Chapter 5

New Demand Management Approaches

In this chapter, we address the identified shortcomings of the TRM, aATP and IR models and show new approaches designed specifically for the characteristics of MTS manufacturing. First, we show an approach based on revenue management ideas and show the optimal demand fulfillment policy. However, for large problem sizes this approach soon reaches its limits since it requires extensive calculation time. In addition, we propose an approximative model based on an adapted version of the SOPA approach which uses ideas from the network revenue management literature, called randomized linear programming (RLP). Due to the high applicability of LP models, randomized linear programming combines a conventional LP with stochastic demand information.

5.1 A Revenue Management Approach

5.1.1 Model formulation

Recall our setting of a make-to-stock manufacturing system facing demand from multiple customer classes. Customer classes differ in the price per unit that they pay. Scheduled inventory replenishments are known. Given this information, the manufacturer decides for each order whether to satisfy it from stock, backorder it at a penalty cost, or reject it. The objective is to maximize profit over a finite planning horizon, taking into account sales revenues, inventory holding costs, and backorder penalties.

In order to achieve this goal the manufacturer has to make a trade-off for each order whether to satisfy it or whether to save the supply for potentially more profitable future orders. We make the following assumptions to model this situation.

Assumption 5.1. Orders from a given customer class follow a compound Poisson process. The order processes of different classes are mutually independent and they are independent of the available supply.
Let $\lambda$ denote the expected number of orders of the Poisson process in a given period, then the probability that exactly $k$ orders arrive can be calculated with

$$f(k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}. \quad (5.1)$$

The Poisson assumption is common in many RM models, specifically in so-called dynamic demand models (see Talluri and van Ryzin (2004)). As Lautenbacher and Stidham (1999), we allow non-unit order sizes, which appears appropriate in a manufacturing environment. In our analysis, we discretize the planning horizon in such a way that the probability of receiving multiple orders within a single period is negligible. Let $T$ denote the length of the planning horizon and $t$ the period index. Moreover, let $c = 1, \ldots, C$ identify the different customer classes.

**Assumption 5.2.** *Inventory replenishments are exogenous and known.*

This assumption reflects the APS planning hierarchy. Inventory replenishments are determined in mid-term and short-term production planning and then serve as input for order promising decisions. Let $x_i$ be the available supply arriving at the beginning of period $i$, $i = 1, \ldots, T$ and let $\bar{x} = (x_1, \ldots, x_T)$ be the vector of all of these replenishments. Note that at time $t$, $x_i$ corresponds with inventory on-hand if $i \leq t$ and with a future scheduled replenishment otherwise. In APS terminology, $\bar{x}$ denotes the ATP quantities.

**Assumption 5.3.** *Order due dates are equal to order arrival times, but orders can be backlogged at a price discount.*

This assumption reflects the MTS context. Customers expect immediate delivery, in principle. Late deliveries are only acceptable at a price discount. Let $r_c$ denote the unit revenue from satisfying an order of class $c$ from stock. Delaying an order gives rise to unit backorder costs $b$ per period. Analogously, holding costs $h$ are incurred for all units of inventory on hand at the end of a period. Note that unit backorder and holding costs are independent of time and customer class.

**Assumption 5.4.** *Partial order fulfillment is allowed.*

This assumption includes splitting an order for partial delivery in different periods. This is a technical assumption, which we need for tractability. We discuss its impact and potential relaxations later on. Let $d$ denote the order quantity, $u_i$ the amount of supply arriving in period $i$ used to satisfy a given customer order, and let $\bar{u} = (u_1, \ldots, u_T)$. Note that for an order arriving in period $t$, $u_i$ corresponds with delivery from stock if $i \leq t$ and with backlogging otherwise.
Table 5.1 summarizes the above notation. We can now formulate our problem as a stochastic dynamic program with state variable $\bar{x}$ and decision variable $\bar{u}$. In principle, one can drop all entries from these vectors, for which $x_i = 0$. I.e., the dimension of the state space corresponds with the number of scheduled replenishments. However, for ease of notation we use $\bar{x}$ and $\bar{u}$ as defined above.

