MODELS FOR OPTIMAL UTILIZATION OF PRODUCTION RESOURCES UNDER DEMAND SELECTION FLEXIBILITY

By

KEVIN MICHAEL TAAFFE

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2004

I dedicate this work to my family and to my future students.

ACKNOWLEDGMENTS

I would like to thank everyone who has provided words of support and encouragement during the past four years. Most importantly, I thank my wife, Mary, for providing unwavering support for my pursuit of this dream. She chose to sacrifice her own desires and needs to assist me by caring for our children and maintaining her business while I completed my degree. I do not know many people who could do that even once . . . and she does it all the time. Furthermore, she has always been my biggest emotional support. After being away from school for so long and working in industry for many years, I found it difficult to resume where I had left off 15 years ago with advanced math and theoretical research. Mary reminded me that it would be hard at times, but she always managed to calm me down and get me back on track. Simply said, I could not have accomplished this goal without her. She is the love of my life, and I will always love her from the bottom of my heart.

Looking back, I could not have asked for a better person to be my thesis advisor than Joe Geunes. He has been a great role model for my future career in academia, and I thank him for all of the experiences we have shared. He allowed me the space to think creatively, but he was always there when I needed help or guidance. Our families became very close over the years, and I hope we continue to stay close for many years to come.

There are many people who have touched my life in a special way since I arrived in Gainesville. From our neighbors who became like family, to my entire church family, and all of the friends I have met along the way, I can honestly say I have never felt such warmth on so many different levels. Every one of these people

has had an impact on who I am today, and I can say that they all have served as daily reminders as to what is truly important in life. It will be a sad farewell when we leave Gainesville, but I have developed many relationships that will never go away. For that I am eternally grateful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

MODELS FOR OPTIMAL UTILIZATION OF PRODUCTION RESOURCES
UNDER DEMAND SELECTION FLEXIBILITY

By

Kevin Michael Taaffe

August 2004

Chair: Joseph Geunes
Major Department: Industrial and Systems Engineering

Optimal demand selection applies to contexts in which an organization has some discretion in deciding the set of demands it will use its resources to satisfy. In such cases, the decision maker wishes to determine the set of downstream demands that provides the best match for its resource capabilities. This steps away from traditional streams of research that ignore the selection decision and assume all demand sources must be satisfied. We focus on developing new models and solution methods for problems that integrate demand selection with the planning and utilization of production resources, for both unlimited and limited production capacities. Capacity limits often restrict the total amount of demand that an organization can satisfy. When total demand for resources exceeds capacity limits, selecting the optimal subset of demand sources is a challenging optimization problem. Even in contexts where capacity limits are typically not a constraining factor, the problem remains difficult due to economies of scale in production and the attractiveness and timing of individual demands. Given a set of heterogeneous downstream demand sources, which may be deterministic or stochastic in nature, along with nonlinear capacity usage costs in volume, we propose models that

provide optimal demand source selections that achieve maximum profitability. In this dissertation, we specifically address demand selection flexibility for the applications areas of general production and inventory planning problems and airport ground holding problems.

CHAPTER 1
INTRODUCTION

This dissertation focuses broadly on models for optimal demand selection. Such models apply to contexts in which an organization has some discretion in deciding the set of demands it will use its resources to satisfy. In such cases, the decision maker wishes to determine the set of downstream demands that provides the best match for its resource capabilities. Capacity limits often restrict the total amount of demand that an organization can satisfy. When total demand for resources exceeds capacity limits, the decision maker must determine the best way in which to allocate its limited resources. Given a set of heterogeneous downstream demand sources, along with nonlinear capacity usage costs in volume, it is not a trivial problem to select the subset of demand sources that will provide the maximum profit to the firm. Even in contexts where capacity limits are typically not a constraining factor, economies of scale in production, combined with time-varying customer demand patterns from customers who have different reservation prices, make the problem of choosing the best set of demands to satisfy a challenging task. We focus on developing new models and solution methods for problems that fall in this general class of demand selection problems.

We explore two applications contexts within the class of demand selection problems:

- General production and inventory planning problems, and
- Airport operations ground holding problems.

We examine several types of production and inventory planning problems in Chapters 2, 3, and 4. In Chapter 2, we consider uncapacitated and capacitated versions of the single-stage, multi-period production planning problem for a producer who can select any number of orders, or demand sources, from a total set of potential demands. The problem has a finite horizon, and the producer has the discretion to choose when to produce and how much demand to satisfy in order to maximize profit. We define this new class of production planning problems with *order selection flexibility* and provide optimization-based modeling and solution methods for these problems. We provide a polynomial-time algorithm for solving the uncapacitated version of the problem, and we propose strong problem formulations and heuristic solution algorithms for several capacitated versions.

In Chapter 3, we extend our discussion of this single-stage production planning problem to address the importance of pricing. Firms that manufacture and sell products with price-elastic demand face the challenge of determining prices, and therefore demand volumes, that provide maximum profit to the firm. Nonlinearities in demand as a function of price and in production costs as a function of demand volumes create complexities in determining pricing strategies that maximize contribution to profit after production. Now, instead of directly selecting the desired demand quantities to satisfy, as shown in Chapter 2, we present a production planning model that implicitly decides, through pricing decisions, the demand levels the firm should satisfy in order to maximize contribution to profit. We present two polynomial-time solution approaches for these problems when production capacities are effectively unlimited, and show that these approaches apply across a range of applicable revenue and cost functions. We also describe a polynomial-time solution approach under time-invariant finite production capacities and piecewise-linear and concave revenue functions in the amount of demand satisfied.

These chapters together illustrate the importance of integrated demand and production planning decisions by enabling a producer to leverage production economies of scale to the greatest extent possible through matching the right amount of demand to production capabilities.

In Chapter 4, we introduce a stochastic version of the demand selection problem in a single-period setting, a problem that we refer to as the selective newsvendor problem. In this problem, a seller faces a long procurement lead time from an external supplier, and must simultaneously decide the markets in which it will sell its product along with the procurement quantity from the external supplier. For each selected market, the seller also determines the amount of marketing effort it will exert in the market, and this marketing effort influences the distribution of demand in the market (e.g., increased marketing effort implies higher expected demand in the market and also impacts the uncertainty of the market's demand). The goal is to choose the markets, advertising levels, and overall procurement quantity that maximizes the seller's expected profit in the selling season.

First, we present solution approaches for market selection decisions in which the marketing levels are fixed or pre-defined by the firm or supplier. We then extend the market selection approach to allow the firm to determine the best level of advertising to apply in each market selected. We illustrate this approach for both the unlimited and limited resources cases, and we evaluate multiple functional forms for the manner in which market demand levels respond to market advertising.

