at time k, $y_k = 0$.

Note that the state has an uncontrollable component like Tetris. The control chosen does not affect y_k.

Control:
Let the decline and accept actions be denoted using the variables 0 and 1.

$$U_k(x_k) = \begin{cases} \{0, 1\} & \text{if } x_k > 0 \\ 0 & \text{if } x_k = 0 \end{cases}$$

Disturbance:
w_k is the offer price made by a customer in the next time period.

$$\mathbb{P}[w_k = p_j] = q_j \, \forall \, j = 1, \ldots, n$$

$$\mathbb{P}[w_k = 0] = 1 - \sum_{j=1}^{n} q_j$$

Airline Revenue Management

MDP Formulation

Dynamics:

$$(x_{k+1}, y_{k+1}) = f_k(x_k, u_k, w_k) = (x_k - u_k, w_k)$$

Rewards:

Terminal rewards are zero since the seats have no value when the flight takes off. For all other time periods,

$$g_k((x_k, y_k), u_k, w_k) = y_k u_k$$

The Bellman equations can be written as

$$J_N((x_N, y_N)) = 0 \,\forall\, x_N, y_N$$

$$J_k((x_k, y_k)) = 0 \,\forall\, x_k = 0, k \in \{0, 1, \ldots, N-1\}$$

$$J_k((x_k, y_k)) = \max_{u_k \in \{0,1\}} \left\{ y_k u_k + \mathbb{E} J_{k+1}((x_k - u_k, w_k)) \right\}$$

$$\forall\, x_k > 0, k \in \{0, 1, \ldots, N-1\}$$

Why does the one-step reward not have an expectation?

Airline Revenue Management

Optimality Conditions

Since, a part of the state is uncontrollable, we can define an ex ante value function

$$\hat{J}_k(x_k) = \sum_{j=1}^{n+1} q_j J_k((x_k, j))$$

where $j = n+1$ is used to represent the no customer situation. Also define the expected marginal value of one additional seat,

$$\Delta \hat{J}_k(x) = \hat{J}_k(x) - \hat{J}_k(x-1)$$

In other words, it is the expected benefit from not selling a seat/having an extra seat when we have x unsold seats at time step k.

When $x_k > 0$, we can thus rewrite the Bellman equations as

$$\begin{aligned}
\hat{J}_k(x_k) &= \sum_{j=1}^{n+1} q_j \max_{u_k \in \{0,1\}} \left\{ p_j u_k + \hat{J}_{k+1}(x_k - u_k) \right\} \\
&= \sum_{j=1}^{n+1} q_j \max_{u_k \in \{0,1\}} \left\{ p_j u_k + \hat{J}_{k+1}(x_k) - \hat{J}_{k+1}(x_k) + \hat{J}_{k+1}(x_k - u_k) \right\}
\end{aligned}$$

Airline Revenue Management

Optimality Conditions

$$= \hat{J}_{k+1}(x_k) + \sum_{j=1}^{n+1} q_{jk} \max_{u_k \in \{0,1\}} \left\{ p_j u_k - \hat{J}_{k+1}(x_k) + \hat{J}_{k+1}(x_k - u_k) \right\}$$

$$= \hat{J}_{k+1}(x_k) + \sum_{j=1}^{n+1} q_{jk} \max_{u_k \in \{0,1\}} \left\{ \left(p_j - \Delta \hat{J}_{k+1}(x_k)\right) u_k \right\}$$

Thus, if a customer offers a price p_j, then it is optimal to accept iff

$$p_j \geq \Delta \hat{J}_{k+1}(x_k)$$

That is, if the price offered is greater than the expected reward from saving the seat, the airlines must accept the offer.

Airline Revenue Management

Structural Results

It turns out that $\Delta\hat{J}_k(x)$ has a few interesting properties.

Proposition

The value functions $\Delta\hat{J}_k(x)$ satisfy

1. $\Delta\hat{J}_k(x+1) \leq \Delta\hat{J}_k(x)$
2. $\Delta\hat{J}_{k+1}(x) \leq \Delta\hat{J}_k(x)$

The first condition implies that the marginal value of one extra seat decreases, i.e., they must be more valuable as they get scarce.

The second condition implies that the marginal value of an extra seat at x seats decreases with time, i.e., a seat is more worthy if there is more time to sell it.

It can also be shown that $\hat{J}_k(x)$ is concave in x for all k! (in a discrete sense)

Airline Revenue Management

Structural Results

Hence, we can derive time-dependent protection levels λ_{jk} (seat-capacity protected for classes $j, j-1, \ldots, 1$) using

$$\lambda_{jk} = \max\left\{ x \;\middle|\; p_j < \Delta \hat{J}_{k+1}(x) \right\}$$

Likewise, booking-limits (number of tickets the airlines is willing to sell to class j customers) can be derived using

$$\beta_{jk} = C - \lambda_{j-1,k}$$

Since $p_1 \geq p_2 \geq \ldots p_{N-1}$, the protection levels and booking limits are nested. These results are clearly more insightful than plain look-up tables.

Airline Revenue Management

Structural Results

Additional Reading:

Subramanian, J., Stidham Jr, S., & Lautenbacher, C. J. (1999). Airline yield management with overbooking, cancellations, and no-shows. Transportation science, 33(2), 147-167.

Your Moment of Zen