The profit $\hat{P}_t(\bar{x}, d, c, \bar{u})$ earned in period $t$ depends on the available supply, order size, customer class, and fulfillment decision as follows

$$
\hat{P}_t(\bar{x}, d, c, \bar{u}) = r_c \sum_{i=1}^{T} u_i - b \sum_{i=1}^{T} u_i (i - t) (1 - \delta_{it}) - h \sum_{i=1}^{T} (x_i - u_i) \delta_{it}, \tag{5.2}
$$

where $\delta_{it}$ is defined as 1 if $i \leq t$ and 0 otherwise, and $\bar{u}$ has to satisfy $u_i \leq x_i$ for all $i$ and $\sum_i u_i \leq d$.

The first term in Equation 5.2 calculates the revenues received from satisfying the current order of class $c$. The second term computes backlogging costs that occur when using supply that arrives later than the customer order, i.e. when $\delta_{it} = 0$. These costs are computed for the total length of the delay $(i-t)$. The third term represents holding costs that are charged for the on-hand inventory at the end of period $t$. Note that unlike the backlogging costs, which are charged for the total customer waiting time, holding costs only cover the current period $t$. To simplify subsequent calculations, we define $P_t(i, c)$ as the incremental profit per unit of supply $i$ used to satisfy one unit of an order of class $c$ in period $t$. Collecting the terms in Equation 5.2 that depend on $u_i$ yields

$$
P_t(i, c) = r_c - b(i - t)(1 - \delta_{it}) + h \delta_{it} \tag{5.3}
$$

and

$$
\hat{P}_t(\bar{x}, d, c, \bar{u}) = \sum_{i=1}^{T} P_t(i, c) u_i - h \sum_{i=1}^{T} x_i \delta_{it}. \tag{5.4}
$$

In addition to the current period’s profit, we also have to take into account the impact of a fulfillment decision $\bar{u}$ on future profits. The state transition is given by $\bar{x} \rightarrow \bar{x} - \bar{u} = (x_1 - u_1, \ldots, x_T - u_T)$. Letting $V_t(\bar{x})$ denote the maximum
Table 5.1: Notation of the Revenue Management Approach

<table>
<thead>
<tr>
<th>Indices:</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t = 1, \ldots, T )</td>
</tr>
<tr>
<td>( i = 1, \ldots, T )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>State variables:</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{x} = (x_1, \ldots, x_T) )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Decision variables:</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{u} = (u_1, \ldots, u_T) )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random variables:</th>
</tr>
</thead>
<tbody>
<tr>
<td>( c )</td>
</tr>
<tr>
<td>( d )</td>
</tr>
<tr>
<td>( F(c, d) )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data:</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r_c )</td>
</tr>
<tr>
<td>( b )</td>
</tr>
<tr>
<td>( h )</td>
</tr>
</tbody>
</table>

Expected profit-to-go from period \( t \) to the end of the planning horizon \( T \) for a given supply vector \( \bar{x} \) we then obtain the following Bellman recursion

\[
V_t(\bar{x}) = E_{d,c} \left[ \max_{0 \leq u_i \leq x_i, \sum_{i=1}^{T} u_i \leq d} \left\{ \sum_{i=1}^{T} (u_i P_t(i,c) - h x_i \delta_{it}) + V_{t+1}(\bar{x} - \bar{u}) \right\} \right] \tag{5.5}
\]

with the boundary condition \( V_{T+1}(\bar{x}) = 0 \).

