Single Resource Capacity Allocation

Part 1

The Context

We are a passenger airline.

We sell seats on air flights accepting bookings.

Our goal is to maximize revenue.

We are subject to capacity constraints, e.g. a certain flight has

Capacity = 100 seats

The basic formula is

Revenue = Price x Quantity

where Quantity is the number of seats sold.

In order to maximize Revenue we can use both levers, price and Quantity.

For the moment, we are going to use only Quantity.

The idea is to divide the available capacity in two or more blocks, allocating each block to a different combination of time, space, sale channel and customer segment.

For example, we can offer seats with these policies:

- 40 seats today, reserving 60 seats for tomorrow;
- 30 seats for passengers form Rome and 70 from Milan ;
- 50 seats for males and 50 for females;
- 80 seats at most for adults, reserving 20 for young people;
- combinations of policies like above.

Each capacity block has an *Expected Revenue*: it is defined afterwards.

We want to maximize the expected revenue for each capacity block.

Prices are assumed as given, for the moment.

The Two-Class Problem

A flight with fixed capacity serves two classes of customers:

1. *Discount customers* who book early.
2. *Full-fare customers* who book later.

What is the rational of such a scheme?

We know some customers are less price-sensitive, e.g. customers flying for business reasons. Others are more price-sensitive, e.g. those travelling for tourism.

We also know that type 1 customers book later, type 2 early.

This scheme is learnt form experience: it is not exact, but approximates reality well enough.

We are using booking time as indicator of what is really important: price-sensitiveness.

The idea is simple: late-booking people are likely to be less price-sensitive, so we offer them seats for a higher price.

Discount customers pay a fare p_d, full-fare customers pay $p_f > p_d$.

Prices p_d and p_f are given: we chose them for some reason, but here it does not matter what reason.

We assume that all discount booking request occur before any full-fare request.

The flight has a limited capacity (i.e. number of seats).

Basic problem: how many discount booking request should we accept at most? This number is named *booking limit*.

Equivalent formulation: how many seats should we protect for full-fare customers? This number is named *protection level*.

Booking limit b and protection level y are bound by the equivalence $y = C - b$, where C is the capacity.

Let $C = 100$: the flight has 100 seats.

If we decide to accept at most 70 request for discount fare p_d during the first round of booking, then $b = 70$ and $y = 30$.

If we receive 65 bookings in the first round for the discount fare p_d , then we accept all them. If we receive 75 bookings, we accept 70 and refuse 5.

Equivalently, we can start reserving 30 seats for the second round, when the fare is p_f. This means to set $y = 30$ and consequently $b = 70$.

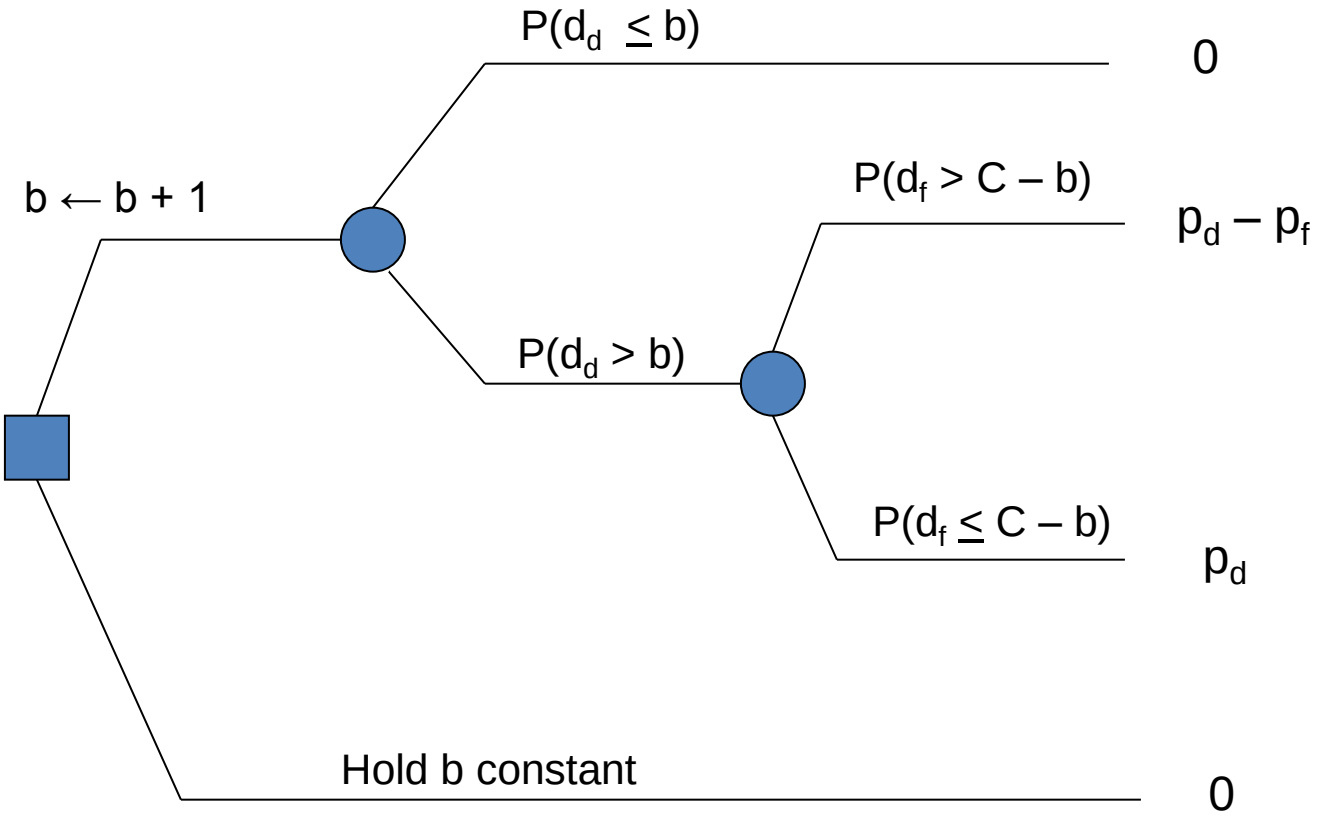
It is important to note that seats available for booking in the first round but left unsold are available for booking in the second round.

