
LECTURE NOTES ON PRODUCTION AND DISTRIBUTION SYSTEMS PLANNING

Ömer S. Benli

California State University, Long Beach

obenli@csulb.edu

2002 -2.0.1

CONTENTS

LIST OF FIGURES	iii
PREFACE	v
1 INTRODUCTION TO MODELING AND OPTIMIZATION IN PRODUCTION AND INVENTORY SYSTEMS	1
1.1 The Nature of Production-Inventory Processes	1
1.2 Continuous vs. Periodic Review Models	3
1.3 Form of Cost Functions	7
1.4 Certainty and Time Dependence of Parameter Estimates	10
1.5 Interaction Among Items: Single vs. Multi Item Models	11
1.6 Concluding Remarks	16
2 BASIC MODELS OF PRODUCTION PLANNING	17
2.1 Static Models	17
2.2 Dynamic Models	21
2.3 Multi Item Models with Demand Interaction	28
2.4 Multi Item Models with Resource Interaction	30
2.5 Concluding Remarks	31
3 PRODUCTION SCHEDULING	35
3.1 Introduction	35
3.2 Sequencing and Scheduling Problems	37
3.3 Scheduling Algorithms and Complexity	39
3.4 Disjunctive Graph Representation	43
3.5 Concluding Remarks	47
3.6 Problems on Scheduling and Sequencing	48
4 LOCATION AND DISTRIBUTION	53

ii LECTURE NOTES ON PRODUCTION AND DISTRIBUTION SYSTEMS PLANNING

4.1 Facility Location Models	54
4.2 Discrete Location Problems	56
4.3 Vehicle Routing and Scheduling	60
REFERENCES	61

LIST OF FIGURES

Chapter 1

1.1	Inventory-Time Plots	5
1.2	A Periodic Review Model	6
1.3	Cost Functions	8
1.4	Product Structures	14
1.5	A Serial Manufacturing Process	15
1.6	A Distribution System	15

Chapter 2

Chapter 3

3.1	Reducibility among scheduling problems	43
3.2	Disjunctive graph representation of a 3-job ($J3 \mid \mid o_j \leq 3 \mid C_{max}$).	45
3.3	A (feasible) schedule for the 3-job ($J3 \mid \mid o_j \leq 3 \mid C_{max}$): (a). Digraph and, (b). Gantt Chart representations	46

Chapter 4

PREFACE

Management of production operations comprises planning, coordinating, and executing all activities that create goods and services with the goal of sustaining long-range profitability. Very simply stated, profitability can be expressed as

$$\begin{aligned}\text{PROFIT} &= \text{REVENUE} - \text{COST} \\ &= \text{PRICE} \times \text{OUTPUT} - \text{COST} \times \text{INPUT} \\ &= \text{INPUT} \times (\text{PRICE} \times \text{PRODUCTIVITY} - \text{COST})\end{aligned}$$

where PRODUCTIVITY is defined as the ratio of OUTPUT to INPUT. Changing nature of business competitiveness made PRICE and COST nonnegotiable. This implies that the key to higher profitability is lies in productivity improvement. These notes deal with planning of production and distribution systems which are essential for the productive operation of any manufacturing or service enterprise.

1

INTRODUCTION TO MODELING AND OPTIMIZATION IN PRODUCTION AND INVENTORY SYSTEMS

1.1 THE NATURE OF PRODUCTION-INVENTORY PROCESSES

Inventory exists as a buffer between customer demand and procurement (production or purchasing) activities. The primary function of a production planning and scheduling system is to decide how this demand is to be met. This is achieved by manipulating the timing and the size of production and purchase orders.

The purpose of procurement is to meet demand. If procurement can be done at the same instant and for the same amount for each demand occurrence, then there is no need for inventories. But this, usually, is not technologically feasible or economical (or, both) for obvious reasons. Therefore, inventories are an unavoidable reality in almost every manufacturing and retail activity.

$$\text{PROCUREMENT} \Rightarrow \boxed{\text{INVENTORY}} \Rightarrow \text{DEMAND}$$

Exercise 1.1 *When uncertainty is present, inventories are used as a protection against risk of stockout (i.e., shortage). In an environment that is perfectly predictable, inventory may be needed to take advantage of the economic features of a particular technology, or to regulate the production process to meet the changing trends in demand. Explain and give examples.*

A production-inventory process can be described in terms of

- variables, and their,
- interrelationships.

Variables¹ can be

- *demand, cost, and technology* related, which are assumed to be uncontrollable (“exogenous”),
- *procurement (production or purchase)* related, which is assumed to be controllable (“endogenous”).

The fundamental relationship among these variables is *inventory* or *material balance* which simply states that procurement either goes into meeting demand or goes into inventory, or both; thus ensuring the flow conservation.

$$\boxed{\text{PROCUREMENT}} = \boxed{\text{INVENTORY}} + \boxed{\text{DEMAND}}$$

The other basic relationship is the *resource constraints* which state that the amount of resources used in the procurement activities cannot exceed the available amount of resources.

$$\boxed{\text{RESOURCES USED}} \leq \boxed{\text{RESOURCES AVAILABLE}}$$

In order to be more explicit about these relationships, it will be convenient to state the *variables* as rates.

Demand rate $d(t)$ = number of items demanded per unit time at time t ,

Procurement rate $x(t)$ = number of items produced or purchased per unit time at time t ,

Inventory rate $i(t)$ = number of items being added to (if positive) or removed from (if negative) inventory per unit time at time t .

Thus, a *production-inventory process* can be described as a *collection of interrelated variables* $\{x(t), d(t), i(t) : t \in \mathcal{T}\}$ where \mathcal{T} is the index set of the process. The index set can be in two forms:

Periodic Review Process $\mathcal{T} = \{t_0, t_1, \dots, t_T\}$, a set of discrete time points (discrete process),

Continuous Review Process $\mathcal{T} = \{0 \leq t \leq T\}$, an interval of the real time axis of length T (continuous process).

¹It is customary to refer to uncontrollable variables as *parameters* and to controllable variables as *decision variables* or, simply, as *variables*.

T is called the *planning horizon* of the process. The planning horizon is called finite if $\mathcal{T} < \infty$, and infinite otherwise. The values of the variables are presented in the form of *tables* for the discrete processes and in the form of *functions* for the continuous processes.

The purpose of modeling a production-inventory process is to optimize an objective function subject to the constraints that describe the physics of the process. In most (single-objective) models, the objective is to maximize profits. This is equivalent to minimizing the costs *only* when the revenues are constant; that is, when the total revenue is not a function of the decision variables.

$$\begin{aligned}\max \{\text{PROFITS}\} &= \max \{\text{REVENUES} - \text{COSTS}\} \\ &= \text{REVENUES} + \max \{- \text{COSTS}\} \\ &= \text{REVENUES} - \min \{\text{COSTS}\}\end{aligned}$$

Example 1.1 (Product Mix Decisions) *In a predominantly make-to-stock environment, there may be a number of end products a company can manufacture and sell in a period. Contribution to profit and overhead from the sale of each end product is likely to be different. The problem is to determine the production plan (the “product mix”) that maximizes the total contribution to profit and overhead during the planning horizon, subject to constraints imposed by resource limitations and considering customer orders already in hand and potential sales (“forecasts”).*

On the other hand, in a make-to-order environment, the demand is exactly known and must be met in its entirety². In this case the revenues from the sale of end items is constant and the objective is to minimize costs subject to satisfaction of demand and the constraints imposed by the resource limitations.

1.2 CONTINUOUS VS. PERIODIC REVIEW MODELS

Let us be more specific about the inventory balance relationship and represent it as the *inventory balance equation*. Consider the continuous process with an index set $\mathcal{T} = \{0 \leq t \leq T\}$. For any two distinct time points, $t_k, t_l \in \mathcal{T}, t_k < t_l$, the inventory balance equation is stated as,

$$\int_{t_k}^{t_l} [x(t) - d(t) - i(t)] dt = 0.$$

The following definitions will be needed:

²Even in the case of shortages this applies. Backorders are met at a later date causing additional costs. Planned lost sales are not really applicable in a make-to-order environment.

Cumulative Production

$$X(t) = X(t_0) + \int_{t_0}^t x(\tau) d\tau$$

where $X(t_0)$ is the cumulative production up to time t_0 which is, without loss of generality, assumed to be equal to zero.

Cumulative Demand

$$D(t) = D(t_0) + \int_{t_0}^t d(\tau) d\tau$$

where $D(t_0)$ is the cumulative demand up to time t_0 which is, without loss of generality, assumed to be equal to zero.

Inventory Level

$$I(t) = I(t_0) + \int_{t_0}^t i(\tau) d\tau$$

where $I(t_0)$ is the *initial* or *beginning* inventory level.

Example 1.2 Let the planning horizon, \mathcal{T} , be 20 units of time in a continuous review process; and

$$x(t) = \begin{cases} 10 \text{ units per unit time,} & 0 \leq t \leq 4 \\ 0 \text{ units per unit time,} & 4 \leq t \leq 20 \end{cases}$$

and

$$d(t) = 3 \text{ units per unit time, } 0 \leq t \leq 20.$$

Figure 1.1 illustrates these functions by plotting $x(t)$, $d(t)$, and $i(t)$ separately, and $X(t)$, $D(t)$, and $I(t)$ on the same graph.

Note that $I(t)$ represents the inventory level at time t . As seen in the above example, this level can be negative. When the inventory level, $I(t)$, reaches zero from above, continuing demand cannot be met from actual inventory. Hence, *shortage* occurs. There are two extreme possibilities:

Backlogging If the customer is willing to wait, possibly for some benefit, his demand is *backordered*, to be met at some future date from the production or purchase to take place at that time.

Lost Sales If the customer is *not* willing to wait and will take his business elsewhere, then the benefit to be incurred from meeting that demand is lost.

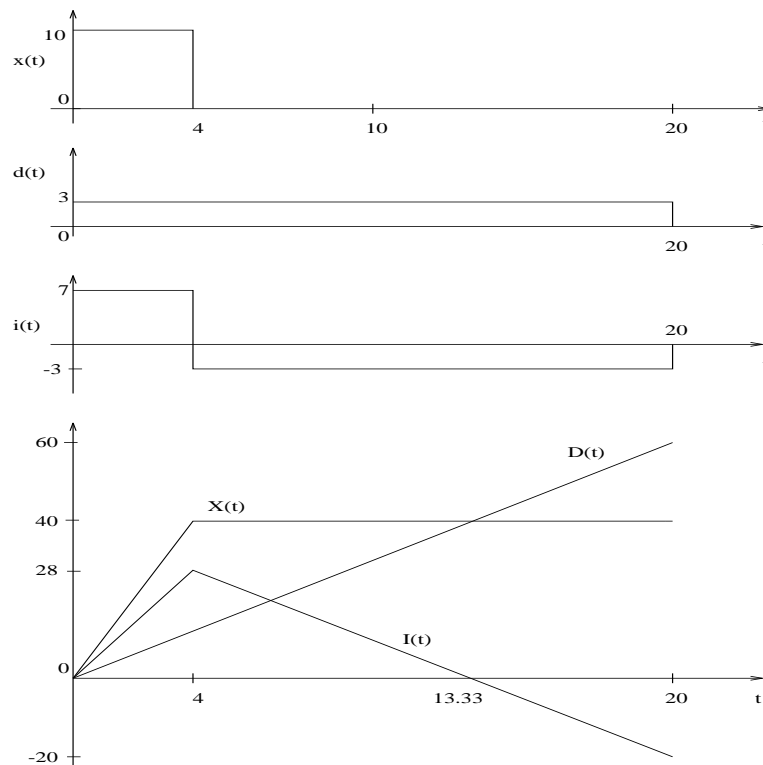


Figure 1.1 Inventory-Time Plots

Between these extremes, other possibilities exist as certain percentage of demand is willing to wait and the remaining will take their business elsewhere.

The fundamental assumption in continuous review models is that procurement decisions can be made at any point during the planning horizon. On the other hand, in the case of *discrete processes*, the index set is composed of discrete time points. The time points of interest are $\{t_0, t_1, \dots, t_T\}$. At all other points the behavior of the process is of no interest. Hence, we want the inventory balance equations to hold only at those points, that is,

$$\int_{t_0}^{t_1} [x(t) - d(t) - i(t)] dt = 0,$$

$$\int_{t_1}^{t_2} [x(t) - d(t) - i(t)] dt = 0,$$

⋮

$$\int_{t_{T-1}}^{t_T} [x(t) - d(t) - i(t)] dt = 0.$$

For ease of notation, define,

$$X_\tau \equiv X(t_\tau) - X(t_{\tau-1}),$$

$$D_\tau \equiv D(t_\tau) - D(t_{\tau-1}),$$

and

$$I_\tau \equiv I(t_\tau).$$

With these definitions and recalling that $X(t_0) \equiv 0$, $D(t_0) \equiv 0$, and $I(t_0)$ being the initial inventory level, we obtain the following inventory balance equations for the periodic review models:

$$X_t - D_t - I_t + I_{t-1} = 0, \quad t = 1, 2, \dots, T.$$

It is customary to refer to these time points, $t = 1, 2, \dots, T$, as *periods*. If each period is considered as a node, then these equations are the node conservation equations. They simply state that, at each period, the beginning inventory plus the production that takes place during that period, must equal to the demand of that period plus the ending inventory. Schematically, a periodic review model can be shown as in Figure 1.2.

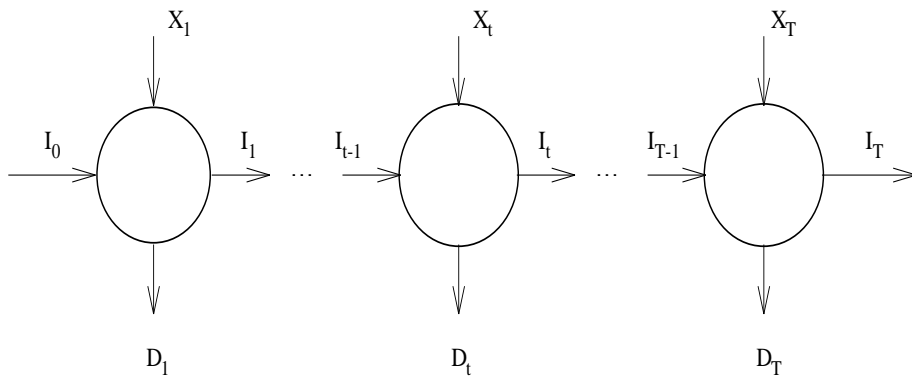


Figure 1.2 A Periodic Review Model

The fundamental assumption in the periodic review models is that the procurement and inventory related costs are accounted only for the state of the process at the points of

interest (i.e. at the end of each period). It should be emphasized that periodic review formulations are not necessarily used as simplified approximations of continuous review models. It may be the case that the real life itself may well be operating according to the periodic review assumptions.

1.3 FORM OF COST FUNCTIONS

In almost every production-inventory model, there are basically two types of costs: procurement related and inventory related. These costs usually expressed as a function of amount produced or purchased and amount stored in inventory, respectively. In general, these functions can take the following forms (See Figure 1.3):

- linear
- piecewise linear (convex, concave, or neither)
- nonlinear (convex, concave, or neither)

Let $X \geq 0$ be the amount produced during a given period and $C(X)$ be the cost of production.

Linear Cost Function

$$C(X) = cX, \quad X \geq 0,$$

where c is the unit cost of production.