We conclude by presenting the airport operations ground holding problem in Chapter 5. This problem involves determining which flights destined for a given airport should be dispatched under uncertainty in future weather. In this context, flights destined for an airport constitute the (future) demand for arrival capacity, while the uncertainty in future weather leads to uncertain and dynamic capacity levels for receiving flights at the destination point. Accepting demands at a given destination can be very costly if the resulting capacity at flight arrival time is low (due to bad weather). Conversely, denying demands by holding flights at their origination points can also be quite costly, particularly if the resulting capacity at the scheduled flight arrival time is high (i.e., when previously predicted bad weather does not materialize at arrival time). The ground holding problem introduced in Chapter 5 addresses this critical issue of optimal flight arrival selection decisions under uncertainty.

The ground holding problem provides an excellent illustration of the benefits of using a stochastic over a deterministic approach in mathematical programming. We summarize these benefits within the chapter. The ground holding problem is also an interesting problem to study due to the number of different potential decision makers influencing the choice of flight demands to ground hold. Since the Federal Aviation Administration, local airport authorities, and individual airlines all have conflicting operational goals, we address new risk aversion models that allow multiple decision makers to achieve acceptable performance at the same time.

As shown in this dissertation, the demand selection problem can appear in many forms, and we provide a thorough discussion of the main focus areas within demand selection, such as deterministic demand vs. stochastic demand, unlimited resources vs. limited resources, and fixed pricing vs. variable pricing. These chapters provide a solid foundation for future research in demand source selection, and we also make several suggestions for research directions in the the concluding remarks.

CHAPTER 2
INTEGRATED ORDER SELECTION AND REQUIREMENTS PLANNING

2.1 Introduction

Firms that produce made-to-order goods often make critical order acceptance decisions prior to planning production for the orders they ultimately accept. These decisions require the firm's representatives (typically sales/marketing personnel in consultation with manufacturing management) to determine which among all customer orders the firm will satisfy. In certain contexts, such as those involving highly customized goods, the customer works closely with sales representatives to define an order's requirements and, based on these requirements, the status of the production system, and the priority of the order, the firm quotes a lead time for order fulfillment, which is then accepted or rejected by the customer (see Yano [85]). In other competitive settings, the customer's needs are more rigid and the customer's order must be fulfilled at a precise future time. The manufacturer can either commit to fulfilling the order at the time requested by the customer, or decline the order based on several factors, including the manufacturer's capacity to meet the order and the economic attractiveness of the order. These "order acceptance and denial" decisions are typically made prior to establishing future production plans and are most often made based on the collective judgment of sales, marketing, and manufacturing personnel, without the aid of the types of mathematical decision models typically used in the production planning decision process.

When the manufacturing organization is highly capacity constrained and customers have firm delivery date requirements, it is often necessary to satisfy a subset of customer orders and to deny an additional set of potentially profitable orders. In some contexts, the manufacturer can choose to employ a rationing scheme in an attempt to satisfy some fraction of each customer's demand (see Lee, Padmanabhan, and Whang [46]). In other settings, such a rationing strategy cannot be implemented; i.e., it may not be desirable or possible to substitute items ordered by one customer in order to satisfy another customer's demand. Thus, it may be necessary for the firm to deny certain customer orders (or parts of orders) so that the manufacturer can meet the customer-requested due dates for the orders it accepts. In contexts where capacity limits are non-binding, it is also not always clear that committing to a particular customer order is in the best interest of the firm, even if the unit price the customer will pay exceeds the variable production cost. This is evident in environments with significant fixed production costs.

Regardless of whether the operation is constrained or unconstrained by production capacity, assessing the profitability of an order in isolation, prior to production planning, leads to myopic decision rules that fail to consider the best set of actions from an overall profitability standpoint. The profitability of an order, when gauged solely by the revenues generated by the order and perceived customer priorities, neglects the impacts of important operations cost factors, such as the opportunity cost of manufacturing capacity consumed by the order, as well as economies of scale in production. Decisions on the collective set of orders the organization should accept can be a critical determinant of the firm's profitability.

Since Wagner and Whitin's [83] seminal paper addressed the basic economic lot-sizing problem (ELSP), numerous extensions and generalizations of this basic problem have followed, including extensions to incorporate backlogging (Zangwill [86]), serial system structures (Love [51]), and multistage assembly and general multistage structures (Afentakis, Gavish, and Karmarkar [2], and Afentakis and Gavish [1]). Intensive research on the capacitated version of the dynamic requirements planning problem began in the 1970's (see Florian and Klein [28], Baker, Dixon, Magazine, and Silver [5], and Florian, Lenstra, and Rinnooy Kan [29]), and has received increased attention recently as a result of the application of strong valid inequalities that enable faster solution of these difficult problems (e.g., Barany, Van Roy, and Wolsey [10], Pochet [61], and Leung, Magnanti, and Vachani [47]). Lee and Nahmias [45], Shapiro [70], and Baker [4] provide excellent overall analyses of the generalizations and solution approaches for dynamic requirements planning problems, including various heuristic approaches that have proven effective for the capacitated version of the problem.

With a few notable exceptions that we later discuss, this past research on dynamic requirements planning problems nearly always assumes that demands are pre-specified by time period and that all demands must be completely filled at the time they occur (or after they occur in models that permit backlogging). In contrast, we consider a requirements planning model that implicitly determines the best demand levels to satisfy in order to maximize contribution to profit. While the uncapacitated version is solvable in polynomial time, as we later discuss, the capacitated version is NP-Hard and therefore requires customized heuristic solution approaches. We propose strong LP formulations of the capacitated version, which often allows solving general capacitated instances via branch-and-bound in reasonable computing time. For those problems that cannot be solved via branch-and-bound in reasonable time, we provide a set of three effective heuristic solution methods. Computational test results indicate that the proposed solution methods for the general capacitated version of the problem are very effective, producing solutions within 0.67% of optimality, on average, for a broad set of 3,240 randomly generated problem instances.

Loparic, Pochet, and Wolsey [50] recently considered a related problem in which a producer wishes to maximize net profit from sales of a single item and does not have to satisfy all outstanding demand in every period. Their model assumes that only one demand source exists in every period, and that the revenue from this demand source is proportional to the volume of demand satisfied. The "order selection" interpretation of the model we present, on the other hand, allows the firm to consider any number of orders (or demand sources) in each period, each with a unique associated per unit revenue (i.e., we allow for customers with different reservation prices). In this respect, their model represents a single-order special case of one of the models we propose. More recently, Lee, Çetinkaya, and Wagelmans [43] considered contexts in which demands can be met either earlier (through early production and delivery) or later (through backlogging) than specified without penalty, provided that demand is satisfied within certain *demand time windows* for the uncapacitated, single-stage lot sizing problem. Their model still assumes ultimately, however, that all pre-specified demands must be filled during the planning horizon.

The remainder of this chapter is organized as follows. Section 2.2 presents a formal definition and mixed integer programming formulation of the general production planning problem with order selection flexibility. We then present a solution approach for the uncapacitated version of the problem that generalizes the Wagner-Whitin [83] shortest path solution method for single-stage dynamic requirements planning problems. In Section 2.3 we consider various mixed integer programming formulations of the capacity constrained problem, along with the advantages and disadvantages of each formulation strategy. We also provide several heuristic solution approaches for each of the capacitated problem instances. Section 2.4 then provides a summary of a set of computational tests used to gauge the effectiveness of the formulation strategies and heuristic solution methods described in Section 2.3.