If $b = 70$ and in the first round we get 65 bookings, we accept them and we still have 35 seats available in the second round.

The key point is the trade-off between two risks:

1. *Spoiling*. Setting booking limit too low, we will turn away discount passenger requiring a seat in the first round. If in the second round we do not see enough full-fare demand to fill the plane, it will depart with empty seats (spoiled seats).
2. *Dilution*. Setting booking limit too high, we will allow too many customers to book at discount price in the first round. If in the second round we see a full-fare demand exceeding still available seats, then we will turn away more profitable passengers, available to pay the full fare.

The trade-off is shown in the graphical form of a decision tree in the following picture.


$P(d_d \leq b)$
0
$b \leftarrow b + 1$
$P(d_f > C - b)$
$p_d - p_f$
$P(d_d > b)$
$P(d_f \leq C - b)$
p_d
Hold b constant
0

The question is *what is the additional revenue brought by an increment of the booking limit* b?

We are speaking of *additional* revenue, also known as *marginal revenue*, the revenue added by one more unit.

Leaves of the tree show the variation in revenue caused by a decision.

The square node is a *decision node*: we can choose between incrementing *b* to *b + 1* and holding *b* constant.

Circle nodes are *event nodes*: something outside our control happens with a certain probability, and we face some consequences.

The decision of holding *b* constant carries no uncertainty: with probability 1, i.e. with certainty, we gain a marginal revenue of 0.

It is trivial: our action is changing nothing, the consequence is getting nothing different.

Probability 1 is omitted for brevity, we could have written it on the edge of the tree.

More interesting is the decision of incrementing the booking limit by one more seat.

Now we accept one more booking: if there is a *(b + 1)th* request, then we accept it, while without the increment we would have refused it.

With d_d and d_f we denote the demand (number of bookings) in the first and second round, respectively.

If $d_d <= b$, then incrementing the booking limit has no effect at all. The additional available seat is not requested, so nothing different happens. The marginal revenue is 0.

If $d_d > b$, then the additional available seat is actually requested, and our decision causes one more seat sold in the first round.

Our revenue in the first round with the increment decision is incremented by p_d (the price of a seat in the first round) compared with our revenue in the first round without that decision.

This is good news. Is there bad news? Maybe.

It depends on d_f, the demand level in the second round.

If $d_f <= C - b$, then we will have no regret: the *(b + 1)th* seat sold in the first round due to our decision would otherwise have been unsold.

No bad news: our decision gives us an additional revenue today without any loss tomorrow. Actually, we have reduced *spoilage*.

Instead, if $d_f > C - b$, then we will have regret: the seat additionally sold today (because our decision) at price p_d would have been sold tomorrow at greater price p_f.

In this scenario we pay an *opportunity cost*: selling today costs a lost sale tomorrow. We incurred in *dilution*.