5.1.2 Structural properties and optimal policy

We now analyze structural properties of the value function of the dynamic program defined in the previous section. This will then allow us to characterize the optimal fulfillment policy. All proofs of this section are given in the appendix. We start by defining marginal profits:

**Definition 5.1.** \( \Delta_i V_t(\bar{x}) := V_t(\bar{x}) - V_t(\bar{x} - \bar{e}_i) \) for \( x_i \geq 1 \),

where \( \bar{e}_i \) denotes the \( i \)-th unit vector. Definition 5.1 concerns the expected marginal value of a unit of supply arriving in period \( i \) or, equivalently, the
opportunity costs of selling this unit. Using this definition, we can rewrite the
Bellman recursion of Equation 5.5 as follows.

\[
V_t(\bar{x}) = E_{d,c} \left[ \max_{0 \leq u_t \leq x_t, \sum_{i=1}^{T} u_t \leq d} \left\{ \sum_{i=1}^{T} (u_i P_t(i, c) - h x_i \delta_{it}) + V_{t+1}(\bar{x} - \bar{u}) \right\} \right]
\]

\[
= E_{d,c} \left[ \max_{0 \leq u_t \leq x_t, \sum_{i=1}^{T} u_t \leq d} \left\{ \sum_{i=1}^{T} (u_i P_t(i, c) - h x_i \delta_{it}) \right. \right.
\]

\[
+ V_{t+1}(\bar{x}) - \sum_{i=1}^{T} \sum_{z=1}^{u_i} \Delta_i V_{t+1} \left( \bar{x} - \bar{\epsilon}_i(z - 1) - \sum_{j=1}^{i-1} (\bar{\epsilon}_j u_j) \right) \left. \right] \right]
\]

\[
= V_{t+1}(\bar{x}) - h \sum_{i=1}^{T} x_i + E_{d,c} \left[ \max_{0 \leq u_t \leq x_t, \sum_{i=1}^{T} u_t \leq d} \left\{ \sum_{i=1}^{T} \left( \sum_{z=1}^{u_i} (P_t(i, c) \right) \right. \right.
\]

\[
- \Delta_i V_{t+1}(\bar{x} - \bar{\epsilon}_i(z - 1) - \sum_{j=1}^{i-1} (\bar{\epsilon}_j u_j)) \left. \right] \right].
\]

(5.6)

Note that this formulation decomposes \( \bar{u} \) into single-unit steps. In this way, the maximization in Equation 5.6 reflects the trade-off between the profit of selling a unit of supply now and the corresponding opportunity cost. Also note that a similar decomposition is well-known for the classical single-leg airline yield management problem (see Talluri and van Ryzin (2004), page 59). What is different in Equation 5.6 is the summation over \( i \), which introduces an additional dimension into the problem.

We now identify properties of the value function that help us evaluate the above maximization expression. The first step is to compare the marginal values of supplies arriving in different periods.

**Proposition 5.1.** For all \( m < n \) and for all \( \bar{x} \) with \( x_m, x_n \geq 1 \) the value function satisfies:

a) \( \Delta_m V_t(\bar{x}) - \Delta_n V_t(\bar{x}) \leq b(\max(n, t) - \max(m, t)) = b(n - m + \delta_{nt}(t - n) + \delta_{mt}(m - t)) \)

b) \( V_t(\bar{x} + \bar{e}_m) - V_t(\bar{x} + \bar{e}_n) \leq b(n - m + \delta_{nt}(t - n) + \delta_{mt}(m - t)) \).

Proposition 5.1 states that the difference between the marginal value of one unit of a supply arriving in period \( m \) and one unit arriving in period \( n \) is bounded by the difference in backordering costs of using each of these supplies in period \( t \). This relationship implies the following important monotonicity property, regarding the alternative fulfillment options.

\[ \Delta \bar{u} \leq \Delta \bar{\epsilon} \]
Proposition 5.2. For all $m < n$ and for all $x$ with $x_m, x_n \geq 1$ it holds that
\[ P_t(m, c) - \Delta_m V_{t+1}(\bar{x}) \geq P_t(n, c) - \Delta_n V_{t+1}(\bar{x}), \forall c. \]

The terms on the left-hand-side of the inequality can be interpreted as the net benefit of the current revenues from selling a unit of supply arriving in period $m$ minus the opportunity cost of not having that unit available in the future. Proposition 5.2 states that this net benefit is decreasing in the arrival time of the supply. Therefore, an order should always be either fulfilled using the earliest available supply or not at all (if the left-hand-side becomes negative). The next important property concerns the concavity of the value function along certain axes.