2.2 Order Selection Problem Definition and Formulation

Consider a producer who manufactures a good to meet a set of outstanding orders over a finite number of time periods, T. Producing the good in any time period t requires a production setup at a cost S_t and each unit costs an additional p_t to manufacture. We let $M(t)$ denote the set of all orders that request delivery in period t (we assume zero delivery lead time for ease of exposition; the model easily extends to a constant delivery lead time without loss of generality), and let m denote an index for orders. The manufacturer has a capacity to produce C_t units in period t, $t = 1, \ldots, T$. We assume that that there is no planned backlogging[1] (i.e., no shortages are permitted) and that items can be held in inventory at a cost of h_t per unit remaining at the end of period t. Let d_{mt} denote the quantity of the good requested by order m for period t delivery, for which the customer will pay r_{mt} per unit, and suppose the producer is free to choose any quantity between zero and d_{mt} in satisfying order m in period t (i.e., rationing is possible, and the customer will take as much of the good as the supplier can provide, up to d_{mt}). The producer thus has the flexibility to decide which orders it will choose to satisfy in each period and the quantity of demand it will satisfy for each order. If the producer finds it unprofitable to satisfy a certain order in a period, it can choose to reject the order at the beginning of the planning horizon. The manufacturer incurs a fixed shipping cost for delivering order m in period t equal to F_{mt} (any variable shipping cost can be subtracted from the revenue term, r_{mt}, without loss of generality). The producer, therefore, wishes to maximize net profit over a T-period horizon, defined as the total revenue from orders satisfied minus total production

[1] Extending our models and solution approaches to allow backlogging at a per unit per period backlogging cost is fairly straightforward. We have chosen to omit the details of this extension.

(setup + variable), holding, and delivery costs incurred over the horizon. To formulate this problem we define the following decision variables:

x_t = Number of units produced in period t,

$$y_t = \begin{cases} 1, & \text{if we setup for production in period } t, \\ 0, & \text{otherwise,} \end{cases}$$

I_t = Producer's inventory remaining at the end of period t,

v_{mt} = Proportion of order m satisfied in period t,

$$z_{mt} = \begin{cases} 1, & \text{if we satisfy any positive fraction of order } m \text{ in period } t, \\ 0, & \text{otherwise.} \end{cases}$$

We formulate the Capacitated Order Selection Problem (OSP) as follows.

[OSP]

$$\text{maximize:} \quad \sum_{t=1}^{T} \left(\sum_{m \in M(t)} (r_{mt} d_{mt} v_{mt} - F_{mt} z_{mt}) - S_t y_t - p_t x_t - h_t I_t \right) \tag{2.1}$$

$$\text{subject to:} \quad I_{t-1} + x_t = \sum_{m \in M(t)} d_{mt} v_{mt} + I_t \quad t = 1, \ldots, T, \tag{2.2}$$

$$0 \leq x_t \leq C_t y_t \quad t = 1, \ldots, T, \tag{2.3}$$

$$0 \leq v_{mt} \leq z_{mt} \quad t = 1, ..., T, m \in M(t), \tag{2.4}$$

$$I_0 = 0, I_t \geq 0 \quad t = 1, \ldots, T, \tag{2.5}$$

$$y_t, z_{mt} \in \{0, 1\} \quad t = 1, ..., T, m \in M(t). \tag{2.6}$$

The objective function (2.1) maximizes net profit, defined as total revenue less fixed shipping and total production and inventory holding costs. Constraint set (2.2) represents inventory balance constraints, while constraint set (2.3) ensures that no production occurs in period t if we do not perform a production setup in the period. If a setup occurs in period t, the production quantity is constrained by the production capacity, C_t. Constraint set (2.4) encodes our assumption regarding the producer's ability to satisfy any proportion of order m up to the amount d_{mt}, while (2.5) and (2.6) provide nonnegativity and integrality restrictions on variables. Observe that we can force any order selection (z_{mt}) variable to one if qualitative and/or strategic concerns (e.g., market share goals) require satisfying an order regardless of its profitability.

2.2.1 The Uncapacitated Order Selection Problem

If a setup occurs in period t, the production quantity is unconstrained by setting C_t equal to a large number. Alternatively, we can set C_t equal to $\sum_{\tau=t}^{T} \sum_{m \in M(\tau)} d_{m\tau}$ without loss of generality, since this is the maximum amount of demand that could be satisfied by period t production (in the absence of backlogging). We denote the resulting uncapacitated order selection problem as [UOSP].

Problem [UOSP] can be represented as a fixed-charge network flow problem as illustrated by the example in Figure 2–1, where $T = 4$ and $M(t) = 2$ for $t = 1, \ldots, T$. The network contains three types of arcs: *production arcs*, *inventory arcs*,

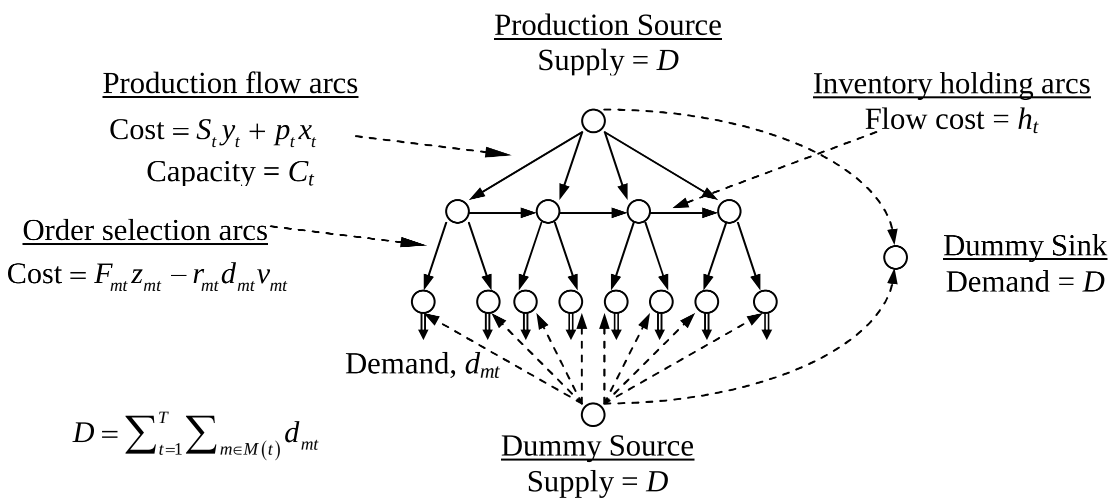


Figure 2–1: Fixed charge network flow representation of UOSP.

and *order selection arcs*. The dummy source node implies that it is not necessary to satisfy all demand—the dummy source can supply the entire demand over the horizon if necessary. We also add a dummy sink node that can receive flow from both the production source and the dummy source. Flow on a production arc implies that a setup occurs in that period, while flow on an order selection arc implies that we satisfy at least some of that order in the corresponding period. Since the flow cost on each arc is concave or linear (and hence also concave), the objective function (2.1) is convex and [UOSP] maximizes a convex function over a polyhedron for a given y, z. This implies that an optimal extreme point solution exists for [UOSP]. Since the problem is a network flow problem, this implies that an optimal *spanning tree* solution exists (see Ahuja et al. [3]), in which the subgraph induced by the arcs with positive flow in a solution forms a spanning tree. We exploit this spanning tree property to derive certain properties of optimal solutions to [UOSP]. Note that [UOSP] generalizes the Wagner-Whitin single-stage requirements planning problem under dynamic demand, whose solution approach we will extend to solve our order selection problem.