Exercise 1.2 Consider the following production planning model:

$$\min_{X, I \geq 0} \sum_{t=1}^T [c_t X_t + h_t I_t]$$

subject to:

$$I_{t-1} + X_t - I_t = D_t, \quad t = 1, 2, \dots, T.$$

Note that inventory and procurement variables are related to each other by the balance equations. In other words, if I tell you the values of all procurement variables, you should be able to tell me the unique values of the inventory variables. So why worry about two sets of variables, one set is enough! Show that the following formulation is equivalent to the formulation given above:

$$\min_{X \geq 0} \sum_{t=1}^T \left(c_t + \sum_{k=t}^T h_k \right) X_t + (\text{constant})$$

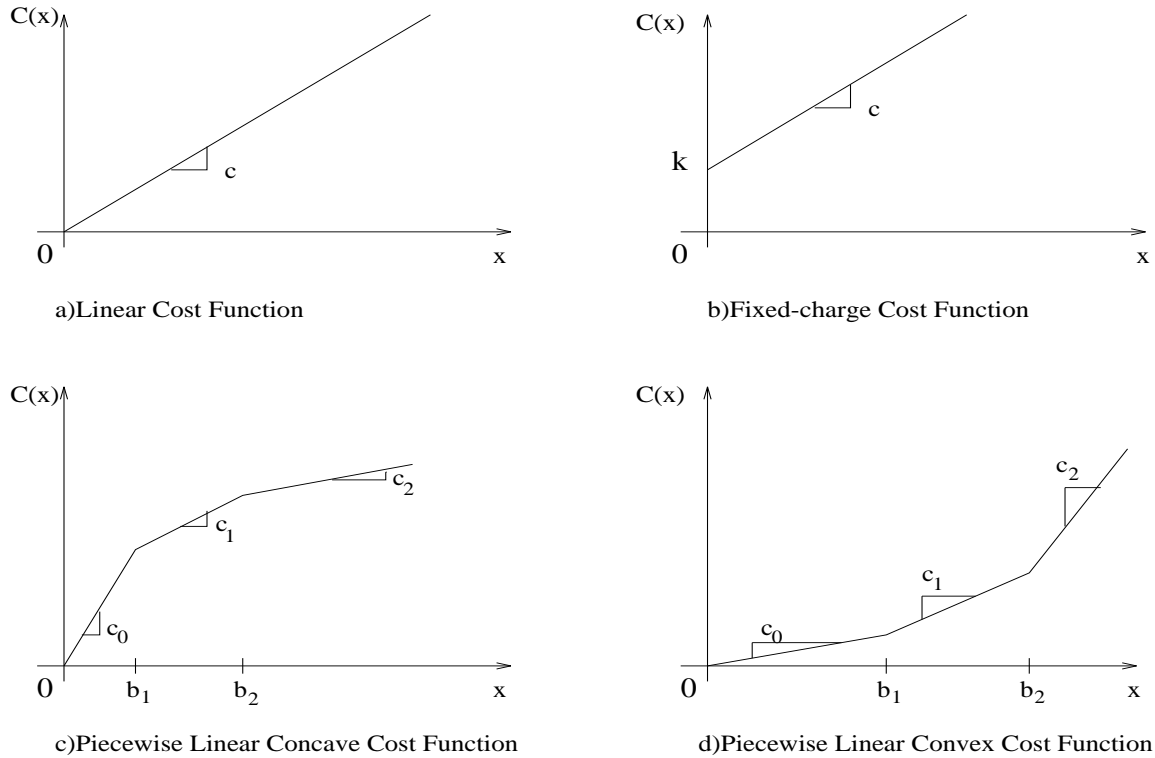


Figure 1.3 Cost Functions

subject to:

$$\sum_{k=1}^t X_k \geq \sum_{k=1}^t D_k - I_0, \quad t = 1, 2, \dots, T.$$

Exercise 1.3 Consider an inventory system where a product is purchased for resale. In period t , the unit purchase price is c_t and the unit selling price is r_t . The product purchased in period t cannot be sold until period $t+1$. Inventory is stored in a warehouse whose capacity is L units. The decision variables in each period are how much to purchase (X_t) and how much to sell (S_t). The planning period is T periods long. Assume that inventory holding costs are negligible. Formulate the problem as a linear program.

Piecewise Linear Cost Function

$$C(X) = a_j + c_j X_j, \quad b_j \leq X \leq b_{j+1}, \quad j = 0, 1, \dots, n,$$

where $a_j = a_{j-1}(c_{j-1} - c_j)b_j$ and $a_0 \equiv 0$. b_1, \dots, b_n are referred to as *breakpoints*. The function is piecewise linear concave if

$$c_0 > c_1 > \dots > c_n$$

and it is piecewise linear convex if

$$c_0 < c_1 < \dots < c_n$$

Example 1.3 (Diseconomies of Scale) *Diseconomies of scale often arise in production planning. This happens when marginal costs are increasing for a minimization problem (or, marginal returns are decreasing for a maximization problem). The practical significance of such a separable piecewise linear convex objective function is that the problem can still be modeled as a linear program (See, for example, pp. 14–18 in [68]).*

Suppose in a particular shop, during a period, the b_1 units can be produced on regular time. An additional $b_2 - b_1$ units can be produced on overtime. Clearly, $c_2 > c_1$. If during the period, more than b_2 units are needed to be procured, then production must be subcontracted to outside suppliers. They are ordered according to increasing unit delivery costs, $c_3 < c_4 < \dots < c_n$, and subcontractors' capacities defining the breakpoints, $b_3 < b_4 < \dots < b_n$.

Let the decision variable be the sum of $n + 1$ auxiliary variables, i.e.,

$$X = X_1 + X_2 + \dots + X_{n+1},$$

where each auxiliary variable is defined over an interval in between the breakpoints, i.e.,

$$0 \leq X_i \leq b_i - b_{i-1}, \quad i = 1, 2, \dots, n + 1.$$

In solving the linear program, the algorithm will always set $X_i = b_i - b_{i-1}$, before taking $X_{i+1} \geq 0$.

Exercise 1.4 (Economies of Scale) *Show that the above reformulation does not work for the concave case, i.e. $c_0 > c_1 > \dots > c_n$.*

Example 1.4 (General Piecewise Representation) *Suppose $\{c_i, i = 1, \dots, n\}$'s are not necessarily increasing. Then, in addition to the auxiliary variables, we need to define n binary variables, $Y_i, i = 1, 2, \dots, n$,*

$$Y_i = \begin{cases} 1 & \text{if } X_i = b_i - b_{i-1}, \text{ i.e. at its upper bound,} \\ 0 & \text{otherwise.} \end{cases}$$

And replace the bound constraints on the auxiliary variables by,

$$(b_i - b_{i-1}) \cdot Y_i \leq X_i \leq (b_i - b_{i-1}) \cdot Y_{i-1}, \quad i = 1, 2, \dots, n.$$

where $Y_0 \equiv 1$ and $Y_n \equiv 0$.

Exercise 1.5 ([16]) *Cost of expanding a plant for the first 4,000 units is \$5 million per 1,000 units of expansion, for the next 6,000 units it is \$1 million per 1,000 units of expansion, and finally it is \$3 million per 1,000 units of expansion for the last 5,000 units. It is not possible to expand the plant by more than 15,000 units. Ignoring the other constraints, formulate the problem as a mixed integer program.*

Exercise 1.6 (Fixed Charge Model) *When a machine is set up to produce a particular item, a setup cost; and when an order is placed, an order cost is incurred. Suppose this fixed cost, k , is independent of the amount procured. If the unit cost of production (or, purchase) is c , the cost of procurement when X units are ordered or produced is given by*

$$C(X) = \begin{cases} k + cX & \text{if } X > 0 \\ 0 & \text{if } X = 0 \end{cases}$$

Show that $C(X)$ is a piecewise linear concave function of X .

1.4 CERTAINTY AND TIME DEPENDENCE OF PARAMETER ESTIMATES

Two major attributes of a production-inventory model are the *certainty* and the *time dependence* of its parameters' estimates. The parameters in these models can be cost, demand, or technology (e.g. input/output) related. Although, in the following, the demand process will be referred to, the discussion equally applies to the other types of uncontrollable variables.

Demand, being a prediction into the future, is estimated as a result of a forecasting effort. Forecasts result either in the expected values or in the form and parameters of a probability distribution. In the terminology of production and inventory systems,

- if *expected values* are forecasted (or, predicted), then the demand process is called *deterministic*,
- if *distributions* are forecasted, the demand process is called *stochastic*.

The process, regardless how it is forecasted, is either *stationary* or *evolutionary*. A stationary process is one whose distribution remains the same as time progresses, because the (random) mechanism producing the process is not changing as time progresses. An evolutionary (sometimes called, the nonstationary) process is one which is not stationary. Thus,

- if the demand process is estimated to be stationary, then the demand is called *static*, and

- if the demand process is estimated to be nonstationary, then the demand is called *dynamic*.

The concept of stationarity is very important, not only in estimating the demand process, but also in estimating the values of all other parameters of the models, such as costs, technological coefficients, etc.

1.5 INTERACTION AMONG ITEMS: SINGLE VS. MULTI ITEM MODELS

Practically all production-inventory systems are multi item processes. There hardly exists a manufacturing or a retail establishment that deals with a single item. George E. Kimball stresses the importance of the concept of an *item* [59]:

In a shoe store, for example, the stock consists of a number of styles of shoes. Within a given style, the shoes come in various sizes. An item here is a pair of shoes of a given style in a given size. The same style in two different sizes makes two different items. Similarly, for an automobile dealer, who carries cars of different models and colors, an item would be a given model in a given color. An oil company processing different crudes must regard each of them as a distinct item.

Items may be distinguished by location alone. If a company carries stock in a number of warehouses, an object must be regarded as an item different from the same object in another warehouse, because moving an object from one location to another requires time and money. In general, two objects are the same item only when they are completely and immediately interchangeable.

It should be remembered that when modeling a production-inventory system as a single item system, the implicit assumption is the nonexistence of any interaction among items in the system. The reasoning behind in analyzing single item models is three-fold:

- Though it may be extremely rare, there are some real life systems that involve a single item,
- Since multi item models can be conceptually complex, analyzing single item models is a useful learning tool before delving into complicated models, and
- In a number of instances, single item models arise as subproblems in analyzing multi item models.

The interaction among items can broadly be classified into,

1. Cost Interaction,
2. Demand Interaction, and
3. Resource Interaction.

Cost interaction occurs in the objective function, demand interaction appears in the inventory balance equations, and resource interaction is the sole reason of existence of the resource constraints.

1.5.1 Cost Interaction

If the cost function³ is completely separable, i.e.,

$$C(X_1, X_2, \dots, X_n) = C(X_1) + C(X_2) + \dots + C(X_n),$$

then there is no cost interaction. Hence, each item $i = 1, 2, \dots, n$ can be treated separately in a single item model, provided that there is no other type of interaction among items.

One type of cost interaction occurs when there exists a discount on the total dollar amount of an order from one supplier regardless of which items are ordered. The other, more common, type of cost interaction is when there exist a “major” and a “minor” fixed charge (setup or order) cost. Suppose a major setup cost, say K_j , incurred when items in a particular group, j , is to be processed in a machining center, and minor setup costs, k_{ji} , incurred for processing each item i in group j . Similar situation may arise in an ordering environment, when items are ordered simultaneously, savings occur in freight, paperwork, and material handling.

Exercise 1.7 *By defining appropriate binary variables, write down a cost function and the related constraints that accounts for both “major” and “minor” fixed charge costs.*

1.5.2 Resource Interaction

In every production-inventory system in which there are more than one item, they have to share limited resources. These resources may be machine time, capital, labor, warehouse space, etc. Therefore, in every model for such systems, there must be resource constraints. If, in a particular model, a resource constraint does not appear, the implicit assumption must have been that ample resources of that type exist in the system.

³This function implies both procurement and inventory related costs.

Consider a periodic review system with N items that share K resources during a planning horizon of T periods. Then, if the additivity and the linearity assumptions hold, the resource constraints can be written as,

$$\sum_{j=1}^N r_{kjt} X_{jt} \leq R_{kt}, \quad k = 1, 2, \dots, K; \quad t = 1, 2, \dots, T.$$

where,

r_{kjt} amount of resource k required in the unit production of item j in period t ,

X_{jt} amount of item j to be produced in period t ,

R_{kt} amount of resource k available in period t .

Exercise 1.8 Show that, under the linearity assumption, in a single item model resource constraints reduce to upper bounds on the procurement amount of the item.

1.5.3 Demand Interaction

In most production-inventory systems, items are related to each other either in the form of *supply-demand* relationship (in order to produce one item you need another item as an input), or in the form of *input-output* relationship (items are stored at different levels, or echelons, such as factory warehouse, wholesaler, retailer, customer; or, when items require different stages of manufacturing)⁴.

This type of interaction arises as a result of *product structures*. A number of product structures are shown in Figure 1.4. These product structures can represent a production process describing how a product is formed from raw materials, purchased items, subassemblies, and assemblies (Figure 1.5). Product structures can also represent the same item at different geographical locations (Figure 1.6). The demand interaction is included in the models by means of modifying the inventory balance equations. In periodic review models, the inventory balance equations for multi item processes can be represented as,

$$X_{jt} - I_{jt} + I_{j,t-1} = D_{jt}, \quad j = 1, 2, \dots, N; \quad t = 1, 2, \dots, T,$$

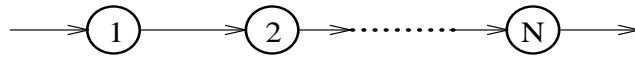
where the production, inventory, and demand variables are further subscripted by j referring to item j . D_{jt} is referred to as *gross demand* and defined as follows,

$$\begin{aligned} D_{jt} &= \bar{D}_{jt} + \hat{D}_{jt} \\ &= \bar{D}_{jt} + \sum_{i \in S(j)} \alpha_{ji} X_{i,t+\delta}, \end{aligned}$$

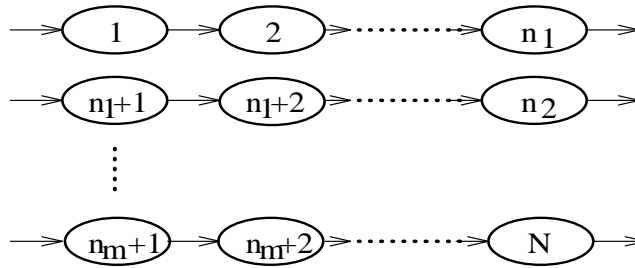
where,

⁴A different kind of demand interaction occurs when there exists *correlated demand* (when the demand of one item is correlated with another), as occurring in marketing phenomena of *substitution* or *reinforcement*.

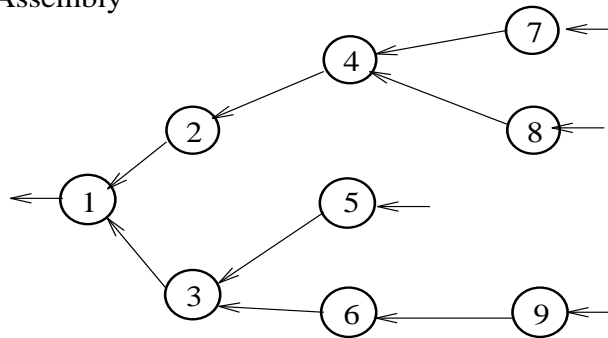
a) Series



b) Parallel



c) Assembly



d) General

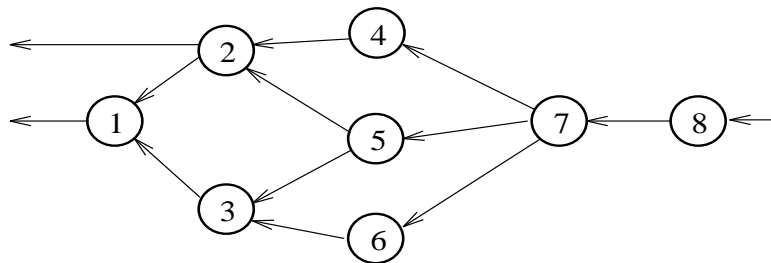


Figure 1.4 Product Structures

\bar{D}_{jt} the external (independent) demand for item j in period t ,

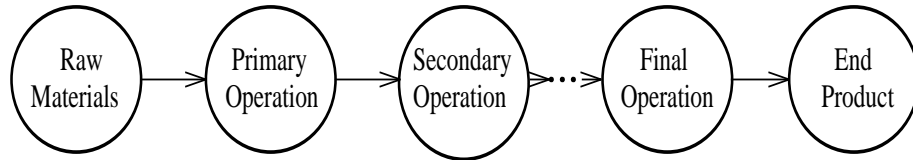


Figure 1.5 A Serial Manufacturing Process

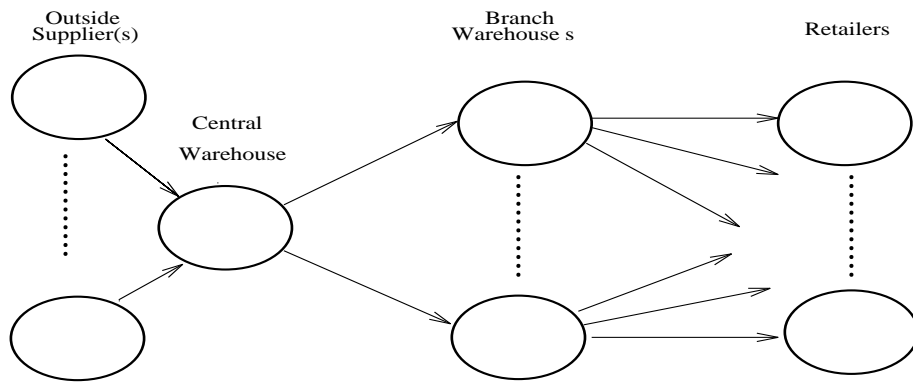


Figure 1.6 A Distribution System

\hat{D}_{jt} the internal (dependent) demand for item j in period t ,

α_{ji} the amount of item j required to produce a unit of item i ,

$S(j)$ the (index) set of items that require item j in their production (the *successor set* of item j),

δ the procurement lead time.