Proposition 5.3. Let $\bar{x}$ be such that $\sum_i x_i \geq 2$. Furthermore, let $m = \min\{i | x_i > 0\}$ and let $n = m$ if $x_m > 1$ and $n = \min\{i | i > m, x_i > 0\}$ otherwise. Then
\[ P_t(m, c) - \Delta_m V_{t+1}(\bar{x}) \geq P_t(n, c) - \Delta_n V_{t+1}(\bar{x} - \bar{e}_m), \forall c. \]

Proposition 5.3 implies in particular that the value function $V_t$ is concave in the quantity of the earliest available supply. This is intuitive since one would expect available supply to have decreasing marginal benefits.

The above properties allow us to characterize the optimal fulfillment policy. The optimal policy turns out to be a generalization of the well-known booking-limit policies in traditional revenue management (Talluri and van Ryzin (2004)). We summarize this result in the following theorem.

Theorem 5.1. Define the following set of critical levels:
For $i = 1, \ldots, T$ let $y_i = \bar{x} - \sum_{k=1}^i \bar{e}_k x_k,$ and let $L_t(c, i, y_i) = \max\{k | P_t(i, c) < \Delta_i V_{t+1}(y_i + k\bar{e}_i)\}.$
Then the following fulfillment decision is optimal in period $t$, given an order quantity $d$ from customer class $c$:
Start with $i = 1$;
\[
\text{set } u_i = \max\left( \min\left( x_i - L_t(c, i, y_i), d - \sum_{k=1}^{i-1} u_k \right), 0 \right); \\
\text{if } u_i < x_i \text{ set } u_k = 0 \text{ for all } k > i \text{ and stop, otherwise repeat for } i+1.
\]

Intuitively, this policy successively consumes units of supply, in the order of their arrival, until the immediate marginal profit drops below the opportunity costs. This very much resembles a traditional booking-limits policy. The values $L_t(c, i, ..)$ set nested protection levels that bar some amount of supply $i$ from consumption by classes $c$ and higher. Note that we have separate protection levels for each class and supply arrival. Also note that the protection levels of supply $i$ depend on the available quantities of subsequent arrivals $y_i$. However, $L_t(c, i, ..)$ is independent of $x_i$ (and of all earlier arrivals $x_j$ for $j < i$) and therefore indeed acts as a protection level. The amount of supply $i$ exceeding
5.1 A REVENUE MANAGEMENT APPROACH

$L_t(c, i, \ldots)$ is available for consumption by customer class $c$ at time $t$. It is worth pointing out that even the most valuable customer class, i.e. $c = 1$, may face non-trivial booking limits for future supplies, i.e. for $i > t$: While this class can always consume supply on hand it is not necessarily optimal to backlog demand from this class, due to the incurred backorder penalties. We illustrate the various protection levels $L_t(c, i, \ldots)$ graphically in an example in the next section.

The following proposition shows that, as in the classical case, protection levels decrease in time. This is intuitive since a shorter remaining planning horizon implies less selling opportunities and therefore available supply is of less value. It is worth mentioning, however, that this result is only true for stationary demand. Unlike in traditional revenue management models, the holding cost term in our model may destroy the monotonicity of the protection levels if the demand distribution changes across periods.

**Proposition 5.4.** The protection levels $L_t(c, i, \bar{y}_i)$ defined in Theorem 5.1 are non-increasing in $t$.

5.1.3 A Numerical Example

Fig. 5.1 illustrates the protection levels of supply one for different customer classes and different levels of the second supply. For example, the line "$c_2, s_2 = 0$" shows the protection levels for class two and no remaining quantities of the second supply. In the first period, 31 units of supply one are protected from consumption of class two (and, accordingly, class three). As the value of supply

![Figure 5.1: Non-Increasing Protection Levels](image-url)
decreases as times goes on, the protection levels decrease. At the end of the planning horizon (period 27), no supplies have to be protected anymore. As seen in Proposition 5.4, protection levels are non-increasing in time.