2.2.2 Solution Properties and Shortest Path Approach for UOSP

The existence of an optimal spanning tree solution for [UOSP] implies the following property:

All-or-nothing order satisfaction property: Given the choice to satisfy any quantity of demand less than or equal to d_{mt} for order m in period t, an optimal solution exists with either v_{mt} equal to 0 or 1 for all m and t; i.e., for each order-period combination (m, t) the producer either provides d_{mt} units or none at all.

We next consider how to extend the Wagner-Whitin dynamic programming solution method for solving UOSP. Their dynamic programming solution method can be equivalently posed as a shortest path problem on a graph containing $T + 1$ nodes (see Figure 2–2). Note that this method relies on the existence of an optimal *Zero-Inventory Ordering* (ZIO) policy in which a setup only occurs in period t if we hold no inventory at the end of period $t - 1$ (the validity of this property can also be shown to hold for [UOSP] as a result of the spanning tree property of the equivalent fixed charge network representation of [UOSP]). Since the Wagner-Whitin approach minimizes total cost, each arc (t, t') in the graph is assigned a cost, $c(t, t')$, where $c(t, t')$ equals the setup cost in period t plus the variable production and holding costs incurred for satisfying all demand in periods $t, t+1, \ldots, t'-1$ using only the setup in period t. This approach ensures that a path exists in the shortest path network for every feasible combination of setups and that the cost of a path corresponds to the minimum cost incurred in using the setups to satisfy all demand.

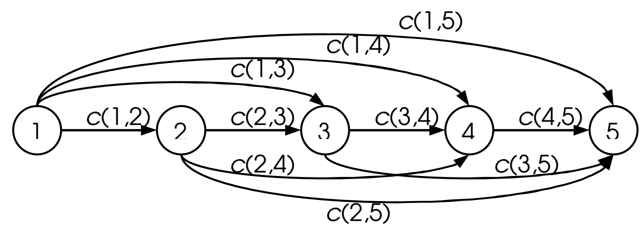


Figure 2–2: Shortest path network structure for UOSP.

Since the ZIO property also holds for the uncapacitated order selection problem, we can solve the UOSP problem using a shortest path graph containing the same structure as that used for solving the Wagner-Whitin problem. The arc length calculation for UOSP, however, requires a new approach. The order selection problem seeks to maximize net profit and so we interpret arc lengths in terms of net contribution to profit and seek the longest path in the graph. The method used for arc length calculation proceeds as follows. We interpret the length of arc (t, t') as the maximum profit possible from satisfying orders in periods $t, \ldots, t'-1$ assuming that the only setup available to satisfy demand in these periods must occur in period t, if at all. Suppose we choose to perform the setup in period t and incur its corresponding cost, S_t. To offset the cost of this setup we will satisfy the demand for order m in period t if and only if

$$(r_{mt} - p_t)d_{mt} \geq F_{mt}; \tag{2.7}$$

i.e., if the net revenue generated from order m is at least as great as the fixed delivery cost for order m in period t. Similarly, for all periods τ such that $t < \tau < t'$, we will satisfy order m in period τ if and only if

$$(r_{m\tau} - p_t - \sum\nolimits_{k=t}^{\tau-1} h_k) d_{m\tau} \geq F_{m\tau}; \tag{2.8}$$

i.e., if the net revenue from order m in period τ, less any holding costs incurred from period t to period τ, exceeds the fixed delivery cost for order m in period τ. Let $O_t(t)$ denote the set of orders in period t such that $r_{mt} \geq p_t$, and let $O_t(\tau)$ denote the set of orders in period τ such that (2.8) holds for $\tau > t$. Then the maximum profit possible if we do a setup in period t and use that setup to satisfy demands in periods $t, \ldots, t'-1$, which we denote by $\mathrm{MP}_S(t,t')$, is given by

$$\begin{aligned} \mathrm{MP}_S(t,t') &= \sum\nolimits_{\tau=t}^{t'-1} \sum\nolimits_{m \in O_t(\tau)} ((r_{m\tau} - p_t)\, d_{m\tau} - F_{m\tau}) \\ &\quad - \sum\nolimits_{\tau=t}^{t'-2} h_\tau \left(\sum\nolimits_{k=\tau+1}^{t'-1} \sum\nolimits_{m \in O_t(\tau)} d_{mk} \right) - S_t. \end{aligned} \tag{2.9}$$

If $\mathrm{MP}_S(t,t')$ is greater than 0 we set the length of arc (t,t') equal to $\mathrm{MP}_S(t,t')$; otherwise we set the length of arc (t,t') equal to zero and assume no setup occurs in period t if the optimal solution (the longest path in the graph) traverses arc (t,t'). After finding the longest path in the graph we can determine which orders to satisfy in each period by checking the elements of the sets $O_t(t)$ and $O_t(\tau)$ for all arcs (t,t') contained in the longest path.

Letting $\bar{m} = \max\limits_{t=1,\ldots,T} \{|M(t)|\}$, the total computational effort of arc cost calculations is bounded by $\mathcal{O}\left(\bar{m}T^2\right)$ and the shortest path calculation is no worse than $\mathcal{O}(T^2)$, so the worst case complexity of this algorithm is bounded by $\mathcal{O}\left(\bar{m}T^2\right)$. Recent work on the uncapacitated lot sizing problem (e.g., Federgruen and Tzur [24] and Wagelmans, van Hoesel, and Kolen [82]) has reduced the complexity of this problem from the $\mathcal{O}(T^2)$ bound to $\mathcal{O}(T \log T)$ (or even $\mathcal{O}(T)$ in certain special cases). These approaches, however, rely on an important property that holds for the ELSP, and this property states that the cumulative demand satisfied as we increase the number of periods in a problem instance is nondecreasing. That is, for the ELSP, the total demand satisfied in a two-period problem is at least as great as that satisfied in a one-period problem (where demand in period 1 is the same in both problem instances). Interestingly, we can show that this property *does not hold* in general for the UOSP problem. Assuming that the holding cost in every period equals zero, we introduce the following data for a three-period problem:

Table 2–1: Counterexample illustrating decreasing cumulative demand satisfaction.