If $S(j) = \emptyset$, for all j , then demand interaction do not exist among items and model reduces to N single item models (provided that no other type of interaction exists among items).

Exercise 1.9 In the above formulation, suppose the procurement lead time, δ , refers to manufacturing lead time. Discuss why, in the presence of resource constraints, and unless there exists technological restrictions to the contrary, manufacturing lead time can always be equal to one period, i.e., $\delta = 1$.

1.6 CONCLUDING REMARKS

Based on their analytical structures, production-inventory models can be classified according to the following characteristics:

CONTINUOUS REVIEW	PERIODIC REVIEW
STOCHASTIC	DETERMINISTIC
MULTI ITEM	SINGLE ITEM
DYNAMIC	STATIC

It can be said that periodic review, deterministic, single item, static models are the easiest for analysis and computation. Whereas, continuous review, stochastic, multi item, dynamic models are the hardest. Generally, production planning problems are formulated as periodic review, deterministic, multi item, dynamic models.

The origins of mathematical inventory theory dates back to F. W. Harris, “How Many Parts to Make at Once”, **Factory, The Magazine of Management**, Vol.10, No. 2, February 1913, pp. 135-136,152 [49]. A historical and entertaining account of Harris’, so called, EOQ model can be found in [37, 38].

It is usually claimed that the cornerstones of modern inventory theory are the following articles:

- Arrow, K. A., T. E. Harris, and J. Marschak, “Optimal Inventory Policy”, **Econometrica**, Vol. 19, pp. 250–272, 1951,
- Dvoretzky, A., J. Kiefer, and J. Wolfowitz, “The Inventory Problem: I. Case of Known Distributions of Demand” pp. 187–222, and “II. Case of Unknown Distributions of Demand” pp.450–466, **Econometrica**, Vol. 20, 1952.

The seminal work of Hadley and Whitin [48] and the model oriented books of Johnson and Montgomery [57] and Hax and Candea [50] are useful references. Practical and heuristic approaches are emphasized in Silver and Peterson [75]. Numerous books on *production and operations management* treat production-inventory systems. For current state of the art in production-inventory systems see [69] and [47].

2

BASIC MODELS OF PRODUCTION PLANNING

In this chapter, a number of basic lotsizing models will be presented. As mentioned earlier, periodic review models are used in formulating production planning problems. Therefore, emphasis will be on periodic review, deterministic models. But it is always useful to associate these models with their continuous review and stochastic counterparts.

2.1 STATIC MODELS

The model we shall investigate in this section is possibly the simplest model of a production-inventory system. The basic assumptions are the following. Items do not have any interaction, it is a single item model. Furthermore, it is *deterministic*, that is, all parameters are assumed to be known constants. The parameters are *stationary*. All of the cost, demand, resource availability, and the technology related (e.g. processing times, input-output coefficients) parameters are assumed to be time independent. Furthermore, we assume that the system will operate indefinitely. Hence, we take the planning horizon to be *infinite*. Procurement decisions can be made at any period and there is no upper bound imposed on the amount procured. Shortages are not allowed. There is a fixed cost of procurement and linear unit variable procurement cost. Since lost sales are not allowed and parameters are stationary, there is no need to consider the variable procurement costs in the optimization: they are *sunk* costs.

Under these assumptions, the production-inventory process to be modeled has a *regenerative property*. Suppose the process starts with zero initial inventory level and suppose we make a specific set of procurement decisions. As a result of these decisions, inventory level will fluctuate. When the inventory level reaches zero level again, a new set of procurement decisions is needed. But, due to the stationarity assumption, none of the parameter values has changed since the last decision, the same set of procurement decisions must again be made. In other words, since nothing has changed, why change the decisions. Hence, the process repeats itself. The times at which inventory level

reaches zero from above are called the *regeneration points*. The portion of the process in between two generation points is called a *cycle*.

We cannot minimize the total costs over an infinite planning horizon¹. Hence, the appropriate objective function is to minimize costs per unit time. Since the process repeats itself identically in every cycle, it is sufficient to compute the costs for one cycle and divide this value by the cycle length in order to obtain the costs per unit time. Our approach, then, to infinite planning horizon models is to compute the costs per unit time as a function of the procurement decision variables and minimize this function with respect to these variables.

Let the parameters of the model be,

D Amount demanded per period (units per period),

k Fixed charge cost of procurement (\$ per procurement),

h Unit variable inventory holding (or, carrying) cost (\$ per unit per period),

and the (decision) variables² be,

X_t Amount to be procured in period t (units),

I_t Amount of inventory carried from period t into period $t + 1$, that is, the ending inventory in period t (units), and $I_0 \equiv 0$.

Exercise 2.1 *If the initial inventory, $I_0 > 0$, which model assumption is violated?*

Our immediate problem is to determine the procurement quantity at the first period, X_1 . Since no shortages are allowed, it must be that $X_1 > D$. Furthermore, it has to be a positive integer multiple of the demand per period, D , that is, $X_1 = nD, n = 1, 2, \dots$. To see this, suppose the contrary, say $X_1 = 1.5D$. Since we have a procurement in period one, we incur the fixed cost k , and since the ending inventory is $.5D$, we have an inventory carrying cost of $.5Dh$. But the beginning inventory is not sufficient to meet the demand in the second period. This necessitates another procurement at period 2, for an amount of at least $.5D$, resulting an additional fixed cost of procurement, k . So, the costs incurred in policy $\{X_1 = 1.5D, X_2 \geq .5D, \dots\}$ is $(2k + .5Dh + \dots)$. But we can do better by a policy $\{X_1 = D, X_2 \geq D, \dots\}$ with a cost $(2k + \dots)$ which is $(.5Dh)$ less than the previous policy. This specific example can easily be generalized and we have the following result:

¹To be precise, this is true only in the absence of *discounting*. Throughout this discussions, discounting of the future costs are not treated.

²Notice that, because of the inventory balance equations, it is sufficient to specify either inventory levels or the procurement amounts.

Result 2.1 *At any period, either there is a beginning inventory or procurement can take place, but not both.*

The problem reduces to finding a value for n , in other words determination of the cycle length. To this end, let us find the procurement and inventory related costs per period as a function of the cycle length, n . Total costs per period is the sum of the fixed procurement cost per period and the inventory holding cost per period. Recall that variable procurement costs can be ignored for the reasons discussed previously and the shortage costs do not exist due to the assumptions of this model.

Since we are placing one order in every n periods, the contribution of the fixed procurement cost to each period is k/n . Amount of ending inventory in each period of the cycle is,

$$I_t = (n - t) \cdot D, \quad t = 1, 2, \dots, n.$$

The total amount of inventory carried in n periods is,

$$\begin{aligned} \sum_{t=1}^n (n - t)D &= \sum_{t=1}^{n-1} (n - t) D \\ &= D \left(\sum_{t=1}^{n-1} n - \sum_{t=1}^{n-1} t \right) \\ &= D [n(n - 1) - (1/2)n(n - 1)] \\ &= (1/2)n(n - 1)D. \end{aligned}$$

Thus, the total costs per period as a function of n , the cycle length, can be stated as,

$$TC/P(n) = k/n + (1/2)hD(n - 1).$$

If we were to treat n as a continuous variable, it is easy to show that the above function is strictly convex and that the minimum occurs at $\hat{n} = \sqrt{(2k)/(hD)}$. But, this value for cycle length will most likely be fractional, thus, the optimal cycle length is given by,

$$n^* = \operatorname{argmin}\{TC/P(\lfloor \hat{n} \rfloor), TC/P(\lceil \hat{n} \rceil)\}$$

And the optimal lotsize is $X^* = D \cdot n^*$.

Exercise 2.2 (Harris' EOQ Model) *Consider a continuous review model with the same set of assumptions as the above model, except that demand is continuous at a constant rate d units per unit time, and that procurement is not restricted to specific time points but can take place at any time. Find the optimal lot size, explicitly defining and indicating the units of each parameter used in the model.*

Exercise 2.3 *Our company produces an item at a constant rate of 5,000 units per month. The production cost per unit is estimated to be \$ 1. These units are delivered to a manufacturer who is subcontracting our company for the manufacture of these units. The delivery of these units is done by single truck which is owned and operated by our company. The truck has a total shipping capacity of 2,500 units, and it costs \$ 25 to make a shipment, irrespective of the quantity shipped. Annual inventory carrying charges are estimated at .50 (\$ per \$ per year). Describe the inventory system graphically and determine how often the shipments be made to the manufacturer.*

Exercise 2.4 (Capacity Constraints) Consider the model in Exercise 2.2. Suppose there exists an upper bound on the amount of procurement, that is $X \leq u$, where u is the upper bound on the procurement level. Show that the constrained optimal lot size, $X_c^* = \min \{X^*, u\}$, where X^* is the unconstrained optimal.

Exercise 2.5 (Continuous Review with Shortages) Consider the model in Exercise 2.2. Suppose shortages are allowed in the form of backorders and assume that there are no lost sales. Let the shortage (backorder) cost be

p Unit shortage (penalty) cost (\$ per unit time per unit short).

Find the optimal lot size and the optimal backorder level.

Exercise 2.6 (Continuous Review with Lost Sales) Consider the model in Exercise 2.2. Suppose shortages are allowed, but none can be backlogged. All shortages are lost sales. Assume that a loss of \bar{p} (\$ per unit short) is experienced. Show that it is never optimal to inventory the item and permit lost sales.

Exercise 2.7 Consider the model in Exercise 2.5. Suppose there exist an additional backlogging cost defined as follows:

\hat{p} Shortage (penalty) cost per unit short, independent of the duration of shortage (\$ per unit short).

Derive the necessary equations to solve explicitly for the optimal lot size and the optimal backorder level.

Let $\alpha \equiv \sqrt{2kh/d}$. Assume $p = 0, \hat{p} > 0$, discuss in detail the resulting inventory systems for the following three cases:

1. $\alpha < \hat{p}$,
2. $\alpha = \hat{p}$,
3. $\alpha > \hat{p}$,

Exercise 2.8 (Continuous Review with Finite Procurement Rate) Consider the model in Exercise 2.5. Suppose in addition to planned shortages we allow for a finite procurement rate,

$x(t) \equiv x$ Number of units procured per unit time.

Find the optimal lot size and the optimal backorder level.

Exercise 2.9 (Finite Planning Horizon) *In infinite planning horizon models, the objective is to minimize total costs per unit time. To this end, total costs per cycle is computed and when this quantity is divided by the cycle length we obtain the total costs per unit time. Beginning points of each cycle constitutes the regeneration points of the process, in the sense that process repeats itself identically at every cycle. Optimization takes place based on the assumption that the process will continue indefinitely. But, if the planning horizon is finite, then this assumption is no longer valid. The optimal values of the decision variables will be the same only if the finite planning horizon is an exact integer multiple of the optimal cycle length computed under the infinite planning horizon assumption—which is highly unlikely to occur. Suggest a scheme that approximates the finite horizon optimal values based on the infinite horizon solutions. How can you measure the performance of this approximation scheme?*

2.2 DYNAMIC MODELS

Production planning requires decision making in order to adapt to the evolving conditions. Inherently, therefore, production planning models must admit nonstationary parameters. In this section, single item models will be discussed. Multi item models will be treated in the following two sections. Single item models are applicable in aggregate production planning problems where there is *full aggregation* of items into a single item. Then there is full aggregation, the single item naturally is *the production rate*.

2.2.1 Linear Cost Models

As it was shown in Example 1.3, when the objective function is a separable piecewise linear convex function, the problem can still be formulated as a linear program, provided that all the constraints are linear and variables can be assumed to be continuous. This provides, in addition to efficient computation, extensive analysis capability by post-optimality procedures. Fortunately, it is not very unrealistic to assume piecewise linear convexity in considering most cost functions in aggregate production planning.

Transportation Formulation

Consider a single item, dynamic, periodic review model. In each period, $t = 1, 2, \dots, T$, the (single) item can be procured from S sources. These sources can be different

suppliers, or modes of production, such as regular time, overtime, and subcontracting from a number of subcontractors. It is natural that these sources have finite capacities. Let the capacity of source $s = 1, 2, \dots, S$ be R_{st} during period $t = 1, 2, \dots, T$, and other parameters and variables are defined by:

D_t Amount demanded in period $t = 1, 2, \dots, T$,

c_{st} Unit cost of procuring from source $s = 1, 2, \dots, S$, in period $t = 1, 2, \dots, T$,

h_t Cost of storing one unit from period t to period $t + 1$,

X_{st} Amount to be procured from source $s = 1, 2, \dots, S$, in period $t = 1, 2, \dots, T$,

I_t Amount of inventory at the end of period $t = 1, 2, \dots, T$.

Then the problem can be formulated as a linear program:

$$\min_{X, I \geq 0} \sum_{s,t} [c_{st}X_{st} + h_t I_t]$$

subject to:

$$\begin{aligned} X_{st} &\leq R_{st}, & s = 1, \dots, S; & t = 1, \dots, T, \\ I_{t-1} + \sum_s X_{st} - I_t &= D_t, & & t = 1, \dots, T, \end{aligned}$$

Exercise 2.10 (Backorders) *In the above formulation, suppose backorders are allowed. That is, let $I_t, t = 1, 2, \dots, T$ be an unrestricted variable. Its objective function coefficient is h_t if it is nonnegative and p_t if it is negative, where p_t is the cost of having one unit on backorder at the end of period t . By defining appropriate nonnegative variables, revise the above linear programming formulation to account for the backorders.*

Exercise 2.11 (Transportation Problem) *Consider the linear programming formulation of problem with backorders. Let*

Y_{stu} *be the amount procured from source $s = 1, \dots, S$ in period $t = 1, \dots, T$ in order to meet the demand in period $u = 1, \dots, T$,*

g_{stu} *be the unit cost of procuring from source $s = 1, \dots, S$ in period $t = 1, \dots, T$ in order to meet the demand in period $u = 1, \dots, T$.*

Express g_{stu} in terms of the original cost parameters and formulate the problem as a transportation problem of linear programming. Show the corresponding transportation network.

Renewable Resources

When resource constraints were discussed earlier, it was implied that resources were nonrenewable, that is, resources can be used until they are exhausted and they cannot be renewed during the planning horizon. In the context of periodic review models, though, it is necessary to view resources in two categories:

Nonrenewable Resources Resources that cannot be renewed during the planning horizon, and

Renewable Resources Resources that are renewed at each period.

One way to explain the difference is the following. Suppose you can at most allocate ten tons of steel for a particular production during the next four weeks. Your planning horizon consists of 4 one-week periods. Your weekly production capacity is known, say 40 hours. At the beginning of each period you have a “renewed” 40 hours of production capacity resource, but have only 10 tons of steel, raw material resource, that can be used during the entire planning horizon of four period. In other words, you cannot use 80 hours of production capacity in the first period and none in the second, whereas, you can use all the steel in the first period and none during the proceeding periods.

Note that the *nonrenewable resources* can be stored in inventory, while the *renewable resources* cannot be inventoried. It may be helpful to think about the goods and services analogy. Goods are like nonrenewable resources which can be stored. On the other hand, renewable resources are like services that cannot be kept in inventory. It will be convenient to treat nonrenewable resources as another item that has a *demand interaction* with the item that requires that resource in its production.

Examples of renewable resources are employment levels and production capacities. From one period to next, it may be possible to change these, at a cost. In the production planning terminology, these costs are referred to as hiring/firing and production capacity change costs, respectively.

Exercise 2.12 *Discuss the intuitive justification of hiring/firing and production capacity change costs being convex (usually, piecewise linear) in most environments.*

It will be convenient to incorporate these level and capacity changes in the formulation by means of balance equations. Similar to inventory balance equations, these resource balance equations simply state that in each period, the resource level is equal to the resource level during the previous period plus the increase in resource level in that period minus the decrease in resource level in that period.