Proposition 5.3 states that the value function is concave. This behavior is illustrated in Fig. 5.2 which shows the values of $V_t(x)$ in period 27 for different remaining quantities of the second supply. In case of $s_2 = 0$ and small amounts of the first supply, each additional unit of supply one contributes much to the expected profit which explains the huge slope of the line from $s_1 = 0$ to 20. After the maximum is reached the slope of the line becomes negative as holding costs are expected that reduce the expected profit. In the extreme case of $s_2 = 100$ each additional unit of supply never contributes in a positive way to the profit and the line steadily decreases.

5.2 Approximations Based on Linear Programming

The idea behind the models presented in this section originates from the work on network revenue management, as described in Talluri and van Ryzin (2004). In network revenue management, resources are bundled together to form the products that customers buy. As an example, airlines offer so-called "origin-destination-fares" (ODF), which often consist of several single flight legs. The complexity of network RM results from the fact that single flight legs might be part of more than one ODF. If two customers want to buy the same flight leg,
for example the first customer from Vienna to Chicago over Amsterdam and the second customer from Vienna to Amsterdam, it has to be decided which customer brings more profit.

Many factors contributed to the development of these approaches. First, optimal solutions of realistically-sized networks can not easily be computed as the curse of dimensionality prevents fast computations. Second, as airlines usually have an enormous amount of demand information available, approximating the optimal solution via deterministic approaches would simply mean a waste of resources. Third, methods of linear programming are known to be efficient and are easily to be implemented. As an introduction to the randomized linear programming approach of Section 5.2.2, we first introduce a deterministic linear programming model (DLP) which also serves as an approximation method to the optimal solution but does not take into account demand variability.

5.2.1 Deterministic Linear Programming

The following deterministic linear programming model resembles the allocation planning model described in Sect. 4.3.2, only differing in the demand constraint (5.8). Instead of restricting the allocated quantities to the demand forecast, the deterministic linear program uses the expected demand \( E[d_{kT}] \) and is formulated as shown in the following.

\[
\max \sum_{k=1}^{K} \sum_{t=1}^{T+1} \sum_{\tau=1}^{T} p_{k\tau} z_{k\tau} \tag{5.7}
\]

subject to

\[
da_{k\tau}^{\min} \leq \sum_{t=1}^{T+1} z_{k\tau} \leq E[d_{kT}] \quad \forall k, \tau = 1, \ldots, T \tag{5.8}
\]

\[
\sum_{k=1}^{K} \sum_{\tau=1}^{T} z_{k\tau} + f_t = ATP_t^1 \quad \forall t = 1, \ldots, T. \tag{5.9}
\]

Despite the fact that the DLP is very efficient to solve and is easily applicable in practical settings, it is not the final answer for stochastic problems as it has an important disadvantage: the DLP neglects demand variations and simply considers the expected demand. All information included in demand distributions are not taken into account.

To illustrate the weaknesses of the DLP approach, consider a case with extreme supply shortage, two customer classes and a large difference between the profits of class one and class two. The DLP will certainly allocate the highest
possible number of units to class one—as this is the most profitable class—which equals the expected demand of this class. The rest will be allocated to the less profitable class two. In the case that the demand of class one exceeds the expected demand, the DLP results in poor profits since not all class one customers are satisfied. The profit loss in not satisfying class one customers is substantial as class two generates only very low profits compared to class one.

This example resembles the situation in the famous newsvendor problem. Here, one has to decide how many newspapers to order on the day before the actual demand is realized. It is assumed that the distribution of the demand is known when the decision about the quantity is made. An intuitive solution to this problem is to order exactly the expected demand as the probability of resulting in too high or too low inventory is minimized. However, this intuitive solution does not always yield the highest profit, only in case the costs of having too much or too less are equal. In a situation in which it is very important to satisfy all customers because costs for loss of goodwill are high, the optimum order quantity will be higher than the expected demand. In the following numerical example, we show that this argumentation is also valid for a simple allocation planning and order promising problem.