Period	**Setup Cost**	**Production Cost**	**Demand**	**Unit Revenue**
1	\$50	\$1.50	20	\$1.80
1	\$50	\$1.25	20	\$4.00
1	\$1,000	\$1.20	10	\$10.00

Consider the period 1 problem alone. If we setup and satisfy all 20 units of demand, the revenue equals \$36, while the setup plus variable production cost equals \$80. Thus we satisfy zero units of demand in the period 1 problem. In the period 1 + period 2 problem, an optimal solution satisfies 20 units in periods 1 and 2, using the setup in period 1, for a total of 40 units of demand satisfied. Finally, for the problem containing periods 1, 2, and 3, it is optimal to setup in period 2 and satisfy the 30 units of demand in periods 2 and 3. This example illustrates why we cannot apply methods previously developed to reduce the complexity of ELSP to $\mathcal{O}(T \log T)$ in an effort to reduce the complexity of our algorithm to, say, $\mathcal{O}(\bar{m}T \log T)$, since cumulative demand satisfied is not necessarily nondecreasing for the UOSP problem (note that it is possible to provide a similar example under which cumulative demand satisfied from period t to T is not necessarily nondecreasing as we move backwards in time, or, as t decreases).

2.3 OSP Models - Limited Production Capacity

We now turn our attention to the capacitated version of our model. We investigate not only the OSP model as formulated above, but also certain special cases and restrictions of this model that are of both practical and theoretical interest. In particular, we consider the special case in which no fixed delivery charges exist (i.e., the case in which all fixed delivery charge (F_{mt}) parameters equal zero). We denote this version of the model as the OSP-NDC. We also explore contexts in which customers do not permit partial demand satisfaction. This is a restricted version of the OSP in which the continuous v_{mt} variables must equal the binary delivery-charge forcing (z_{mt}) variable values, and can therefore be substituted out of the formulation; let OSP-AND denote this version of the model (where AND implies all-or-nothing demand satisfaction). Observe that for the OSP-AND model we can introduce a new revenue parameter $R_{mt} \equiv r_{mt}d_{mt}$, where the total revenue from order m in period t must now equal $R_{mt}z_{mt}$. Table 2–2 defines our notation with respect to the different variants of the OSP problem.

Table 2–2: Classification of model special cases and restrictions.

Model	**Fixed Delivery Charges Exist**	**Partial Order Satisfaction Allowed**
OSP	Y	Y
OSP-NDC	N	Y
OSP-AND	U	N

Y = Yes; N = No.
U: Model and solution approaches unaffected by this assumption.

We distinguish between these model variants not only because they broaden the model's applicability to different contexts, but also because they can substantially affect the model's formulation size and complexity, as we next briefly discuss. Let $M_\Sigma = \sum_{t=1}^{T} |M(t)|$ denote the total number of customer orders over the T-period horizon, where $|M(t)|$ is the cardinality of the set $M(t)$. Note that formulation [OSP] contains $M_\Sigma + T$ binary variables and $M_\Sigma + 2T$ constraints, not including the binary and nonnegativity constraints. The OSP-NDC model, on the other hand, in which $F_{mt} = 0$ for all order-period (m, t) combinations, allows us to replace each z_{mt} variable on the right-hand-side of constraint set (2.4) with a 1, and eliminate these variables from the formulation. The OSP-NDC model contains only

T binary variables and therefore requires M_Σ fewer binary variables than [OSP], a significant reduction in problem size and complexity. In the OSP-AND model, customers do not allow partial demand satisfaction, and so we require $v_{mt} = z_{mt}$ for all order-period (m,t) combinations; we can therefore eliminate the continuous v_{mt} variables from the formulation. While the OSP-AND, like the OSP, contains $M_\Sigma + T$ binary variables, it requires M_Σ fewer total variables than [OSP] as a result of eliminating the v_{mt} variables. Table 2–3 summarizes the size of each of these variants of the OSP with respect to the number of constraints, binary variables, and total variables.

Table 2–3: Problem size comparsion for capacitated versions of the OSP.

	OSP	**OSP-NDC**	**OSP-AND**
Number of Constraints[a]	$M_\Sigma + 2T$	$M_\Sigma + 2T$	$2T$
Number of Binary Variables	$M_\Sigma + T$	T	$M_\Sigma + T$
Number of Total Variables	$2M_\Sigma + 3T$	$M_\Sigma + 3T$	$M_\Sigma + 3T$

[a] Binary restriction and nonnegativity constraints are not included.

Based on the information in this table, we would expect the OSP and OSP-AND to be substantially more difficult to solve than the OSP-NDC. As we will show in Section 2.4, the OSP-AND actually requires the greatest amount of computation time on average, while the OSP-NDC requires the least.

Note that the OSP-AND is indifferent to whether fixed delivery charges exist, since we can simply reduce the net revenue parameter, $R_{mt} \equiv r_{mt}d_{mt}$, by the fixed delivery-charge value F_{mt}, without loss of generality. In the OSP-AND then, the net revenue received from an order equals $R_{mt}z_{mt}$, and we thus interpret the z_{mt} variables as binary "order selection" variables. In contrast, in the OSP, the purpose of the binary z_{mt} variables is to force us to incur the fixed delivery charge if we satisfy any fraction of order m in period t. In this model we therefore interpret the z_{mt} variables as fixed delivery-charge forcing variables, since their objective function coefficients are fixed delivery cost terms rather than net revenue terms, as in the OSP-AND. Note also that since both the OSP-NDC and the OSP-AND require only one set of order selection variables (the continuous v_{mt} variables for the OSP-NDC and the binary z_{mt} variables for the OSP-AND), their linear programming relaxation formulations will be identical (since relaxing the binary z_{mt} variables is equivalent to setting $z_{mt} = v_{mt}$). The OSP linear programming relaxation formulation, on the other hand, explicitly requires both the v_{mt} and z_{mt} variables, resulting in a larger LP relaxation formulation than that for the OSP-NDC and the OSP-AND. These distinctions will play an important role in interpreting the difference in our ability to obtain strong upper bounds on the optimal solution value for the OSP and the OSP-AND in Section 2.4.3. We next discuss solution methods for the OSP and the problem variants we have presented.

2.3.1 OSP Solution Methods

To solve the OSP, we must decide which orders to select and, among the selected orders, how much of the order we will satisfy while obeying capacity limits. We can show that this problem is NP-Hard through a reduction from the capacitated lot-sizing problem as follows. If we consider the special case of the OSP in which $\sum_{t=1}^{j} C_t \geq \sum_{t=1}^{j} \sum_{m \in M(t)} d_{mt}$ for $j = 1, \ldots, T$ (which implies that satisfying all orders is feasible) and $\min\limits_{t=1,\ldots,T, m \in M(t)} \{r_{mt}\} \geq \max\limits_{t=1,\ldots,T} \{S_t\} + \max\limits_{t=1,\ldots,T} \{p_t\} + \sum_{t=1}^{T-1} h_t$ (which implies that it is profitable to satisfy all orders in every period), then total revenue is fixed and the problem is equivalent to a capacitated lot-sizing problem, which is an NP-Hard optimization problem (see Florian and Klein [28]).