Exercise 2.13 *Show that in the (optimal) solution of a linear program that models resource level changes, it is not possible to have both an increase and a decrease in a resource level.*

Example 2.1 (Production Capacity Change) *Let*

R_t *be the production capacity in period* t ,

W_t *be the increase in production capacity from period* $t - 1$ *to period* t ,

Z_t *be the decrease in production capacity from period* $t - 1$ *to period* t ,

w_t *be the cost to increase in production capacity by one unit from period* $t - 1$ *to period* t ,

z_t *be the cost to decrease in production capacity by one unit from period* $t - 1$ *to period* t .

Then the production capacity balance equation is

$$R_{t-1} + W_t - Z_t - R_t = 0, \quad t = 1, 2, \dots, T,$$

where R_0 *is the given initial production capacity. The objective function needs to be appended by the appropriate cost terms.*

Exercise 2.14 *Give a complete formulation for a production planning problem with cost of production capacity change.*

Exercise 2.15 *Particularly in process industries, it is common to face with a situation where there are costs associated with production rate changes. These should not be confused with production capacity changes. For example, if a process is being operated at a rate of X_t in period t , unless the $X_t = X_{t+1}$, then we incur production rate changeover costs. By defining appropriate variables, formulate this problem.*

2.2.2 Convex Cost Models

For all practical purposes, convex cost functions can be approximated, for any desired level of accuracy, by piecewise linear convex functions. The resulting models can be solved as linear programs. Original work on convex cost models was done by Veinott[78]; details can be found in [80, 50, 57]. Importance of detailed study of convex cost models lies in their contribution to the *planning horizon analysis*³. Very simply stated, the

³The planning horizon analysis is also an equally challenging problem in concave and linear cost functions.

planning horizon analysis deals with the issue of determining the shortest horizon to consider such that the current decision remains optimal. These results constitute the core of *rolling horizon* concept. In order to appreciate the importance of convexity, consider the following exercise.

Exercise 2.16 *Suppose in Exercise 2.11, backlogging is not allowed. Develop a “one-pass procedure” for the optimal solution.*

Exercise 2.17 (Where-or-When Production Problem, [80]) *This problem has various interpretations. Suppose a company has N plants and must manufacture a total of D units in a given period. Let X_i be the amount produced in plant $i = 1, \dots, N$. Furthermore, assume that the cost of production in plant i is given by*

$$C_i(X_i) = (1/c_i) \cdot X_i^2, \quad i = 1, \dots, N,$$

where $c_i > 0$ is estimated from historical accounting information. Then the problem is

$$\min_{X \geq 0} \sum_{i=1}^N C_i(X_i)$$

subject to:

$$\sum_{i=1}^N X_i = D.$$

Another interpretation is in terms of choosing optimal production levels at a single plant over a number of periods with the requirement that the total production over the planning period is D . Note that only production costs are considered and the inventory holding costs are assumed to be negligible.

1. Use a dynamic programming formulation to show that given the values X_i , $i = 1, \dots, n-1$, the optimal value for $X_n = c_n(D - \sum_{i=1}^{n-1} X_i) / (\sum_{i=n}^N c_i)$.
2. Find the same optimality condition by applying the Karush-Kuhn-Tucker Optimality Conditions.

2.2.3 Concave Cost Models

The problem is a single item, periodic review model with nonstationary parameters but no bounds on variables other than the nonnegativity restrictions. Assume that the procurement and inventory costs are concave. The arguments will equally apply when backlogging is permitted, but for ease of presentation let us assume that the shortages

are not allowed. Then the model can be formulated as the following mathematical program:

$$\left\{ \min_{X, I \geq 0} \left[\sum_t C_t(X_t) + H_t(I_t) \right] \mid I_{t-1} + X_t - I_t = D_t, t = 1, \dots, T, \right\}$$

where, the $C_t(\cdot)$ and $H_t(\cdot)$ are concave functions of procurement and inventory amounts in period t , respectively.

Consider the following results,

Result 2.2 *Minimum of a concave function subject to linear constraints will occur at one of the extreme points of the feasible region defined by the linear constraints.*

Result 2.3 *Every basic feasible solution to a set of linear constraints corresponds to an extreme point of the feasible region defined by those constraints.*

Therefore, it is sufficient to search for the optimal solution only among the basic feasible solutions to the system of equations

$$\{I_{t-1} + X_t - I_t = D_t, t = 1, \dots, T\}.$$

Any basic feasible solution to this set of linear equations can have at most T positive variables from the set

$$\{X_t, I_t, t = 1, \dots, T\}.$$

Based on these, the following result follows:

Result 2.4 *In a basic feasible solution to the above problem, there is either inventory carried into a period or procurement takes place at that period, but not both. That is,*

$$I_{t-1} \cdot X_t = 0, t = 1, \dots, T.$$

For example, in a four period problem, let $X_1, X_2, I_1 > 0$. Since there can be at most four positive variables in a basic feasible solution, there is no way of satisfying the demands of periods 3 and 4, D_3, D_4 , by choosing only one more positive variable out of I_2, I_3, X_4 . In other words, if a period “uses” two positive variables (i.e. one more than it deserves), then there must be a period into which there is no inventory being carried and no procurement taking place—thus making it impossible to meet its demand.

An immediate consequence of this result is that the procurement occurs only when inventory reaches zero (as is the case in its continuous review counterpart). These

points are referred to as *regeneration points* and periods in between two regeneration points constitute a cycle. This is similar to the static case, except that, since parameters are nonstationary, the cycle lengths are not necessarily the same.

An *efficient* dynamic programming formulation of this problem is based on the following result:

Result 2.5 *If there is a procurement to take place in any period, $t = 1, \dots, T$, it is equal to $\{D_t\}$ or $\{D_t + D_{t+1}\}$ or ... or $\{D_t + D_{t+1} + \dots + D_T\}$.*

Wagner-Whitin's[81] original dynamic programming formulation assumes fixed charge procurement and linear inventory holding costs with no backlogging, though numerous extensions and generalizations of the approach has been made since then (see [47] for an overview). The significance of Wagner-Whitin's work lies as much in developing a planning horizon theorem as providing an efficient solution to the above problem⁴.

The functional equation for periods, $t = 1, \dots, T$, can be defined as follows:

$$\begin{aligned} f_t(I_{t-1}) &= \text{the minimum cost of satisfying demand from period } t \text{ to } T, \\ &\quad \text{given the initial inventory, } I_{t-1}. \\ &= \min_{X_t} \{C_t(X_t) + H_t(I_t) + f_{t+1}(I_t)\} \\ &= \min_{X_t} \{C_t(X_t) + H_t(I_{t-1} + X_t - D_t) + f_{t+1}(I_{t-1} + X_t - D_t)\} \end{aligned}$$

where I_0 and I_T are given and $f_{T+1} \equiv 0$.

Exercise 2.18 *Suppose there are capacity constraints on procurement. Is the above dynamic programming formulation still applicable? Why or why not?*

Exercise 2.19 *Suppose you are faced with the demands $D = \{5, 7, 11, 3\}$ for the next four periods. You have to choose two alternative ways of meeting this demand: in-house production or outside purchase (once you decide on an alternative, all four periods' demand should be met that way). Shortages are not allowed and holding cost is \$1.00 per unit per period for both alternatives.*

In-house production data:

⁴In the words of [77] "The Wagner-Whitin Legacy ... started a small industry."

Period, i	Setup Cost, $\$k_i$ per setup	Production Cost, $\$c_i$ per unit
1	5	1
2	7	1
3	9	2
4	7	2

Outside Purchase data (The order cost is zero):

Range, in number of units ordered	Unit Cost $\$c_i$ in period			
	1	2	3	4
1 to 3	1	2	2	3
4 to 11	1	4	5	4

Which alternative would you choose and why?

These dynamic lot sizing models discussed in this section often occur as subproblems of Material Requirements Planning (MRP) Systems. See [7, 19] for an overview of these issues.

2.3 MULTI ITEM MODELS WITH DEMAND INTERACTION

The concepts of an *item* and *product structures* were introduced in Section 1.5.3. The models to be discussed in this section arise due to these product structures. In the MRP literature, they are commonly known as *infinite loading* models. No resource constraints are included (unless one treats a nonrenewable resource as another item) in these models. Since these models primarily model production environments with input/output relationship among items, other than the end items, it does not make sense to permit shortages. For example, an assembly is not physically complete without an item it must contain. Shortage of a single unit may shut down the entire production line.

Basically, these models minimize the sum of production and inventory holding costs subject to inventory balance equations where demands per period reflect both internal and external demands.

The general form of the model is given below:

$$\min_{X, I \geq 0} \left[\sum_{j,t} C_{jt}(X_{jt}) + H_{jt}(I_{jt}) \right]$$

subject to:

$$X_{jt} - I_{jt} + I_{j,t-1} - \sum_{i \in S(j)} \alpha_{ji} X_{i,t+\delta_j} = \bar{D}_{jt}, \quad j = 1, 2, \dots, N; \quad t = 1, 2, \dots, T.$$

The definitions of variables are repeated for convenience:

X_{jt} amount of item $j = 1, \dots, n$ to be produced in period $t = 1, \dots, T$,

I_{jt} amount of inventory of item $j = 1, \dots, n$ at the end of period $t = 1, \dots, T$,

\bar{D}_{jt} the external (independent) demand for item j in period t ,

α_{ji} the amount of item j required to produce a unit of item i ,

$S(j)$ the (index) set of items that require item j in their production (the *successor set* of item j),

δ_j the production lead time of item $j = 1, \dots, n$.

Even when the cost functions are assumed to be linear, dimension of such models make their routine application either infeasible or very expensive with conventional LP software [53]. For example, for a ten-period problem, with 100 products, each having 100 predecessor parts, the resulting linear program will have on the order of 100,000 constraints and 200,000 variables. It is computationally prohibitive to consider the fixed charge production costs, as they are required in MRP environments. For problems of practical size, use of heuristic multilevel lotsizing procedures are unavoidable. The reader is referred to [7] for an overview of heuristics.

Exercise 2.20 *Because of the production lead times, certain variables defined above may be eliminated from the model. Explicitly define those variables to be set to zero, for specific product structures.*

Exercise 2.21 ([19]) *In systems like infinite-loading MRP, the production lead times are used as a means of handling the finite production capacities. But lead times and lot-sizes are intimately related. Lead times are a consequence a particular lotsizing decision and cannot be used a priori to generate accurate lotsizes. Discuss.*

Exercise 2.22 *Show how a nonrenewable resources can be treated as an item in a demand interaction model. Assume there are no other resource constraints. Give a detailed formulation for a small example.*

2.4 MULTI ITEM MODELS WITH RESOURCE INTERACTION

A single period case of a multi item model with resource constraints was discussed in Example 1.1. When there are multi periods, clearly, one must incorporate the inventory balance equations. If we assume that there does not exist any demand interaction, i.e. $S(j) = \emptyset, \forall j$, these would be similar in structure to single item balance equations where variables are appended with subscripts identifying the item they refer to.

The general form of the model is given below:

$$\min_{X, I \geq 0} \left[\sum_{j,t} C_{jt}(X_{jt}) + H_{jt}(I_{jt}) \right]$$

subject to the inventory balance equations:

$$X_{jt} - I_{jt} + I_{j,t-1} = D_{jt}, \quad j = 1, 2, \dots, N; \quad t = 1, 2, \dots, T,$$

and, the resource constraints,

$$\sum_{j=1}^N r_{kjt} X_{jt} \leq R_{kt}, \quad k = 1, 2, \dots, K; \quad t = 1, 2, \dots, T.$$

where, the definitions are repeated for convenience,

X_{jt} amount of item $j = 1, \dots, n$ to be produced in period $t = 1, \dots, T$,

I_{jt} amount of inventory of item $j = 1, \dots, n$ at the end of period $t = 1, \dots, T$,

D_{jt} the demand for item $j = 1, \dots, n$ in period $t = 1, \dots, T$,

r_{kjt} amount of resource k required in the unit production of item j in period t ,

R_{kt} amount of resource $k = 1, 2, \dots, K$ available in period $t = 1, 2, \dots, T$.

Exercise 2.23 (Multi Period Product Mix Problem) *Formulate a profit maximization version of the above problem. Define the necessary parameters.*

2.5 CONCLUDING REMARKS

The key elements of a production system are customers, material, and capacity. Such a system requires, in the least, the following objectives:

- maximize customer service,
- minimize inventory investment, and
- maximize utilization of capacity for efficient plant operation.

An effective production planning and control system aims to resolve these conflicting (“incompatible”) objectives. In order to achieve this, one can use *multi criteria optimization*. For example, simultaneously taking into account the following objective functions:

- minimize shortages,
- minimize inventory holding, and
- maximize machine utilization.

The dimensions and complexities of most production planning problems make their analysis by means of multi criteria optimization either impossible or computationally prohibitive. Practical considerations make it necessary to utilize *satisficing*⁵ approach. Ackoff [1] describes the satisficing procedure as follows:

The most important objective (judged in some qualitative way) is selected as the basis for the measure of performance of a course of action. Minimal levels of acceptable performance (subjectively determined) are imposed as restrictions on an acceptable solution. Such a level may, for example, be the level of performance attained in the past.

Consider a case in which a decision maker wants to reduce costs and improve service. [...] The decision maker may be willing to settle for this criterion of performance: to minimize costs subject to the condition that the quality of service does not deteriorate below the current level. If service is improved, fine; this is a bonus. But it will not be permitted to get worse.

Such a procedure combines the principles of optimization and satisficing; optimization with respect to costs, satisficing with respect to service. If alternative optima relative to costs are found, that action which does best relative to service is selected.

⁵*satisfying + sufficing*

Exercise 2.24 According to [52], the eminent management scientist and Nobel Laureate in Economics, Herbert Simon, has introduced the concept that the goal of an operations research study should be to “satisfice” rather than optimize. Discuss.

This approach results in the classical formulation of production planning problems: *minimization of production and inventory related costs subject to satisfying the demand without violating the resource constraints*. Considering a planning horizon of, probably, years and each eight-hour shift as a period, with thousands, if not ten thousands of interacting items, such a formulation turns out to be a large scale mathematical program. The problem with this *monolithic* approach, in addition to prohibitive computational burden, is massive detailed data requirements.

In order to resolve this dilemma, hierarchical approaches based on aggregation are suggested (see, for example, [64]). Basically, production planning decision hierarchy is designed to be composed of echelons such as

- plant sizing and location,
- equipment type and amount,
- production allocation among plants,
- plant capacity planning,
- item production schedules.

Each higher level in the hierarchy imposes constraints on the lower level and each lower level sends feedback information to the higher level. One major problem is how to treat the interactions among these hierarchical levels. On one extreme, this can be as loose as in Manufacturing Resources Planning (MRPII) or, on the other extreme, as mathematically structured as in the price- or resource-directive decomposition approaches of mathematical programming. Another major problem is how to treat the aggregation and disaggregation issues. Levels of aggregation is required for both items and production resources. Although there is considerable amount of ongoing research in this area, the theory is far from complete [77].

Exercise 2.25 Consider a multi item, periodic review model with resource and demand interaction. Suppose periods are of different lengths. For example, the planning horizon consists of ten periods of length one-day, followed by five periods of length one-week, and so on. Suggest a way of handling inventory balance equations.

Exercise 2.26 ([57]) The company you work for produces a number of products. You are responsible for the planning the production of only two of these products: let us say Product A and Product B. Each of these two products require two operations. For

both products, the first operation is done in Department 1 and the second operation in Department 2. You are pretty sure that the production rates are:

Production Rate in Units per Hour		
Product	Operation 1	Operation 2
A	2	4
B	1	5

It is a terribly competitive business you are in: no matter how much you force the Sales Department, the best you can get out of them is their guesstimate of the demand for the next three months. They say, to the best of their abilities, the demand for A will 100, 200, and 150 units, and for B, 180, 220, and 100 units. But they threaten you with your job if you allow for shortages!

Then you fight with Production Department for capacity allotments. After hours of negotiations all you can get from the production people is the following:

Capacity in Machine Hours				
Month	Department 1		Department 2	
	Regular Time	Overtime	Regular Time	Overtime
1	250	50	100	20
2	250	50	50	20
3	200	50	50	20

But you have a feeling that, with a little help from your boss, you can make small changes in these allotments...

Your company operates in a free zone, the labor rates are highly volatile, the same can be said about inventory carrying charges. But the cost people did their best and come out with the following cost data. You are left with the problem of distributing the unit cost figure over operations one and two.

Product	Unit Cost(\$)		Unit Inventory Cost per Period (\$)
	Regular Time	Overtime	
A	30	31	2
B	28	32	3

Last time you checked the finished goods inventory, you noticed that there were 10 units of Product A and 20 units of Product B.