**Example 5.1.** Consider a two-customer class problem with an expected demand per period of class one and two of \(E[D] = 5\). Available supply is 10 units per period. For the sake of simplicity, all units that are not consumed are lost at the end of a period. We consider three scenarios of different profitabilities: (1) class one yields a profit of \(p_1 = 5\), class two of \(p_2 = 1\), (2) \(p_1 = 2\) and \(p_2 = 1\) and (3) \(p_1 = 1\) and \(p_2 = 1\). Furthermore, we illustrate the effects of different allocation strategies, from the one extreme making all units available to the first class to the other extreme case of allocating everything to the second class. We do not consider nesting, i.e. class one is not allowed to consume from class two. To calculate profitability of each scenario and allocation strategy, we simulated a Poisson distributed demand stream over 10,000 periods.
5.2 APPROXIMATIONS BASED ON LINEAR PROGRAMMING

Fig. 5.3 illustrates the effects of the different allocation strategies. In case of equal profitabilities, the best allocation strategy is the one from the DLP approach, which means allocating the expected demand quantities. However, if class one is more profitable as class two, it is beneficial to make more units available to class one. In these cases, the DLP strategy does not perform well as the optimum moves away from the expected demand quantities. These results are in compliance with the previously discussed newsvendor problem.

5.2.2 Randomized Linear Programming

We have seen that the optimal solution may not be reached by simply inserting the expected demand in the AP step. The RLP approach combines the easy to use LP formulation of the DLP, and the available stochastic demand information. The idea is to repetitively solve the DLP, not with the expected, but with a random demand drawn from the known stochastic demand distribution. Let \( D_{ktr} \) denote a random variable following a known stochastic distribution of the demand quantity. Then we can formulate the randomized linear programming problem as

\[
H_t(D_{ktr}^i) = \max \sum_{k=1}^{K} \sum_{t=1}^{T+1} \sum_{\tau=1}^{T} p_{ktr} z_{ktr}^i \tag{5.10}
\]

subject to

\[
d_{ktr}^{min} \leq \sum_{t=1}^{T+1} z_{ktr}^i \leq D_{ktr}^i \quad \forall k, \tau = 1, \ldots, T \tag{5.11}
\]

\[
\sum_{k=1}^{K} \sum_{\tau=1}^{T} z_{ktr}^i + f_t = ATP_t^i \quad \forall t = 1, \ldots, T. \tag{5.12}
\]

Note that the resulting optimal solution \( H_t^*(D_{ktr}^i) \) is a random variable. The optimal solution of the stochastic problem can be approximated by calculating the expected value \( E[H_t^*(D_{ktr}^i)] \). We estimate the allocated quantities \( z_{ktr} \) by using the weighted average of the resulting quantities over \( N \) simulation runs, as shown in 5.13.

\[
z_{ktr} = \frac{\sum_{i=1}^{N} H_t^*(D_{ktr}^i) \times z_{ktr}^i}{\sum_{i=1}^{N} H_t^*(D_{ktr}^i)} \tag{5.13}
\]

In the subsequent numerical analysis of the RLP approach, we choose \( N = 30 \) which showed robust results in some preliminary tests. This is in line with...
the chosen parameters of Talluri and van Ryzin (1999, p. 213), although he mentions that this behavior might also be problem dependent.

Numerical studies of RLP applications in the airline industry do not show promising results. It is often the case that the DLP approach dominates the RLP solution. De Boer et al. (2002) analyzed this result in more detail and found an answer to this problem. The authors state that “... the observed domination of the deterministic model over probabilistic techniques is a fortunate by-product of ignoring the uncertainty related to demand. This phenomenon is based on nesting”.

As the RLP approach is an alternative version for the allocation planning step, we still have to decide about how to consume the resulting aATP quantities. In Section 4.3.3 we discussed different consumption rules and mentioned that SOPA_D might be fragile to stochastic demand streams. In contrast, SOPA_A aggregates the aATP quantities in order to compensate forecast errors. Thus, we choose SOPA_D as the consumption rule after running the RLP because RLP already considers stochastic demand.