These considerations justify marginal revenue figures in the picture: note that the second leave shows a negative marginal revenue.

The expected marginal revenue is the differential value of our decision, i.e. the expected additional revenue due to our decision of incrementing the booking limit.

We compute it as the sum of possible outcomes (i.e. marginal values on leaves of the decision tree), each weighted with the probability of this outcome to occur.

Let us denote the action of incrementing *b* by one seat with *h(b)* and the expected value of our decision with *E[h(b)]*.

Let us define

$$F_d(x) = P(d_d \leq x)$$

This is the *distribution probability function* of the demand in the first round. It gives the probability of demand not exceeding *x* bookings. Analogously, $F_f(x) = P(d_f \leq x)$.

The formula for expected marginal revenue of incrementing the booking limit can be written as

$$E[h(b)] = F_d(b)0 + [1 - F_d(b)]\{\left[1 - F_f(C - b)\right](p_d - p_f) + F_f(C - b)p_d\}$$

Simplifying:

$$E[h(b)] = [1 - F_d(b)]\{p_d - \left[1 - F_f(C - b)\right]p_f$$

If the term on the right-hand side is greater than zero, then increasing the booking limit we increase the expected revenue.

If it is less than zero, increasing the booking limit we decrease the revenue.

Our decision criterion is:

Increase the booking limit from b *to* b + 1 *if and only if*:

$$E[h(b)] = [1 - F_d(b)]\{p_d - \left[1 - F_f(C - b)\right]p_f > 0$$

By definition, *[1 − F_d(b)]* cannot be negative, because it is a probability, more precisely the probability of demand in the second round exceeding *b*.

It could be zero if it were impossible to have a demand exceeding *b*: let us assume it is not the case (the line of reasoning is not really affected by this assumption).

We can suppress that term and state the decision criterion again:

Increase the booking limit from b *to* b + 1 *if and only if*:

$$p_d - \left[1 - F_f(C - b)\right]p_f > 0$$

Rewriting the equation, we get:

Increase the booking limit from b *to* b + 1 *if and only if*:

$$1 - F_f(C - b) < \frac{p_d}{p_f}$$

In natural language, this means what follows.

Incrementing the booking limit is a good decision (incrementing expected revenue) if and only if the probability of the demand in the second round exceeding the number of seats left excluded from booking in the first round is less than the ratio between the discount and the full fare.

Note that the ratio on the right-hand side is between 0 and 1, which is consistent with the left-hand side being a probability.

To reach a more intuitive formulation, let us define

$$1 - F_f(C - b) \stackrel{\text{def}}{=} R$$

The interpretation of *R* is the risk of dilution, i.e. the risk of tomorrow regret of having sold the *(b+1)th* seat today.

The decision criterion is now

Increase the booking limit from b *to* b + 1 *if and only if*:

$$Rp_f < p_d$$

In natural language, we require the certain (probability 1) additional revenue of today p_d be greater than the possible (probability R) tomorrow opportunity cost p_f.

We are comparing two revenue figures, one certain and one uncertain. Or, if you prefer to think so, we are comparing a revenue figure with a cost figure.

In both case, we are doing an inter-temporal comparison.

Indeed, our decision causes a conflict between our interested of today (selling one more seat) with our interest of tomorrow (having one more seat available for sale).

You can also see it as a conflict between the interest of today-us and the interest of tomorrow-us.

In some sense, this is an auction: the first and the second round are the bidders, we are the auctioneer assigning the seat to the better bid.

The key point to understand is that we are *maximizing the expected revenue*.

We imagine two scenarios:

1. We hold booking limit constant.
2. We increment booking limit by 1 unit.

For each scenario we estimate:

a. The probability it happens.

b. The revenue it brings if it happens.

The expected revenue of each scenario is the product of *a* and *b* points.

The scenario with the greater expected revenue is the winner. We assign the marginal unit of resource, i.e. the *(b + 1)th* seat, to the winner.

The *marginalistic analysis* gave us a criterion to choose between incrementing or not incrementing a certain level of booking limit.

Now we know how to decide if it is better to hold constant a certain given b or to increment it to $b + 1$.

Is this sufficient to find the optimal value of b?

This is our ultimate goal: to find the booking limit giving us the maximum possible revenue.

The answer is *yes*. Using this criterion we can find the optimal booking limit.

Let us start with booking limit $b = 0$.

We ask if it better to increment b to 1. The decision criterion says *yes*.

Now we state $b \leftarrow 1$.