You are up for promotion and you really want to show off your expertise in modeling and solving production planning problems, making extensive use of the powers of post

optimality analysis of linear programming. Your company has all the necessary software and support personnel and 24-hour open computing facilities. Although you have a very important engagement for the coming weekend, you have to submit the report in a week!

3

PRODUCTION SCHEDULING

3.1 INTRODUCTION

The domain of scheduling is material flow systems. “Material” in these systems may be

- items to be manufactured (as in production scheduling),
- tasks to be processed in a computing systems (as in computer scheduling),
- activities to be completed in order to terminate a project (as in project scheduling),
- services to be rendered (as in crew scheduling or classroom scheduling)

Our primary concern is production scheduling and, therefore, manufacturing terminology and examples will be used. It is common to define scheduling as “the art of assigning production resources to production activities in order to insure the termination of activities in a reasonable amount of time”. The aim of scheduling theory is to devise solution procedures for allocation over time of the production resources in the form *machines* to production activities referred to as *jobs*, subject to constraints on machines and jobs. The major constraints are limits on the capacity of resources and technological restrictions on the order in which jobs can be performed. Usually, these constraints are too complicated to be stated explicitly in lot sizing models. Therefore, in analytical models of production planning surrogate or aggregate constraints are used, which are simplified versions of true resource constraints. The exact capabilities of production resources can only be determined as the result of production scheduling process.

Typically, the number of feasible allocations or schedules will be finite, but very large. If all the relevant information on jobs, machines, and optimality criterion are known in advance, the scheduling problem becomes a combinatorial optimization problem. Most of these problems are notoriously difficult problems.

Sequencing simply is the ordering of a collection of jobs to be performed, whereas scheduling is concerned with the assignment of times for each job that tells when it is to be performed. If the criterion of a scheduling problem is a *regular measure of performance*, that is, when the criterion cannot be improved by inserting idle times on the machines, and thus, jobs can be started on machines as soon as possible, then a sequence determines a schedule, and vice versa.

In deterministic scheduling, all job characteristics (number of jobs, routing of jobs, etc.) and all shop characteristics (number of machines, their availabilities, etc.) are fixed and known in advance. In stochastic scheduling those characteristics vary randomly. Based on the assumption on the availability of information on the future requirements, the scheduling environment can be static or dynamic. In a *static environment*, the scheduling problem is defined with respect to a finite set of fully specified requirements; no additional requirements will be added to this set, nor will any of the specifications be altered. In *dynamic environment*, the scheduling problem is defined not only for the known requirements, but also with respect to the anticipations for additional requirements and specifications generated over the planning horizon.

A good schedule is the one which utilizes the resources efficiently, responds to demands rapidly, and conforms to prescribed due dates closely. Ideally, as any production plan, the schedules should be evaluated based on their costs, such as,

- fixed costs for setups and changeovers,
- variable production and overtime costs,
- inventory holding costs,
- shortage costs for not meeting the due dates and for stocking out,
- expediting costs for implementing the schedule in a dynamic environment,
- system costs for generating the schedule and monitoring the progress of the schedule.

But, in a detailed production scheduling environment, it is difficult, if not impossible, to assign such costs to each schedule alternative. Hence, it is customary to evaluate schedules based of their performances on job and shop related variables. Examples of schedule performance measures are

- utilization level of production resources,
- percentage of late jobs,
- the average or maximum tardiness of jobs,
- the average or maximum time jobs spend in the shop.

3.2 SEQUENCING AND SCHEDULING PROBLEMS

Deterministic, static scheduling problems may arise whenever n jobs J_j ($j = 1, \dots, n$) have to be processed on s stages S_ℓ ($\ell = 1, \dots, s$) each of which may have m_ℓ parallel machines $M_{k\ell}$ ($k = 1, \dots, m_\ell$). We assume that

- machines of a stage are not shared with any other stage i.e., if \mathcal{M}_ℓ denotes the set of machines at stage S_ℓ (where $|\mathcal{M}_\ell| = m_\ell$) then $\mathcal{M}_\ell \cap \mathcal{M}_h = \emptyset \quad \forall \ell \neq h$,
- each machine $M_{k\ell}$ of a stage S_ℓ can process at most one job J_j at a time and,
- unless otherwise stated each job J_j can be processed on at most one machine $M_{k\ell} \in \mathcal{M}_\ell$ and stage S_ℓ at a time.
- following data can be specified for each job J_j :
 - a number of operations o_j ,
 - a sequence of operations $\{O_{1j}, \dots, O_{o_j,j}\}$, where O_{ij} has to be processed on one of $m_{s_{ij}}$ parallel machines of a stage s_{ij} with $s_{(i-1)j} \neq s_{ij} \quad \forall i = 2, \dots, o_j$,
 - a processing requirement p_{kij} of each O_{ij} on k -th ($k = 1, \dots, m_{s_{ij}}$) machine of s_{ij} ,
 - a *ready time* or *release date* r_j on which J_j becomes available for processing,
 - a *due date* d_j by which J_j should ideally be completed,
 - a *deadline* \bar{d}_j by which J_j must be completed,
 - a *weight* w_j indicating the relative importance of J_j ,
 - a nondecreasing real function f_j of the completion time C_j , indicating the cost $f_j(C_j)$ incurred if J_j is completed at C_j .

Given such an instance, a scheduling problem can be modeled as determining the schedule \mathcal{S} that minimizes f_{max} or $\sum f_j$ such that in \mathcal{S}

1. $f_{max} = \max_{1 \leq j \leq n} \{f_j(C_j)\}$ and $\sum f_j = \sum f_j(C_j)$,
2. prescribed $\{O_{1j}, \dots, O_{o_j,j}\}$ for each job J_j is preserved,
3. each of $m_{s_{ij}}$ parallel machines in stage s_{ij} processes one operation O_{ij} at a time,
4. each operation O_{ij} requiring the stage s_{ij} are processed on one and only one of $m_{s_{ij}}$ parallel machines at a time,
5. possibly some other constraints on each job and/or shop are satisfied.

The large variety of sequencing and scheduling problems makes it necessary to introduce a systematic notation that classifies these problems. The small differences in the problem assumptions makes significant difference in the solvability of these problems. Therefore, a detailed specification of each problem essential. The scheme that is presented below is based on [61], which was first suggested in [45] and also appears in [13, 12]. The version given here is adapted from [4]. In this classification, each scheduling problem is represented by a 4-tuple $\alpha \mid \beta \mid \gamma \mid \delta$:

- α identifies the production environment:
 - $\alpha = 1$: a single stage problem.
 - $\alpha = Fs$: a *flow shop* problem in which $o_j \leftarrow s$ and $s_{ij} \leftarrow s_i \quad \forall J_j$ and O_{ij} . If s is not given the general class of flow shop scheduling problems will be represented.
 - $\alpha = Js$: a *job shop* problem which is the general case defined at the beginning of this section.
 - $\alpha = Os$: an *open shop* problem which is same as the flow shop problem except in this case the order of operations is immaterial, i.e. $\{O_{1j}, \dots, O_{o_j, j}\}$ represents a set of operations but not necessarily their sequence.
- β identifies the machine environment at each stage of production. If we let \circ denote the *empty symbol* then the possible configurations are:
 - $\beta = \circ$: the problem with single machine at each stage of production.
 - $\beta = 1$: single machine at a particular stage of production; $p_{1ij} \leftarrow p_{ij}$.
 - $\beta = Pm_\ell$: Identical m_ℓ parallel machines at stage S_ℓ ; $p_{kij} \leftarrow p_{ij} \quad \forall M_{k\ell} \in \mathcal{M}_\ell$. If m_ℓ is not specified then the general class of problems in which there is an arbitrary number (m_ℓ) of parallel machines at stage S_ℓ , is implied.
 - $\beta = Qm_\ell$: Uniform m_ℓ parallel machines at stage S_ℓ ; $p_{kij} \leftarrow p_{ij}/t_{k\ell}$ for a given speed $t_{k\ell}$ of machine $M_{k\ell} \in \mathcal{M}_\ell$.
 - $\beta = Rm_\ell$: Unrelated m_ℓ parallel machines at stage S_ℓ .
- γ identifies further assumptions of the scheduling problem such as:
 - $\gamma = pmtn$: job preemption is allowed, i.e. the processing of any operation may be interrupted and resumed at a later time.
 - $\gamma = prec$: A precedence relation (\rightarrow) between the jobs is specified. It is derived from an acyclic directed graph G with the vertex set $\{1, \dots, n\}$. If G contains a directed path from j to k we write $J_j \rightarrow J_k$ and require that J_j is completed before J_k can start.
 - $\gamma = r_j$: nonzero ready times, that may differ for each job, j , are specified.
 - $\gamma = o_j \leq o$: constant upper bound on number of operations for all J_j is specified (valid only if $\alpha = Js$).

- δ identifies the optimality criterion of the scheduling problem. Commonly used performance measures are:

$f_j(C_j)$	f_{max}	$\sum f_j$
C_j	C_{max} (makespan)	$\sum w_j C_j$ (mean weighted flow time)
$C_j - d_j$	L_{max} (maximum lateness)	—
$\max\{0, C_j - d_j\}$	—	$\sum w_j T_j$ (mean weighted tardiness)
$U_j = \begin{cases} 0 & \text{If } C_j \leq d_j, \\ 1 & \text{otherwise.} \end{cases}$	—	$\sum w_j U_j$ (mean weighted number of tardy jobs)

These performance measures are called *regular* in the sense that each δ is a monotone function of the completion times C_1, C_2, \dots, C_n . That is $C_j \leq C'_j \ \forall \ j \Rightarrow \delta(C_1, C_2, \dots, C_n) \leq \delta(C'_1, C'_2, \dots, C'_n)$.

3.2.1 Examples

$1|P||C_{max}$: refers to a class of single stage scheduling problems in which n jobs are scheduled on m identical parallel machines so as to minimize makespan.

$1|Pc||C_{max}$: refers to a class of $1|P||C_{max}$ problems in which the number of machines is a constant c .

$1|Q||C_{max}$: refers to a class of single stage scheduling problems in which n jobs are scheduled on m uniform parallel machines so as to minimize makespan.

$1|P|r_j, \overline{d_j}||C_{max}$: refers to a class of single stage scheduling problems in which n jobs are to be scheduled on m identical parallel machines so as to minimize makespan. In a feasible schedule no job can start before its ready time r_j and each job must be completed by its deadline $\overline{d_j}$.

$J||C_{max}$: refers to a class of job shop scheduling problems in which the aim is to minimize makespan. It is assumed that in the job shop there is a single machine at each stage.

3.3 SCHEDULING ALGORITHMS AND COMPLEXITY

Given a scheduling problem, we need an algorithm to solve it. How do we know that the algorithm is a “good” one? A useful measure of performance would be “the rate of growth of the time or space required to solve larger and larger instances of a problem”

[2]. Therefore, we need a number to measure the size of a problem. One convenient measure is the size of the input data. For example, the size of a scheduling problem might be the number of jobs to sequence, n . More precisely, it is the number of digits in the data of the problem instance when it is encoded in binary form. The *time complexity* or the *running time* of the algorithm is the time needed by the algorithm (e.g., number of elementary operations such as additions and comparisons) expressed as a function of the problem size. The *space complexity* is similarly defined.

If a scheduling algorithm solves, in the worst case, a problem with an input size of n in time cn^2 for some constant c that is independent of the size and the data of the problem instance, then we say the time complexity of (or, the computational effort required by) the algorithm is $O(n^2)$, read “order n^2 ”. In general, an algorithm may have the complexity of $O(p(n))$. If $p(\cdot)$ is a fixed polynomial function in the size of the problem, then that algorithm is said to be *polynomially bounded* or, in the words of Edmonds [34], it is a “good” algorithm and the problem is “well solved”.

In order to appreciate the importance of polynomial boundedness, consider the following example adapted from [2]. Suppose we have a single stage scheduling problem with two identical parallel machines with the objective of minimizing schedule length, i.e., $1|P2||C_{max}$. Consider two algorithms: \mathcal{A}_1 and \mathcal{A}_2 . The first one, \mathcal{A}_1 , lists the n processing times in the nonincreasing order and assigns the jobs according to this list to the machine with the least amount of work, breaking the ties arbitrarily. It may not find the optimal solution, but it is a good heuristic algorithm with a time complexity of $O(n \log n)$. \mathcal{A}_2 , on the other hand, explicitly enumerates all possible solutions, with a time complexity of $O(2^n)$. Assume that one unit of time on our computer is equal to one millisecond. The limits on the problem size as determined by the growth rate are shown below:

<i>Algorithm</i>	<i>Time Complexity</i>	<i>Maximum Problem Size</i>		
		1 sec	1 min	1 hour
\mathcal{A}_1	$n \log n$	140	4893	2.0×10^5
\mathcal{A}_2	2^n	9	15	21

Suppose that we replace our computer with another one, which is ten times faster than the current one. The increase in the size of the problem that we can solve due to tenfold speed-up is shown below:

<i>Algorithm</i>	<i>Time Complexity</i>	<i>Maximum Problem Size Before Speed-up</i>	<i>Maximum Problem Size After Speed-up</i>
\mathcal{A}_1	$n \log n$	s_1	$\approx 10s_1$
\mathcal{A}_2	2^n	s_2	$s_2 + 3.3$

Note that with algorithm \mathcal{A}_2 , the size of the problem that can be solved after tenfold speed-up only increases by three, whereas with algorithm \mathcal{A}_1 , the size increases almost tenfold. Instead of an increase in speed, consider the effect of using more efficient algorithm. Suppose you have only one minute to solve the problem. Using the faster algorithm, you can solve a problem more than 325 times larger than the one you can solve with the slower algorithm.

At present, roughly 10% of the scheduling problems are “well solved”. It is also known that for 80% of the problems it is highly unlikely that polynomially bounded algorithms can be developed. The remaining 10% are open. This means that if we require optimal solutions, for roughly 90% of the problem types in scheduling, we have to resort to enumerative methods, such as branch & bound or dynamic programming, which may require exponential time in the worst case. If, in practice, this turns out to be computationally prohibitive, then one has to use a *heuristic* algorithm.

In order to be more concrete about these solvability issues, we need some definitions from the theory of *NP-completeness*. This theory deals with so-called *recognition* problems. These problems are also referred to as *feasibility* or *decision* problems. Given an optimization problem, corresponding feasibility problems, requiring a yes/no answer, can be constructed in the following manner. Say, we have a minimization problem, $\min_{x \in \mathcal{X}} f(x)$, then the feasibility problem is asking for the existence of a feasible $x \in \mathcal{X}$ such that $f(x) \leq z$ for some threshold value z .

Suppose we have an “oracle” that returns a “yes” or a “no” answer to our feasibility problem, then we can solve any “reasonable” minimization problem by consulting the oracle polynomial number of times. It is reasonable to expect that we have a finite lower bound, a , and a finite upper bound, b , for the optimum value of the objective function. That is, the oracle answers with a “no” to the feasibility problem with $z = a$ and “yes” to the feasibility problem with $z = b$. Let $u = b - a$ be the initial *interval of uncertainty*. It is also reasonable to be content with an integer interval of uncertainty and the nearest integer for the optimal objective function value. Apply a *bisection search* to the interval of uncertainty, by asking the oracle the feasibility of $z = a + u/2$. Depending on the yes/no answer, the new interval of uncertainty is either $[a, a + u/2]$ or $[a + u/2, b]$. Thus, the interval of uncertainty is reduced by a factor of 2. After performing ℓ bisections, the interval of uncertainty is reduced by a factor of 2^ℓ . If we want the final interval of uncertainty to be of length 1, then we have to perform ℓ bisections:

$$\frac{u}{2^\ell} = 1,$$

or

$$\ell = \log u,$$

Thus, we can obtain the optimal solution after $\log u$ number of calls to the oracle.

Therefore, for the purposes of computational complexity, it is sufficient to be concerned with the feasibility problems. Consider the following definitions.

Class \mathcal{P} A feasibility problem belongs to class \mathcal{P} if, for any instance of the problem, a “yes” or a “no” answer can be *determined* by a polynomial algorithm.

Class \mathcal{NP} A feasibility problem belongs to class \mathcal{NP} , if the feasibility of a given structure can be *checked* in polynomial time.