Given that the OSP is NP-Hard, we would like to find an efficient method for obtaining good solutions for this problem. As our computational test results in Section 2.4 later show, we were able to find optimal solutions using branch-and-bound for many of our randomly generated test instances. While this indicates that the majority of problem instances we considered were not terribly difficult to solve, there were still many instances in which an optimal solution could not be found in reasonable computing time. Based on our computational test experience in effectively solving problem instances via branch-and-bound using the CPLEX 6.6 solver, we focus on strong LP relaxations for the OSP that provide quality upper bounds on optimal net profit quickly, and often enable solution via branch-and-bound in acceptable computing time. For those problems that cannot be solved via branch-and-bound, we employ several customized heuristic methods, which we discuss in Section 2.3.3. Before we discuss the heuristics used to obtain lower bounds for the OSP, we first present our reformulation strategy, which helps to substantially improve the upper bound provided by the linear programming relaxation of the OSP.

2.3.2 Strengthening the OSP Formulation

This section presents an approach for providing good upper bounds on the optimal net profit for the OSP. In particular, we describe two LP relaxations for the OSP, both of which differ from the LP relaxation obtained by simply relaxing the binary restrictions of [OSP] (constraint set (2.6)) in Section 2.2. We will refer to this simple LP relaxation of [OSP] as OSP-LP, to distinguish this relaxation from the two LP relaxation approaches we provide in this section.

The two LP relaxation formulations we next consider are based on a reformulation strategy developed for the UOSP. In Chapter 3, we will present a "tight" formulation of a similar problem to the UOSP, for which we show that the optimal LP relaxation solution value equals the optimal (mixed integer) UOSP solution value. We discuss this reformulation strategy in greater detail by first providing a tight linear programming relaxation for the UOSP. We first note that for the UOSP, an optimal solution exists such that we never satisfy part of an order; i.e., v_{mt} equals either 0 or 1. Thus we can substitute the v_{mt} variables out of [OSP] by setting $v_{mt} = z_{mt}$ for all t and $m \in M(t)$.

Next observe that since $I_t = \sum_{j=1}^{t} x_j - \sum_{j=1}^{t} \sum_{m \in M(j)} d_{mj} z_{mj}$, we can eliminate the inventory variables from [OSP] via substitution. After introducing a new variable production and holding cost parameter, c_t, where $c_t \equiv p_t + \sum_{j=t}^{T} h_j$, the objective function of the UOSP can be rewritten as

$$\text{maximize} \sum_{j=1}^{T} \sum_{m \in M(j)} R_{mj} z_{mj} + \sum_{t=1}^{T} h_t \left(\sum_{j=1}^{T} \sum_{m \in M(j)} d_{mj} z_{mj} \right) - \sum_{t=1}^{T} (S_t y_t + c_t x_t) \quad (2.10)$$

We next define ρ_{mt} as an adjusted revenue parameter for order m in period t, where $\rho_{mt} = \sum_{j=t}^{T} h_j + R_{mt}$. Our reformulation procedure requires capturing the exact amount of production in each period allocated to every order. We thus define x_{mtj} as the number of units produced in period t used to satisfy order m in period j, for $j \geq t$, and replace each x_t with $\sum_{j=t}^{T} \sum_{m \in M(j)} x_{mtj}$. The following formulation provides the "strong" linear programming relaxation of the UOSP.

[UOSP′]

$$\text{maximize:} \quad \sum_{j=1}^{T} \sum_{m \in M(j)} \rho_{mj} d_{mj} z_{mj} - \sum_{t=1}^{T} \left(S_t y_t + c_t \sum_{j=t}^{T} \sum_{m \in M(j)} x_{mtj} \right) \qquad (2.11)$$

subject to:

$$\sum_{t=1}^{j} x_{mtj} - d_{mj} z_{mj} = 0 \qquad j = 1, \ldots, T, m \in M(j), \quad (2.12)$$

$$\left(\sum_{m \in M(j)} d_{mj} \right) y_t - \sum_{m \in M(j)} x_{mtj} \geq 0 \quad t = 1, \ldots, T, j = t, \ldots, T, \quad (2.13)$$

$$-z_{mj} \geq -1 \qquad j = 1, \ldots, T, m \in M(j), \quad (2.14)$$

$$y_t, x_{mtj}, z_{mj} \geq 0 \qquad t = 1, \ldots, T,$$

$$j = t, \ldots, T, m \in M(j). \quad (2.15)$$

Note that since a positive cost exists for setups, we can show that the constraint $y_t \leq 1$ is unnecessary in the above relaxation, and so we omit this constraint from the relaxation formulation. It is straightforward to show that [UOSP′] with the additional requirements that all z_{mj} and y_t are binary variables is equivalent to our [OSP] when production capacities are infinite. (We will also show in Chapter 3 that by disaggregating the setup forcing constraints (2.13), the resulting formulation has zero integrality gap through a dual solution approach.)

To obtain the LP relaxation for the OSP (when production capacities are finite), we add finite capacity constraints to [UOSP′] by forcing the sum of x_{mtj} over all $j \geq t$ and all $m \in M(j)$ to be less than the production capacity C_t in period t. That is, we can add the following constraint set to [UOSP] to obtain an equivalent LP relaxation for the OSP:

$$\sum_{j=t}^{T} \sum_{m \in M(j)} x_{mtj} \leq C_t \qquad t = 1, \ldots, T. \tag{2.16}$$

Note that this LP relaxation approach is valid for all three variants of the OSP: the general OSP, the OSP-NDC, and the OSP-AND. Observe that the above constraint can be strengthened by multiplying the right-hand-side by the setup forcing variable y_t. To see how this strengthens the formulation, note that constraint set (2.13) in [UOSP′] implies that

$$\sum_{j=t}^{T} \sum_{m \in M(j)} x_{mtj} \leq \sum_{j=t}^{T} \left(\sum_{m \in M(j)} d_{mj} \right) y_t \qquad t = 1, \ldots, T.$$

To streamline our notation, we define the following. Let $X_{tT} = \sum_{j=t}^{T} \sum_{m \in M(j)} x_{mtj}$ and $D_{tT} = \sum_{j=t}^{T} \left(\sum_{m \in M(j)} d_{mj} \right)$ for $t = 1, \ldots, T$ denote aggregated production variables and order amounts, respectively. Constraint set (2.16) can be rewritten as

$$X_{tT} \leq C_t \qquad t = 1, \ldots, T,$$

and the aggregated demand forcing constraints (2.13) can now be written as $X_{tT} \leq D_{tT} \cdot y_t$. If we do not multiply the right-hand-side of capacity constraint set (2.16) by the forcing variable y_t, the formulation allows solutions, for example, such that $X_{tT} = C_t$ for some t, while X_{tT} equals only a fraction of D_{tT}. In such a case, the forcing variable y_t takes the fractional value $\frac{X_{tT}}{D_{tT}}$, and we only absorb a fraction of the setup cost in period t. Multiplying the right-hand-side of (2.16) by y_t, on the other hand, would force $y_t = \frac{X_{tT}}{C_t} = 1$ in such a case, leading to an improved upper bound on the optimal solution value. We can therefore strengthen the LP relaxation solution that results from adding constraint set (2.16) by instead using the following capacity forcing constraints.

$$X_{tT} \leq \min\{C_t, D_{tT}\} \cdot y_t \qquad t = 1, \ldots, T. \tag{2.17}$$

Note that in the capacitated case we now explicitly require stating the $y_t \leq 1$ constraints in the LP relaxation, since it may otherwise be profitable to violate production capacity in order to satisfy additional orders. We refer to the resulting LP relaxation with these aggregated setup forcing constraints as the [ASF] formulation, which we formulate as follows.