The next step is to ask if it is better to increment b to 2.

At each step, if the answer is positive, then we increment b.

We eventually stop because:

- Either $b = C$, i.e. the booking limit reached the capacity. A further increment would be desirable, but it is not possible.
- Or the decision criterion says *no*. A further increment is not desirable.

In both cases, the last level of b we reached is the optimal one.

Let us name this last booking limit as b^*.

It is either

$$b^* = C$$

or

$$b^* = \max x \text{ such that } 1 - F_f(C - x) < p_d / p_f$$

The concept is very general:

We are in the optimal situation if each unitary step in any direction is impossible or makes our situation worse.

Actually, this intuitive statement is true only if some conditions are satisfied. In many practical problems it is the case. The problem at hand, finding the optimal booking limit, belongs to the class of lucky cases, where these conditions are satisfied.

Hill-climbing Algorithm

The previous marginalistic analysis suggests an algorithm for computing the optimal booking limit.

1. Set b ← 0.
2. If b = C, set b* ← C and stop.
3. Compute $E[h(b)] = p_d - [1 - F_f(C - b)]p_f$.
4. If $E[h(b)] \leq 0$ or $F_d(b + 1) = 1$ set b* ← b and stop.
5. If $E[h(b)] > 0$ and $F_d(b + 1) < 1$, set b ← b + 1 and stop.

The concept is simple: start with booking limit 0 and increment it by one until another increment decrease the expected revenue, or the capacity is completely used.

This kind of algorithm is named *hill-climbing*.

At each step we increment something (our decision lever, here the booking limit) if this move increase the outcome (here the expected revenue).

We stop as soon as a further move is not useful.

These algorithms are generally simple to implement, but suffer of a major drawback: they reach a *local optimum* (best among immediate neighbors), not necessarily a *global optimum* (best among all).

For some problems, there exist only one optimum, which is both local and global. The two-class booking limit problem belongs to this class of problems, so we can use it safely.

Littlewood's Rule

The previous algorithm find the value b^* such that

$$1 - F_f(C - b^*) = \frac{p_d}{p_f}$$

Using the optimal protection level y^*, the equivalent formula is

$$1 - F_f(y^*) = \frac{p_d}{p_f}$$

We previously saw this equation, which is known as *Littlewood's Rule*. It is of big historical importance: in 1972 it was the seminal result for a theory which is now extremely rich and complex.

To better grasp the intuitive meaning of the rule, think of its behavior when the price ratio tends to 0 or 1, or when the F_f tends to 0 or 1.

The cumulative distribution function of the full-fare demand F_f can be expressed as a mathematical function or empirically with a data table.

If we can compute its inverse F_f^{-1}, then the optimal protection level can be explicitly represented as

$$y^* = \min[{F_f}^{-1}\left(1 - \frac{p_d}{p_f}\right), C]$$

Note that the optimal protection level (or, equivalently, the optimal booking limit) does not depend on the forecast of discount demand. The optimal choice is simply the reserve for the second booking round exactly the number of requests that will arrive, leaving remaining seats for the first round.

If the full-fare is 300€ and the discount fare is 150€, then the optimal protection level is the number of bookings which has probability 50% of being exceeded during the second booking round.

If the cumulative distribution function F_f is an analytical probability distribution (e.g. a Gaussian Normal Distribution), then we can compute the protection level using well-known methods from probability calculus.

If it is represented as data table, we choose the value giving circa 50% as probability of excess.

Even if we do not have a data table at disposition, the rule helps us providing a clear question: find an *y* whose risk of excess is about 50% and choose it as protection level.

Newsvendor Problem

Littlewood's Rule for Capacity Allocation is a special case of the Newsvendor Problem, studied as early as 1888.

A newspaper salesperson needs to determine how many newspapers to purchase at the beginning of the day to satisfy the day's uncertain demand.

He faces different costs if he purchases too much or too little:

- Too many copies: at the end of the day some copies are unsold and worthless.
- Too few copies: he will sell out and turn away potentially profitable customers.

We name *overage cost* the cost per unit of purchasing too many items, and *underage cost* the unit cost of purchasing too few.

If the newsvendor buys newspapers for 20 cents and sells them at 25 cents per unit, then the overage cost is $O = 20$ and the underage cost is $U = 5$.

The overage cost is an actual monetary cost.

The underage cost is virtual, an opportunity cost, i.e. the profit missed due to non-optimal decision ($U = 25 – 20$).

The optimal order quantity is y^* satisfying

$$F(y^*) = \frac{U}{U + O}$$