For example, the feasibility version of the problem $1|P2||C_{max}$ is a member of \mathcal{NP} . If you give me a structure, that is list of jobs assigned to each machine, and a z , I can, in polynomial time, “check” if maximum of the sums of processing times in each machine is less than or equal to z . But, suppose the problem is a variant of the above problem: *all* possible assignments of jobs to two identical machines such that the makespan is less than or equal to z . Then, if you give me, say, two sets of assignments, and claim that they constitute *all* the assignments whose makespan is less than or equal to z , I can check if those two are feasible, but cannot be sure if they are the *all* possible assignments, without computing all possible assignments myself, and I don’t have a way of doing the latter in polynomial time.

It follows from the definitions that $\mathcal{P} \subseteq \mathcal{NP}$.

A problem Π_1 is reducible to another problem Π_2 if for any instance of Π_1 , an instance of Π_2 can be constructed in polynomial time such that solving the instance of Π_2 will solve the instance of Π_1 as well. The reducibility of Π_1 to Π_2 implies that Π_1 can be considered as a special case of Π_2 .

A summary of reducibility among scheduling problems are illustrated in Figure 3.1. In this figure each graph \mathcal{G}_i represents a different characteristic of a scheduling problem and a 5-tuple (v_1, \dots, v_5) , where v_i is a vertex of graph \mathcal{G}_i , represents a particular scheduling problem. In \mathcal{G}_i , the directed path from Π' to Π shows the reducibility in terms of the characteristic v_i . Note that graph \mathcal{G}_3 indicates that there is no reducibility relation between *pmtn* and \circ .

Now the final two definitions in our discussion of computational complexity,

\mathcal{NP} -hard A problem P is called \mathcal{NP} -hard if every problem, P' , in class \mathcal{NP} can be *reduced* to P .

\mathcal{NP} -complete A problem P is called \mathcal{NP} -complete if it is \mathcal{NP} -hard and belongs to class \mathcal{NP} .

All of the above implies the following: if a polynomial time algorithm can be found for any \mathcal{NP} -complete problem, then every problem in class \mathcal{NP} can be solved in polynomial time, and thus proving $\mathcal{P} = \mathcal{NP}$. But this is very unlikely, \mathcal{NP} contains some very difficult combinatorial problems, such as, integer programming and many other problems in graphs theory, network design, sets and partitions, automata and languages, etc. which received considerable amount of research without finding any polynomial

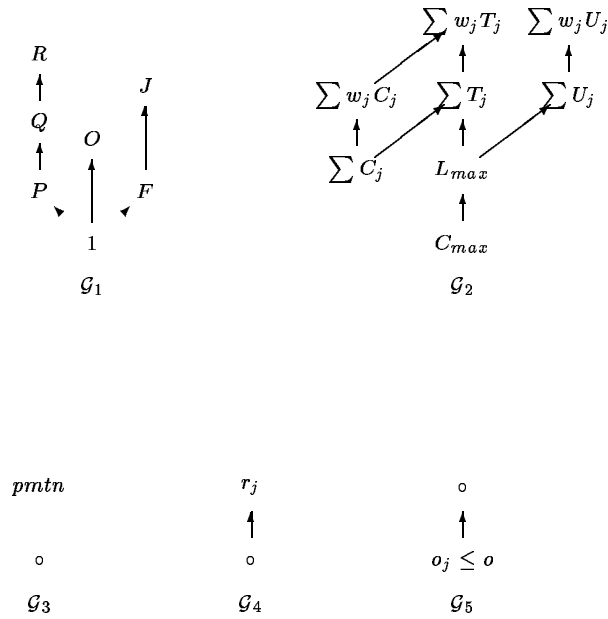


Figure 3.1 Reducibility among scheduling problems

algorithm for these problems[43]. We know $\mathcal{P} \subseteq \mathcal{NP}$, the major unresolved problem is whether $\mathcal{P} \subset \mathcal{NP}$ or $\mathcal{P} = \mathcal{NP}$.

In order to show the \mathcal{NP} -completeness of a problem in \mathcal{NP} , one has to reduce it to a known \mathcal{NP} -complete problem (for an illustration as to how this is done see, e.g. [39]). The practical significance of showing the feasibility version of an optimization problem to be \mathcal{NP} -complete is that one should not pursue the search for a good optimizing algorithm for such a problem and be content with finding a good approximating (i.e. heuristic) algorithm.

3.4 DISJUNCTIVE GRAPH REPRESENTATION

Gantt Charts are very useful visual tools for representing schedules. Another graphical representation of scheduling problems is by means of disjunctive graphs due to [72]. It is widely used in representing single machine multi stage makespan minimization scheduling problems. In the following, this scheme is presented in the context of the

basic job shop scheduling problem, $(Js ||| C_{max})$. Clearly, in this configuration s_{ij} can equivalently be referred as the machine required by the operation O_{ij} . If we let

- $N = \{0, 1, \dots, n \sum_j o_j, *\}$ denote the set of operations (with 0 and * being the dummy “start” and “finish” operations),
- M denote the set of machines,
- $A = \{(i, \ell) \mid i \rightarrow \ell\}$ denote the precedence relation \rightarrow , among the operations i and ℓ of each job,
- $E_k = \{(i, \ell), (\ell, i) \mid \forall i \ \& \ \ell \text{ performed on machine } k\}$ denote set of all possible precedence relations among the operations i and ℓ performed on machine k .
- t_i and p_i (with p_0 and p_* being 0) denote the start and processing times of an operation $i \in N$, respectively,

then the scheduling problem can be restated as

$$\begin{aligned} & \min t_* \\ & \text{s.t.} \\ & \quad t_\ell - t_i \geq p_i \quad \forall (i, \ell) \in A; \\ & \quad t_i \geq 0 \quad \forall i \in N; \\ & \quad t_\ell - t_i \geq p_i \vee t_i - t_\ell \geq p_\ell \quad \forall (i, \ell) \in E_k, k \in M. \end{aligned}$$

Any feasible solution of this problem is called a schedule. In the above formulation the first inequality preserves the precedence constraints and the last inequality (disjunctive constraint) guarantees that each machine processes one job and each job is processed by a single machine at a time. This formulation suggests that the job shop scheduling problem can be represented via a *disjunctive graph* $G = (N, A, E)$ with node set N , ordinary (conjunctive) arc set A and, disjunctive arc set $E = \bigcup_{k \in M} E_k$. Figure 1 illustrates disjunctive graph representation of a 3-job $(J3 \mid \mid o_j \leq 3 \mid C_{max})$. The nodes of G correspond to operations, the directed arcs in A correspond to precedence relations among operations of a job and the disjunctive arcs in E_k correspond to precedence relations among operations performed on machine k . The number on the node i is the processing time of the operation i . Let

- A *selection* $S_k = \{(u, v) \mid (v, u) \in E_k \setminus S_k\}$. If S_k is acyclic then it corresponds to a unique sequence of operations on machine k .
- A *complete selection* $S = \bigcup_{k \in M} S_k$. S is called acyclic if the digraph $D_S = (N, A \cup S)$ obtained by replacing E with S is acyclic. Acyclic D_S represents a unique feasible schedule for the problem. The makespan of this schedule is the longest path of D_S . For the above example, figure 2 depicts a digraph and *Gantt chart* representations of a feasible schedule.

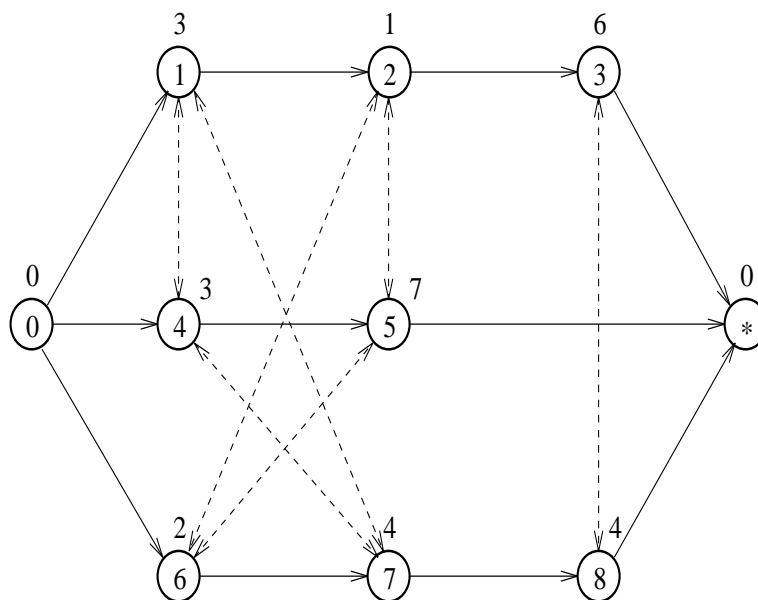
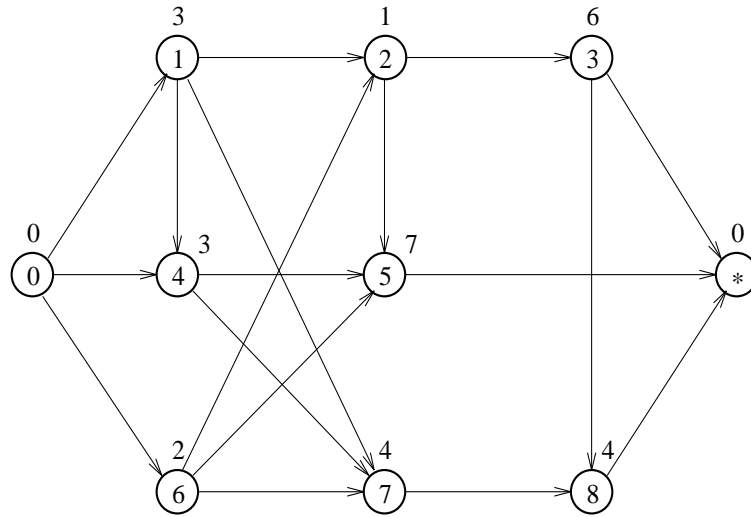
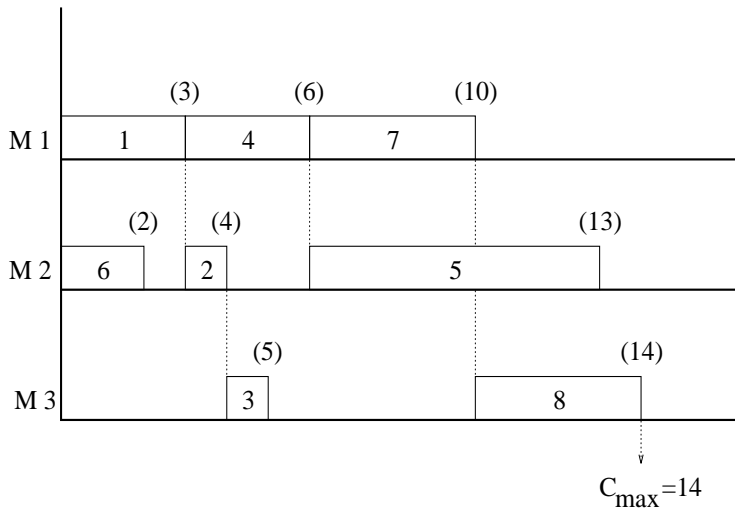


Figure 3.2 Disjunctive graph representation of a 3-job ($J3 \mid \mid o_j \leq 3 \mid C_{max}$).

With the above notation, ($J3 \mid \mid \mid C_{max}$) can be restated as determining an acyclic $S \subset E$ that minimizes the length of a longest path in the digraph D_S .



(a)



(b)

Figure 3.3 A (feasible) schedule for the 3-job ($J3 \mid \mid o_j \leq 3 \mid C_{\max}$): (a). Digraph and, (b). Gantt Chart representations

3.5 CONCLUDING REMARKS

The seminal work on the theory of machine scheduling is [26]. The classic texts are [6] and [42], both of them seem to be currently out of print. A recent book by Baker, [8], does provide a very readable introduction to the basic results in the area. Other book containing most of recent results and having extensive discussions on computational complexity and analysis of heuristics is [12]. [50] has an introductory chapter on scheduling. Heuristic methods are thoroughly covered in [66]. A recent survey of the area, together with classification, complexity results and open problems can be found in [61]. Complexity results of static scheduling problems can also be found in [51].

The advances in manufacturing automation and the changing nature of business competitiveness created demand for viable production scheduling systems. In order to achieve this one needs to integrate production planning (e.g. lot sizing) with production scheduling and sequencing in structurally sound manner with acceptable computational burden.

Origins of lot sizing dates back to [49] and that of scheduling to the early work of H. L. Gantt who developed the *Gantt Charts* during the World War I to handle scheduling problems associated with loading cargo to Allied ships. Not much done until 1950's, when there were rigorous mathematical analyses in both areas; such as the work of Arrow and his colleagues [5] and Dvoretzky and his colleagues [31, 32] in inventory theory and Johnson [58] and Smith [76] in scheduling theory. Although there were few isolated studies, such as [33] and [60], which more or less attacked the lotsizing and sequencing problems jointly, the two areas developed quite independent of each other. Holstein [55] probably was the first to argue explicitly for the necessity of integrating these two areas.

After the publication of the critical articles on the state-of-the-art in these areas by the eminent researchers such as Wagner [79], Silver [74], and Graves [46], 1980's has witnessed the resurgence of research in integrating systems. In North America, the Hierarchical Production Planning was developed by a group at MIT (see [11] for an overview), a group at Cornell proposed a framework [63], developed COSMOS [67], and the resulting by-products such as XCELL+ [25] and PRS [20]. In Europe, GRAI Method was developed at the Automation Laboratory of University of Bordeaux 1 [30]. Also during this decade, the emerging production management paradigms, such as MRPII, OPT, and JIT, received considerable attention by both practitioners and researchers in the area (see [19] for an overview).

Some of the recent work concerning the integration issues are [28], [29], [70], [71], [23].

The central role played by mixed integer programming in production planning and scheduling models is stressed in [73]. As stated in [9], the successful solution of large-scale mixed integer programs requires formulations whose LP-relaxations give a good approximation to the convex hull of feasible solutions. Stronger mixed integer program-

ming formulations have led to computational breakthroughs in number of large-scale problems; see, for example, [54] and [56]. The recent methodologies of branch-and-cut and branch-and-price, by themselves may not be sufficient in solving the huge mixed integer programming problems resulting from integrated production planning models. Coarse grained parallel computing may offer the needed additional computational capability [73, 18].

3.6 PROBLEMS ON SCHEDULING AND SEQUENCING

1. Suppose that 12 jobs must be processed in an open shop with six stages having a single machine in each. How many different sequences are there? If your computer could evaluate 100 schedules every second, how much time would be required in order to evaluate all feasible sequences.
2. For each of the problems below, indicate precisely what or who would correspond to jobs and who or what would correspond to machines. In each case discuss what objectives might be appropriate and special priorities that might exist.
 - Treating patients in a hospital emergency room.
 - Unloading cargo from ships at port.
 - Serving users on a time-shared computer system.
 - Transferring long-distance phone calls from one city to another.
3. Seven jobs are to be processed through a single machine. The processing times and the due dates are given below:

Job	1	2	3	4	5	6	7
Processing Time	3	6	8	4	2	1	7
Due Date	4	8	12	15	11	25	21

Determine the sequence that minimizes

- a Mean flow time,
 - b number of tardy jobs,
 - c Maximum lateness,
 - d Makespan.
4. Consider $F3||C_{max}$ problem with the following processing time data:

Job	Machine		
	A	B	C
1	4	2	6
2	2	3	7
3	6	5	6
4	3	4	8

What is the makespan? Draw a Gantt chart for your solution.

5. Suppose you own a auto paint shop. You have currently five cars waiting to be painted. You can only paint one car at a time. The time it takes to paint these cars and your expected profit from each car are:

Car	Painting time (hr.s)	Profit (\$100)
A	6	2
B	2	1
C	5	4
D	4	1
E	3	2

In which order would you paint the cars? Why?

6. Solve a $F3||C_{max}$ problem with the following processing times

Job	Machine		
	A	B	C
1	12	5	13
2	6	10	13
3	9	11	18
4	17	16	4

7. Consider an $F3||C_{max}$ problem:

Job	Machine			Due Date
	A	B	C	
1	7	5	3	5
2	6	9	8	12
3	13	6	5	16
4	2	4	7	20

- a Draw a Gantt Chart for the permutation sequence {4-2-3-1}.
- b Determine the makespan, the mean flow time, maximum lateness, and maximum tardiness for the sequence in (a).