[ASF]

$$\text{maximize:} \quad \sum_{j=1}^{T} \sum_{m \in M(j)} \rho_{mj} d_{mj} z_{mj} - \sum_{t=1}^{T} \left(S_t y_t + c_t \sum_{j=t}^{T} \sum_{m \in M(j)} x_{mtj} \right)$$

subject to: Constraints (2.12 – 2.15,2.17)

$$y_t \leq 1 \qquad t = 1, \ldots, T. \tag{2.18}$$

We can further strengthen the LP relaxation formulation by disaggregating the demand forcing constraints (2.13) (see Erlenkotter [23], who uses this strategy for the uncapacitated facility location problem). This will force y_t to be at least as great as the maximum value of $\frac{x_{mtj}}{d_{mj}}$ for all $j = t, \ldots, T$ and $m \in M(j)$. The resulting Disaggregated Setup Forcing (DASF) LP relaxation is formulated as follows.

[DASF]

$$\text{maximize:} \quad \sum_{j=1}^{T} \sum_{m \in M(j)} \rho_{mj} d_{mj} z_{mj} - \sum_{t=1}^{T} \left(S_t y_t + c_t \sum_{j=t}^{T} \sum_{m \in M(j)} x_{mtj} \right)$$

subject to: Constraints (2.12, 2.14, 2.15,2.17, 2.18)

$$x_{mtj} \leq d_{mj} y_t \qquad t = 1, \ldots, T, \quad j = t, \ldots, T, \quad m \in M(j). \tag{2.19}$$

Each of the LP relaxations we have described provides some value in solving the capacitated versions of the OSP. Both the OSP-LP and ASF relaxations can be solved very quickly, and they frequently yield high quality solutions. The DASF relaxation further improves the upper bound on the optimal solution value. But as the problem size grows (i.e., the number of orders per period or the number of time periods increases), [DASF] becomes intractable, even via standard linear programming solvers. We present results for each of these relaxation approaches in Section 2.4. Before doing this, however, we next discuss methods for determining good feasible solutions, and therefore lower bounds, for the OSP via several customized heuristic solution procedures.

2.3.3 Heuristic Solution Approaches for OSP

While the methods discussed in the previous subsection often provide strong upper bounds on the optimal solution value for the OSP (and its variants), we cannot guarantee the ability to solve this problem in reasonable computing time using branch-and-bound due to the complexity of the problem. We next discuss three heuristic solution approaches that allow us to quickly generate feasible solutions for OSP. As our results in Section 2.4 report, using a composite solution procedure that selects the best solution among those generated by the three heuristic solution approaches provided feasible solutions with objective function values, on average, within 0.67% of the optimal solution value. We describe our three heuristic solution approaches in the following three subsections.

2.3.3.1 Lagrangian Relaxation Based Heuristic

Lagrangian relaxation (Geoffrion [32]) is often used for mixed integer programming problems to obtain stronger upper bounds (for maximization problems) than provided by the LP relaxation. As we discussed in Section 2.3.2, our strengthened linear programming formulations typically provide very good upper bounds on the optimal solution value of the OSP. Moreover, as we later discuss, our choice of relaxation results in a Lagrangian subproblem for which we can find an optimal extreme point solution equivalent to the solution found for our LP relaxation. This implies that the upper bound provided by our Lagrangian relaxation scheme will not provide better bounds than our LP relaxation. Our purpose for implementing a Lagrangian relaxation heuristic, therefore, is strictly to obtain good feasible solutions using a Lagrangian-based heuristic. Because of this we omit certain details of the Lagrangian relaxation algorithm and implementation, and describe only the essential elements of the general relaxation scheme and how we obtain a heuristic solution at each iteration of the Lagrangian algorithm.

Under our Lagrangian relaxation scheme, we add (redundant) constraints of the form $x_t \leq My_t$, $t = 1, \ldots, T$ to [OSP] (where M is some large number), eliminate the forcing variable y_t from the right-hand side of the capacity/setup forcing constraints (2.3), and then relax the resulting modified capacity constraint (without the y_t multiplier on the right-hand side) in each period. The Lagrangian relaxation subproblem is then simply an uncapacitated OSP (or UOSP) problem. Although the Lagrangian multipliers introduce the possibility of negative unit production costs in the Lagrangian subproblem, we retain the convexity of the objective function, and all properties necessary for solving the UOSP problem via a shortest path network approach still hold (see Section 2.2.2). We can therefore solve the Lagrangian subproblems in polynomial time. Because we have a tight formulation of the UOSP (as we will prove in Chapter 3), this implies that the

Lagrangian solution will not provide better upper bounds than the LP relaxation. We do, however, use the solution of the Lagrangian subproblem at each iteration of a subgradient optimization algorithm (see Fisher [26]) as a starting point for heuristically generating a feasible solution, which serves as a candidate lower bound on the optimal solution value for OSP.

Observe that the subproblem solution from this relaxation will satisfy all constraints of the OSP except for the relaxed capacity constraints (2.3). We therefore call a feasible solution generator (FSG) at each step of the subgradient algorithm, which can take any starting capacity-infeasible (but otherwise feasible) solution and generate a capacity-feasible solution. (We also use this FSG in our other heuristic solution schemes, as we later describe.) The FSG works in three main phases. Phase I first considers performing additional production setups (beyond those prescribed by the starting solution) to try to accommodate the desired production levels and order selection decisions provided in the starting solution, while obeying production capacity limits. That is, we consider shifting production from periods in which capacities are violated to periods in which no setup was originally planned in the starting solution. It is possible, however, that we still violate capacity limits after Phase I, since we do not eliminate any order selection decisions in Phase I.

In Phase II, after determining which periods will have setups in Phase I, we consider those setup periods in which production still exceeds capacity and, for each such setup period, index the orders satisfied from production in the setup period in nondecreasing order of contribution to profit. For each period with violated capacity, in increasing profitability index order, we shift orders to an earlier production setup period, if the order remains profitable and such an earlier production setup period exists with enough capacity to accommodate the order. Otherwise we eliminate the order from consideration. If removing the order from the setup period will leave excess capacity in the setup period under consideration, we consider shifting only part of the order to a prior production period; we also consider eliminating only part of the order when customers do not require all-or-nothing order satisfaction. This process is continued for each setup period in which production capacity is violated until total production in the period satisfies the production capacity limit. Following this second phase of the algorithm, we will have generated a capacity-feasible solution. In the third and final phase, we scan all production periods for available capacity and assign additional profitable orders that have not yet been selected to any excess capacity if possible. The Chapter Appendix in Section 2.6 contains a detailed description of the FSG algorithm.