- c Instead of a single machine in the first stage, suppose you have four identical parallel machines (in stages two and three, you still have a single machine each). That is, the problem is $F3|P4, 1, 1||C_{max}$. What is the improvement on the makespan?
- d What is the improvement if you have four machines in every stage? i.e. $F3|P4||C_{max}$.
8. We know that mean lateness and mean tardiness are not necessarily equal, are maximum tardiness and maximum lateness always equal? Explain.
9. A manufacturer of charm bracelets has five jobs to schedule for a leading customer. Each job requires a stamping operation followed by a finishing operation, which can begin on an item immediately after its stamping is complete. The table below shows operation times per item (in minutes) for each job in the order. In addition, preparations for each job at the stamping facility require a setup before processing begins, as shown in the table below. Find a schedule that completes all five jobs as soon as possible.

Job	Items in Lot	(min/item)		Setup Times
		Stamp.	Finish.	
1	20	2	8	100
2	25	2	5	250
3	100	1	2	60
4	50	4	2.5	60
5	40	3	6	80

10. Consider the following $F3|||C_{max}$ problem

Jobs	1	2	3	4	5	6
$p_{j,1}$	6	12	4	3	6	2
$p_{j,2}$	7	2	6	11	8	14
$p_{j,3}$	3	3	8	7	10	12

Given that the C_{max} of the sequence $\{3-5-6-4-2-1\}$ is 57, find an optimal sequence.

11. Consider the following three job problem with all ready times equal to zero:

Job	1	2	3
Processing Time	8	6	4
Weight	2	1	2
Due Date	10	10	10

Assume that preemption is not allowed.

- (a) Sequence the jobs such that weighted mean flow time is minimized.

- (b) Evaluate your schedule by each of the following objectives:
- i. makespan,
 - ii. weighted flow time,
 - iii. weighted lateness,
 - iv. weighted tardiness,
 - v. maximum flow time,
 - vi. maximum lateness,
 - vii. maximum tardiness, and
 - viii. weighted number of tardy jobs.

12. Solve the following $F3||C_{max}$ problem by branch & bound.

Jobs	1	2	3
$p_{j,1}$	1	2	3
$p_{j,2}$	3	2	1
$p_{j,3}$	2	1	1

13. (a) State in words exactly what the problem $(J3 || o_j \leq 3 | C_{max})$ is.
- (b) Given that $(J3 || o_j \leq 3 | C_{max})$ is NP-hard, based ONLY on the reducibility graphs, Figure (3.1), what can you say about,
- $(F3 || o_j \leq 3 | C_{max})?$
 - $(J3 || o_j \leq 3 | L_{max})?$
 - $(J3 || o_j \leq 3, pmtn | C_{max})?$
- (c) $(J3 || o_j \leq 3 | C_{max})$ is an optimization problem, state its corresponding decision (or, feasibility, or, recognition) problem.

4

LOCATION AND DISTRIBUTION

Location and distribution problems are intimately related. In location problems we wish to locate a number of facilities at certain points in order to optimize a given measure of performance. The facilities to be located can be plants, warehouses, distribution centers, machines on shop floor, transistors on a chip, etc. Whatever the system is, these facilities have some form of interaction among themselves and possibly with other facilities whose locations are fixed. For example, a number of items produced at certain plants are to be shipped to warehouses at a number of locations which in turn ship these items to regional distribution centers that serve individual customers at various final destinations. Distribution problems, on the other hand, determine how this flow of material should take place so that a given performance measure is optimized.

The simplest distribution problem is the transportation problem of linear programming. We wish to determine the amounts to be shipped from supply points at fixed locations with known amounts of available material, to demand points at fixed locations with known amounts of demand so as to minimize the total transportation costs. In this problem, since the locations of the facilities (i.e., the supply and demand points) are fixed, the distances among facilities are given parameters of the problem. The decision variables in the problem are the amount of interaction (i.e., the number of units shipped) among facilities. In the quadratic assignment problem, on the other hand, locations are the unknowns of the problem and the amount of interactions among facilities are the givens.

We can think of locations of the facilities implying a measure of *proximity* among facilities and the amounts of interaction between two facilities the facilities implying a measure of *intensity* between these facilities. Measures of proximity can be the distance or the time required to travel or the unit cost of shipment between facilities. Measures of intensity among facilities can be amount of material shipped per unit time or amount of communication required between a pair of facilities or some weight denoting the relative importance of one facility to the other.

4.1 FACILITY LOCATION MODELS

Let \mathcal{F} be the index set of facilities whose locations are the decision variables in the problem. In general, the cardinality of \mathcal{F} can also be decision variable of the problem. Define \mathcal{G} to the index set of facilities whose locations are known in advance and hence are the parameters of the problem. If we denote the level of proximity between any two facilities, i and j , by d_{ij} and the level of intensity by w_{ij} , then their joint contribution to the objective function can be expressed by $(d_{ij}w_{ij})$.

Depending upon the “geographical” restrictions on where the facilities are to be located, the location problems can be considered in three broad classes:

Planar Location Facilities can be located at any point on the plane. Number of possible locations are infinite.

Network Location Facilities can be located at any point on a given network. All the nodes and every point on the arcs of the network are possible candidate points, again resulting in infinite number of possible locations.

Discrete Location Facility locations can be chosen among a finite set of given candidate points.

In planar location problems, the measure of proximity, d_{ij} , is the planar distance, ℓ_p . Let (x_i, y_i) denote the coordinates of a point, i , on the plane. Given any two points, i and j , on the plane, the ℓ_p distance between them is given by

$$[|x_i - x_j|^p + |y_i - y_j|^p]^{1/p}, \text{ for } p \geq 1,$$

where,

$p = 1$, gives the rectilinear distance value, and

$p = 2$, is the Euclidean distance; when

$1 < p < 2$, the value of the distance lies between the rectilinear and Euclidean distance values, and for

$p > 2$, the distance value lies below the Euclidean and continues to decrease as p increases. Finally, as

$p \rightarrow \infty$, it takes the limiting value of so called Tchebychev distance,

$$\max \{|x_i - x_j|, |y_i - y_j|\},$$

which is a convenient measure, for example, in modeling location problems involving the location of items moved in and out of an automated storage and retrieval systems [40].

In network location problems distance is measured as the shortest distance in the network; and in discrete location problems, it is part of the given data set.

In location problems there are two primary measures of performance:

Minimax which minimizes the maximum contribution among all pairs of facilities; also referred to as p -centers problem,

$$\min \max_{(i,j)} \{d_{ij}w_{ij}\}, \quad i, j \in \mathcal{F} \cup \mathcal{G},$$

Minisum which minimizes the sum of contributions of all facility pairs ; also referred to as p -medians problem,

$$\min \sum_{(i,j)} \{d_{ij}w_{ij}\}, \quad i, j \in \mathcal{F} \cup \mathcal{G}.$$

Exercise 4.1 Which measure of performance would you choose in locating

- fire stations in a major city,
- a tool crib on a shop floor?

Exercise 4.2 Formal mathematical models of location problems go at least as back as Pierre de Fermat (1601–1665) who proposed a basic form of the Euclidean distance minisum problem by issuing the challenge [82]:

...let he who does not approve of my method, attempt the solution of the following problem: given three points in the plane, find the fourth point such that the sum of its distances to the three given points is a minimum.

Formulate the Fermat's problem using the notation introduced in this chapter and suggest a solution approach.

The standard textbook on location problems is [41]. An excellent review of field by the same authors can be found in the article [40]. More advanced books on location problems are [62] and [65]. Discrete location models are reviewed in [3]. An overview of representative problems in location research is given in [17]. Algorithms and solution results for network location problems are comprehensively covered in [21, 22].

In most of the location problems, it is desirable to be close to the facilities to be located. However, some facilities are “undesirable” or “semi-desirable”, such as nuclear reactors, military installations, and polluting plants. It may be meaningful to use maximin and maxisum objective in locating such facilities. A comprehensive review of related research can be found in [36].

4.2 DISCRETE LOCATION PROBLEMS

Let m denote the number of candidate facility locations, indexed by $i = 1, \dots, m$. Suppose there are n demand centers, indexed by $j = 1, \dots, n$, where the services or the products of the facilities are required. Let the amount demanded at demand center j be d_j which is in units of quantity per period, such as tons per month, boxes per day, etc. A candidate facility location may have a limited capacity for production of goods or services, denote this capacity by k_i , in same units as those for d_j .

There are two sets of decision required: selection of locations for the facilities, and how much each open facility will ship to each demand center. Define the following decision variables, for $i = 1, \dots, m$, and $j = 1, \dots, n$:

$$y_i = \begin{cases} 1 & \text{if a facility is located at } i, \\ 0 & \text{otherwise.} \end{cases}$$

x_{ij} = amount to be transported from i to j , per unit time.

The objective function is minimization of the total costs of fixed costs of locating facilities and the variable costs of transportation from these supply points to demand centers. Let f_i be the fixed cost corresponding the candidate location i . The parameters f_i 's are in units of costs per unit time, where unit time is same as the unit used in d_j ; such as dollars per year. These costs may include amortized capital costs, annualized lease costs, insurance, taxes, and other overhead costs whose total is not directly related to the facility throughput [40]. Finally, the variable costs of transporting one unit from i to j is denoted by c_{ij} .

Then, the discrete facility location problem can be modeled as the following mixed integer linear programming problem:

$$\min \sum_i f_i y_i + \sum_{i,j} c_{ij} x_{ij}$$

subject to:

$$\begin{aligned} \sum_i x_{ij} &\geq d_j, \quad j = 1, \dots, n, \\ \sum_j x_{ij} &\leq k_i y_i, \quad i = 1, \dots, m. \end{aligned}$$

where $x \geq 0$ and $y = 0, 1$.

Note that when we choose the locations by fixing the values of y_i 's, i.e. projecting the problem onto the space of y -variables, the resulting is a transportation problem. We can use a solution strategy in which we fix each y_i to 0 or 1, and then solve the resulting transportation problem. We then can fix another set of y 's and solve the transportation problem, etc. If we were to solve for all sets of y , the solution giving

the minimum objective function value would be the optimum solution. However, if we can use the transportation problem to help us calculate a new set of y 's which leads to a lower value of the objective function, then we can converge to the optimal solution without having to check explicitly all possible sets of y . This is exactly what Benders' Decomposition does when applied to location and distribution problems.

Exercise 4.3 (Dual of the Transportation Problem [68]) *Suppose ore is available at m mines and is required at n plants. Let $b_j (> 0)$ denote the minimum amount of ore required at plant $j = 1, \dots, n$, and let $a_i (> 0)$ denote the maximum amount of ore that can be shipped from mine $i = 1, \dots, m$. Let c_{ij} be the cost of shipping one unit of ore from the i th mine to the j th plant. The problem is to determine how much ore to ship from each mine to each plant (x_{ij}) in order to meet the requirements at the plants with minimum transportation cost.*

Then a trucking company offers to take all the available ore at the mines and deliver the required amounts ore at the plants. Let u_i be the price they agree to pay for one unit of ore at the i th mine and let v_j be the price they agree to sell one unit of ore at the j th plant. Discuss an economic interpretation of the dual problem as that of maximizing the net revenue of the trucking company.

4.2.1 Benders' Decomposition

Benders' Decomposition [10] is a resource directive decomposition for solving large scale linear (mixed-integer) programming problems with coupling (or, complicating) variables. Whereas, Dantzig-Wolfe Decomposition [27] is a price directive decomposition for solving large scale mathematical programming problems with coupling (or, complicating) constraints. Therefore, it is natural to view these as duals of one another.

Decomposition algorithms, an addition to being computational techniques, can also be interpreted as descriptive theoretical models of planning. In a multi-divisional firm or in a multi-sector economy, the master problem can represent the actions of the coordinating center, while the subproblems correspond to the divisions or sectors. This interpretation can also be viewed as the simulation of decentralized planning procedures subject to central resource constraints and centrally defined objectives. It is important to point out that the decomposition algorithm, in this context, is the descriptive model of the preparation of the planning decisions, not the model of the everyday operation and of everyday control of the economic system.

A brief outline of Benders' Decomposition as applied to the discrete location problems is given below. Rewrite the location problem of the previous section in matrix notation as,

$$\min_{x \geq 0, y = 0, 1} \{cx + dy | Ax + Dy \geq b\}.$$

Define $\mathcal{Y} = \{y = 0, 1 \mid \exists x \geq 0, Ax \geq b - Dy\}$. Projecting onto the y -variables,

$$\min_{y \in \mathcal{Y}} \{dy + \min_{x \geq 0} [cx \mid Ax \geq b - Dy]\}.$$

The inner minimization in the above problem is balanced transportation problem that is always feasible and has finite optimum. [Why?]. Therefore we can replace it with its dual,

$$\min_{y \in \mathcal{Y}} \{dy + \max_{u \geq 0} [u(b - Dy) \mid uA \leq c]\}$$

where u is the dual vector corresponding to the constraints $\{Ax \geq b - Dy\}$. Let $\{u^p, p = 1, \dots, P\}$ be the extreme points of $\{u \geq 0 \mid uA \leq c\}$, then

$$\begin{aligned} & \min_{y \in \mathcal{Y}} \{dy + \max_{1 \leq p \leq P} [u^p(b - Dy)]\}, \text{ or} \\ & \min_{y \in \mathcal{Y}, w} \{dy + w \mid w \geq u^p(b - Dy), p = 1, \dots, P\}. \end{aligned}$$

This problem which is equivalent to the original problem, has P constraints which can be a huge number. But it is likely that at optimality only a small subset of these constraints will be active. A natural solution strategy to apply for such problems is *relaxation*. Basically this solution strategy starts solving the problem with a small subset of the constraints, other constraints being “relaxed”. The optimal solution thus obtained is checked if it satisfies the “relaxed” constraints. If it does, then we are done, otherwise, the most violated constraint is appended to the subset. [We may also discard some of the amply satisfied constraints (i.e. ones with the greatest slacks), when and if it is desirable to keep the size of the problem reasonable.] And the program is solved with the updated set of constraints. Thus, in this approach, instead of solving a large problem once, a number of smaller problems are solved many times.

The relaxed original problem is referred to as the *master problem* and the problem of determining the most violated constraint is called the *subproblem*. Let $S \subseteq \{u^p, k = 1, \dots, P\}$. Then the master problem is:

$$\min_{y \in \mathcal{Y}, w} \{dy + w \mid w \geq u^p(b - Dy), u^p \in S\}.$$

Let its optimal solution be (y°, w°) . And the subproblem is given by:

$$\max_{u \geq 0} \{u(b - Dy^\circ) \mid uA \leq c\},$$

and its dual is the transportation problem,

$$\min_{x \geq 0} \{cx \mid Ax \geq b - dy^\circ\}.$$

Let their optimal solutions, respectively, be u° and x° . Finally, if $w^\circ \geq u^\circ(b - Dy^\circ)$, then (y°, x°) is optimal, otherwise $S \leftarrow S + \{u^\circ\}$, and resolve the master problem. For a detailed analysis of the use of Benders Decomposition in distribution systems design see Geoffrion and Graves [44].

Exercise 4.4 Write down a pseudocode for Benders’ Decomposition of the discrete location problem discussed in the previous section.

4.2.2 An Institutional Interpretation

It will be illustrative to discuss the master and sub-problems in terms of the following institutional interpretation. Consider two involved parties: The Producer and The Transporter. The Producer's concern is to determine at which sites the plants are to be constructed so as the total (fixed and variable) costs are minimized.

After determining the initial location vector, i.e. sites at which plants are to be constructed, The Producer subcontracts the remainder of the work, namely the production levels and transportation flows, to The Transporter. The Transporter's concern, after receiving the "statement of work" is to determine the production levels at the sites where plants are to be constructed and the transportation flows of commodities from the plants to the customers so as to minimize the transportation costs. This is accomplished by The Transporter by solving a transportation problem. After solving this problem, a "bid" is sent by The Transporter to the Producer stating how much the transportation will cost. In addition, The Transporter also sends a "proposal" which contains information about how the bid would be affected for certain changes in the statement of work.

The Producer, based on the proposal (and the bid) received from The Transporter, and other restrictions such as the configurational constraints on the opening of plants, checks whether the locations that were specified in the statement of work to The Transporter are optimal levels. If they are not yet optimal, a new statement of work is drafted and send to The Transporter. This iterative process continues until the location vector is optimal (or satisfactorily close to optimal), then the problem is solved, which is the termination of the overall solution procedure.

Exercise 4.5 *Suppose we have the following warehouse location problem. There are three customers, $j = 1, 2, 3$, with daily demands of 400, 200, and 300 units, respectively. There are three candidate sites, $i = 1, 2, 3$, for leasing warehouses. Warehouse fixed costs would be \$1,000 per day at each location for a capacity of 1,000 units or less. The unit transportation costs from each candidate site to each customer location are given below,*

\$/unit	Customer		
	1	2	3
Site			
1	1	4	5
2	4	1	3
3	6	4	1

Solve the problem and identify explicitly the information flow between the master and the subproblems in every major iteration.

4.3 VEHICLE ROUTING AND SCHEDULING

Design of a distribution system resulting in the number, locations, and sizes of facilities, and assignment of facilities to demand zones is basically a strategic issue. Routing and scheduling of vehicles on such a system is crucial from tactical and operational viewpoint. A system that is optimal in the strategic sense may not be the best choice when tactical and operational aspects are considered [15]. Given a transportation network and demands for service at various points in the network, “spatial” configuration of vehicle movements are the concern of *routing* problems; when “temporal” decisions are also important, that is, if explicit consideration is given to the times at which various locations are visited, we are faced with a vehicle scheduling problem [14].

The basic problem can be stated as follows [35]: A set of customers, each with a known location and a known requirement for some commodity, is to be supplied from a single facility by vehicles of known capacity. The problem is to design the minimum cost routes of these vehicles, subject to:

1. The requirements of all the customers must be met.
2. The total load allocated to each vehicle may not exceed its capacity.
3. The total time (or, distance) for each vehicle to complete its route may not exceed some predetermined value.
4. There is an earliest and a latest time within which a customer can accept a delivery.

When the last three constraints are relaxed, the problem reduces to traveling salesman problem, which is known to be *strongly NP-Hard*. The additional constraints can only make the solution even more difficult. If the last two constraints are relaxed, the problem is to determine the minimum number of vehicles, which is closely related to the classical knapsack problem. See [24] for various formulations and extensive bibliography in this area.

REFERENCES

- [1] Russel L. Ackoff. *Scientific Method: Optimizing Applied Research Decisions*. Wiley, 1962.
- [2] Alfred V. Aho, John E. Hopcroft, and Jeffrey D. Ullman. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, 1974.
- [3] C. H. Aikens. Facility location models for distribution planning. *European J. Operational Research*, 22:263–279, 1985.
- [4] H. Cemal Akyel. *Minimizing Schedule Length on Identical Parallel Machines: An Exact Algorithm*. PhD thesis, The Institute of Engineering and Science, Bilkent University, 1991.
- [5] K. A. Arrow, T. E. Harris, and J. Marschak. Optimal inventory policy. *Econometrica*, 19:250–272, 1951.
- [6] K. R. Baker. *Introduction to Sequencing and Scheduling*. Wiley, 1974.
- [7] K. R. Baker. Requirements planning. In S. C. Graves, A. H. G. Rinnooy Kan, and P. H. Zipkin, editors, *Logistics of Production and Inventory*, volume 4 of *Handbooks in Operations Research and Management Science*, chapter 9, pages 571–628. North-Holland, 1993.
- [8] K. R. Baker. *Elements of Sequencing and Scheduling*. The Amos School of Business Administration, Dartmouth College, Hanover, N.H., 1994.
- [9] C. Barnhart, E. L. Johnson, G. L. Nemhauser, M. W. P. Savelsbergh, and P. H. Vance. Branch-and-Price: Column generation for solving huge integer programs. In J. R. Birge and K. G. Murty, editors, *Mathematical Programming: State of the Art 1994*, pages 186–207. MPS, The University of Michigan, 1994.
- [10] J. F. Benders. Partitioning procedures for solving mixed variables problems. *Numer. Math.*, 4:238–252, 1962.
- [11] G. R. Bitran and D. Tirupati. Hierarchical production planning. In S. C. Graves, A. H. G. Rinnooy Kan, and P. H. Zipkin, editors, *Logistics of Production and Inventory*, volume 4 of *Handbooks in Operations Research and Management Science*, chapter 9, pages 523–568. North-Holland, 1993.
- [12] J. Blazewicz, K. Ecker, G. Schmidt, and J. Weglarz. *Scheduling in Computer and Manufacturing Systems*. Springer-Verlag, 1993.

- [13] J. Blazewicz, J. K. Lenstra, and A. H. G. Rinnooy Kan. Scheduling subject to resource constraints: Classification and complexity. *Discrete Appl. Math.*, 5:11–24, 1983.
- [14] L. Bodin, B. Golden, A. Assad, and M. Ball. Routing and scheduling of vehicles and crews: The state of the art. *Computers and Operations Research*, 10:63–211, 1983.
- [15] J. H. Bookbinder and K. E. Reece. Vehicle routing considerations in distribution system design. *European J. Operational Research*, 37:204–213, 1988.
- [16] S. P. Bradley, A. C. Hax, and T. L. Magnanti. *Applied Mathematical Programming*. Addison-Wesley, 1977.
- [17] M. L. Brandeau and S. S. Chiu. An overview of representative problems in location research. *Mgmt Sci.*, 35:645–674, 1989.
- [18] R. W. Brown, J. F. Shapiro, and P. J. Waterman. Parallel computing for production scheduling. *Manuf. Syst.*, 6:56–64, 1988.
- [19] Jimmie Browne, John Harhen, and James Shivnan. *Production Management Systems: A CIM Perspective*. Addison-Wesley, 1988.
- [20] C-WAY, 2359 N. Triphammer Road, Ithaca, New York 14850, USA. *A Guided Tour of PRS: The Production Reservation System*, 1990.
- [21] B. Ç. Tansel, R. L. Francis, and T. J. Lowe. Location on networks: A survey—part i. the p -center and p -median problems. *Mgmt Sci.*, 29:482–497, 1983.
- [22] B. Ç. Tansel, R. L. Francis, and T. J. Lowe. Location on networks: A survey—part ii. exploiting the tree network structure. *Mgmt Sci.*, 29:498–511, 1983.
- [23] Shi-Chung Chang and Fu-Shiung Hsieh. Integrated order and production scheduling for flow shops. In G. Doumeingts, J. Browne, and M. Tomljanovich, editors, *Computer Applications in Production and Engineering: Integration Aspects*, pages 621–628. IFIP, Elsevier Science Publishers B. V. (North-Holland), 1991.
- [24] N. Christofides. Vehicle routing. In E. L. Lawler, J. K. Lenstra, A. H. G. Rinnooy Kan, and D. B. Shmoys, editors, *The Traveling Salesman Problem*, chapter 12. Wiley, 1985.
- [25] R. Conway, W. L. Maxwell, J. O. McClain, and S. L. Worona. *User's Guide to XCELL+: Factory Modeling System*. The Scientific Press, third edition, 1990.
- [26] Richard W. Conway, William L. Maxwell, and Louis W. Miller. *Theory of Scheduling*. Addison-Wesley, 1967.
- [27] G. B. Dantzig and A. Wolfe. The decomposition algorithm for linear programming. *Econometrica*, 29:767–778, 1961.

- [28] Stephane Dauzere-Peres and Jean B. Lasserre. *An Integrated Approach in Production Planning and Scheduling*, volume 411 of *Lectures Notes in Economics and Mathematical Systems*. Springer-Verlag, 1994.
- [29] Stephane Dauzere-Peres and Jean-Bernard Lasserre. Integration of lotsizing and scheduling decisions in a job-shop. *European J. Operational Research*, 75:413–426, 1994.
- [30] G. Dougmeings, L. Pun, M. Mondain, and D. Breuil. Decision making system for production control, planning, and scheduling. *Int. J. Prodn Res.*, 16:137–152, 1978.
- [31] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. The inventory problem: I. case of known distributions of demand. *Econometrica*, 20:187–222, 1952.
- [32] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. The inventory problem: II. case of unknown distributions of demand. *Econometrica*, 20:450–466, 1952.
- [33] B. P. Dzielinski and R. E. Gomory. Optimal programming of lot lot sizes, inventory, and labor allocations. *Mgmt Sci.*, 11:874–890, 1965.
- [34] J. Edmonds. Paths, trees, and flowers. *Canad. J. Math.*, 17:449–467, 1965.
- [35] S. Eilon, C. D. T. Watson-Gandy, and N. Christofides. *Distribution Management: Mathematical Modelling and Practical Analysis*. Hafner, 1971.
- [36] E. Erkut and S. Neuman. A survey of analytical models for locating undesirable facilities. *European J. Operational Research*, 40:275–291, 1989.
- [37] Donald Erlenkotter. An early classic misplaced: Ford W. Harris’s economic order quantity model of 1915. *Mgmt Sci.*, 35:898–900, 1989.
- [38] Donald Erlenkotter. Ford Whitman Harris and the economic order quantity model. *Ops Res.*, 38:937–946, 1990.
- [39] M. Florian, J. K. Lenstra, and A. H. G. Rinnooy Kan. Deterministic production planning: Algorithms and complexity. *Mgmt Sci.*, 26:669–679, 1980.
- [40] R. L. Francis, L. F. McGinnes, and J. A. White. Locational analysis. *European J. Operational Research*, 12:220–252, 1983.
- [41] R. L. Francis, L. F. McGinnes, and J. A. White. *Facility Layout and Location*. Prentice-Hall, second edition, 1992.
- [42] Simon French. *Sequencing and Scheduling: An Introduction to the Mathematics of the Job-Shop*. Ellis Horwood Series in Mathematics and Its Applications. John Wiley & Sons, 1982.
- [43] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Freeman, 1979.

- [44] A. M. Geoffrion and G. W. Graves. Multicommodity distribution system design by benders decomposition. *Mgmt Sci.*, 20:822–844, 1974.
- [45] R. L. Graham, E. L. Lawler, J. K. Lenstra, and A. H. G. Rinnooy Kan. Optimization and approximation in deterministic sequencing and scheduling theory: A survey. *Annals of Discrete Mathematics*, 5:287–326, 1979.
- [46] S. C. Graves. A review of production scheduling. *Ops Res.*, 29:646–675, 1981.
- [47] S. C. Graves, A. H. G. Rinnooy Kan, and P. H. Zipkin, editors. *Logistics of Production and Inventory*, volume 4 of *Handbooks in Operations Research and Management Science*. Elsevier Science, 1993.
- [48] G. Hadley and T. M. Whitin. *Analysis of Inventory Systems*. Prentice-Hall, 1963.
- [49] Ford W. Harris. How many parts to make at once. *Factory, The Magazine of Management*, 10:135–136,152, 1913. Reprinted in *Ops Res.*, 38:947–950, 1990.
- [50] A. C. Hax and D. Candea. *Production and Inventory Management*. Prentice-Hall, 1984.
- [51] Jeffrey W. Herrmann, Chung-Yee Lee, and Jane L. Snowdon. A classification of static scheduling problems. In Panos M. Pardalos, editor, *Complexity in Numerical Optimization*, pages 203–253. World Scientific Publishing Co., P.O.Box 128, Farrer Road, Singapore 9128, 1993.
- [52] Frederick S. Hillier and Gerald J. Lieberman. *Introduction to Operations Research*. Holden-Day, fourth edition, 1986.
- [53] James K. Ho and William A. McKenney. Triangularity of the basis in linear programs for material requirements planning. *Opns. Res. Lett.*, 7:273–278, 1988.
- [54] K. Hoffman and M. Padberg. LP-based combinatorial problem solving. *Annals of Operations Research*, 4:145–194, 1985.
- [55] W. K. Holstein. Production planning and control integrated. *Harvard Business Review*, May-June 1968.
- [56] Ellis L. Johnson. Modeling and strong linear programs for mixed integer programming. In Stein W. Wallace, editor, *Algorithms and Model Formulations in Mathematical Programming*, pages 1–43, Berlin Heidelberg, 1989. NATO ASI Series, Vol. F51, Springer-Verlag.
- [57] Lynwood A. Johnson and Douglas C. Montgomery. *Operations Research in Production Planning, Scheduling, and Inventory Control*. Wiley, 1974.
- [58] S. M. Johnson. Optimal two and three-stage production schedules with setup times included. *Naval Research Logistics Quarterly*, 1:61–68, 1954.
- [59] George E. Kimball. General principles of inventory management. *J. of Manufacturing and Opns. Mgmt.*, 1:119–130, 1988.

- [60] L. S. Lasdon and R. C. Terjung. An efficient algorithm for multi-item scheduling. *Ops Res.*, 19:946–969, 1971.
- [61] E. L. Lawler, J. K. Lenstra, A. H. G. Rinnooy Kan, and D. B. Shmoys. Sequencing and scheduling: Algorithms and complexity. In S. C. Graves, A. H. G. Rinnooy Kan, and P. H. Zipkin, editors, *Logistics of Production and Inventory*, volume 4 of *Handbooks in Operations Research and Management Science*, chapter 9, pages 445–522. North-Holland, 1993.
- [62] R. F. Love, J. G. Morris, and G. O. Wesolowsky. *Facilities Location: Models & Methods*. North-Holland, 1988.
- [63] W. Maxwell, J. Muckstadt, L. J. Thomas, and J. Vander Eecken. A modeling framework for planning and control of production in discrete parts manufacturing and assembly systems. *Interfaces*, 13:92–104, 1983.
- [64] Harlan C. Meal. Putting production decisions where they belong. *Harvard Business Review*, 62:102–111, 1984.
- [65] P. B. Mirchandani and R. L. Francis, editors. *Discrete Location Theory*. Wiley, 1990.
- [66] T. E. Morton and D. W. Pentico. *Heuristic Scheduling Systems with Applications to Production Systems and Project Management*. Wiley, 1993.
- [67] J. A. Muckstadt, P. L. Jackson, W. T. Martin, S. Bellantoni, and R. Ferstenberg. Cornell simulator of manufacturing operations (COSMOS). Technical Report 684, SORIE, Cornell University, 1986.
- [68] Katta G. Murty. *Linear Programming*. Wiley, 1983.
- [69] Evan L. Porteus. Stochastic inventory theory. In D. P. Heyman and M. J. Sobel, editors, *Stochastic Models*, volume 2 of *Handbooks in Operations Research and Management Science*, chapter 12, pages 605–652. North-Holland, 1990.
- [70] C. N. Potts and L. van Wassenhove. Integrating scheduling with batching and lot-sizing: a review of algorithms and complexity. *Journal of Operational Research Society*, 48:395–406, 1992.
- [71] D. B. Pressmar. LP-models in production planning and control. In G. Fandel, Thomas Gullledge, and Albert Jones, editors, *New Directions for Operations Research in Manufacturing*, pages 91–100, Berlin – Heidelberg, 1992. Springer-Verlag. Proceedings of a Joint US/German Conference, Gaithersburg, Maryland, USA, July 30–31, 1991.
- [72] B. Roy and B. Sussmann. Les problèmes d’ordonnancement avec contraintes disjointives. Note DS no. 9 bis, SEMA, Montrouge, 1964.

- [73] J. F. Shapiro. Mathematical programming models and methods for production planning and scheduling. In S. C. Graves, A. H. G. Rinnooy Kan, and P. H. Zipkin, editors, *Logistics of Production and Inventory*, volume 4 of *Handbooks in Operations Research and Management Science*, chapter 9, pages 371–444. North-Holland, 1993.
- [74] E. A. Silver. Operations research in inventory management: A review and critique. *Ops Res.*, 29:628–645, 1981.
- [75] Edward A. Silver and Rein Peterson. *Decision Systems for Inventory Management and Production Planning*. Wiley, second edition, 1985.
- [76] W. E. Smith. Various optimizers for single-stage production. *Naval Research Logistics Quarterly*, 3:59–66, 1956.
- [77] L. J. Thomas and J. O. McClain. An overview of production planning. In S. C. Graves, A. H. G. Rinnooy Kan, and P. H. Zipkin, editors, *Logistics of Production and Inventory*, volume 4 of *Handbooks in Operations Research and Management Science*, chapter 9, pages 333–370. North-Holland, 1993.
- [78] Arthur F. Veinott, Jr. Production planning with convex costs: A parametric study. *Mgmt Sci.*, 10:441–460, 1964.
- [79] H. M. Wagner. Research portfolio for inventory management and production planning systems. *Ops Res.*, 28:445–475, 1980.
- [80] Harvey M. Wagner. *Principles of Operations Research*. Prentice-Hall, 1969.
- [81] Harvey M. Wagner and Thomson M. Whitin. Dynamic version of the economic lot size model. *Mgmt Sci.*, 5:89–96, 1958.
- [82] G. O. Wesolowsky. The Weber problem: History and perspectives. *Location Science*, 1:5–23, 1993.