2.3.3.2 Greatest Unit Profit Heuristic

Our next heuristic solution procedure is motivated by an approach taken in several well-known heuristic solution approaches for the ELSP. In particular, we use a similar "myopic" approach to those used in the Silver-Meal [72] and Least Unit Cost (see Nahmias [56]) heuristics. These heuristics proceed by considering an initial setup period, and then determining the number of consecutive period demands (beginning with the initial setup period) that produce the lowest cost per period (Silver-Meal) or per unit (Least Unit Cost) when allocated to production in the setup period. The next period considered for a setup is the one immediately following the last demand period assigned to the prior setup; the heuristics proceed until all demand has been allocated to some setup period. Our approach differs from these approaches in the following respects. Since we are concerned with the profit from orders, we take a greatest profit rather than a lowest cost approach. We also allow for accepting or rejecting various orders, which implies that we need only consider those orders that are profitable when assigning orders to a production period. Moreover, we can choose not to perform a setup if no selection of orders produces a positive profit when allocated to the setup period. Finally, we apply our "greatest unit profit" heuristic in a capacitated setting, whereas a modification of the Silver-Meal and Least Unit Cost heuristics is required for application to the capacitated lot-sizing problem.

Our basic approach begins by considering a setup in period t (where t initially equals 1) and computing the maximum profit per unit of demand satisfied in period t using only the setup in period t. Note that, given a setup in period t, we can sort orders in periods $t, \ldots, T$ in nonincreasing order of contribution to profit based solely on the variable costs incurred when assigning the order to the setup in period t (for the OSP when fixed delivery charges exist we must also subtract this cost from each order's contribution to profit). Orders are then allocated to the setup in nonincreasing order of contribution to profit until either the setup capacity is exhausted or no additional attractive orders exist. After computing the maximum profit per unit of demand satisfied in period t using only the setup in period t, we then compute the maximum profit per unit satisfied in periods $t, \ldots, t+j$ using only the setup in period t, for $j = 1, \ldots, j'$, where period j' is the first period in the sequence such that the maximum profit per unit in periods $t, \ldots, t+j'$ is greater than or equal to the maximum profit per unit in periods $t, \ldots, t+j'+1$. The capacity-feasible set of orders that leads to the greatest profit per unit in periods $t, \ldots, j'$ using the setup in period t is then assigned to production in period t, assuming the maximum profit per unit is positive. If the maximum profit per unit for any given setup period does not exceed zero, however, we do not assign any orders to the setup and thus eliminate the setup.

Since we consider a capacity-constrained problem, we can either consider period $j'+1$ (as is done in the Silver-Meal and Least Unit Cost heuristics) or period $t+1$ as the next possible setup period following period t. We use both approaches and retain the solution that produces higher net profit. Note that if we consider period $t + 1$ as the next potential setup period following period t, we must keep track of those orders in periods $t + 1$ and higher that are already assigned to period t (and prior) production, since these will not be available for assignment to period $t + 1$ production. Finally, after applying this greatest unit profit heuristic, we apply Phase III of the FSG algorithm (see the Chapter Appendix in Section 2.6) to the resulting solution, in an effort to further improve the heuristic solution value by looking for opportunities to effectively use any unused setup capacity.

2.3.3.3 Linear Programming Rounding Heuristic

Our third heuristic solution approach uses the LP relaxation solution as a starting point for a linear programming rounding heuristic. We focus on rounding the setup (y_t) and order selection (z_{mt}) variables that are fractional in the LP relaxation solution (rounding the order selection variables is not, however, relevant for the OSP-NDC problem, since the z_{mt} variables do not exist in this special case). We first consider the solution that results by setting all (non-zero) fractional y_t and z_{mt} variables from the LP relaxation solution to one. We then apply the second and third phases of our FSG algorithm to ensure a capacity feasible solution, and to search for unselected orders to allocate to excess production capacity in periods where the setup variable was rounded to one.

We also use an alternative version of this procedure, where we round up the setup variables with values greater than or equal to 0.5 in the LP relaxation solution, and round down those with values less than 0.5. Again we subsequently apply Phases II and III of the FSG algorithm to generate a good capacity-feasible solution (if the maximum setup variable value takes a value between 0 and 0.5, we round up only the setup variable with the maximum fractional variable value and apply Phases II and III of the FSG algorithm). Finally, based on our discussion in

Section 2.3.2, we have a choice of three different formulations for generating LP relaxation starting solutions for the rounding procedure: formulation [OSP] (Section 2.2), [ASF] (Section 2.3.2), or [DASF] (Section 2.3.2). As our computational results later discuss, starting with the LP relaxation solution from the [DASF] formulation provides solutions that are, on average, far superior to those provided using the other LP relaxation solutions. However, the size of this LP relaxation also far exceeds the size of our other LP relaxation formulations, making this formulation impractical as problem sizes become large. We use the resulting LP relaxation solution under each of these formulations and apply the LP rounding heuristic to all three of these initial solutions for each problem instance, retaining the solution that provides the highest net profit.

2.4 Scope and Results of Computational Tests

This section discusses a broad set of computational tests intended to evaluate our upper bounding and heuristic solution approaches. Our results focus on gauging both the ability of the different LP relaxations presented in Section 2.3.2 to provide tight upper bounds on optimal profit, and the performance of the heuristic procedures discussed in Section 2.3.3 in providing good feasible solutions. Section 2.4.1 next discusses the scope of our computational tests, while Sections 2.4.2 and 2.4.3 report results for the OSP, OSP-NDC, and OSP-AND versions of the problem.

2.4.1 Computational Test Setup

This section presents the approach we used to create a total of 3,240 randomly generated problem instances for computational testing, which consist of 1,080 problems for each of the OSP, OSP-NDC, and OSP-AND versions of the problem. Within each problem version (OSP, OSP-NDC, and OSP-AND), we used three different settings for the number of orders per period, equal to 25, 50, and 200. In order to create a broad set of test instances, we considered a range of setup cost values, production capacity limits, and per unit order revenues.[2] Table 2–4 provides the set of distributions used for randomly generating these parameter values in our test cases. The total number of combinations of parameter distribution settings shown in Table 2–4 equals 36, and for each unique choice of parameter distribution settings we generated 10 random problem instances. This produced a total of 360 problem instances for each of the three values of the number of orders per period (25, 50, and 200), which equals 1,080 problem instances for each problem version. As the distributions used to generate production capacities in

Table 2–4: Probability distributions used for generating problem instances.

Parameter	**Number of Distribution Settings**	**Distributions used[a] for Parameter Generation**
Setup cost (varies from period-to-period)	3	U[350,650]
		U[1750,3250]
		U[3500,6500]
Per unit per period holding cost[b]	2	$0.15 \times p/50$
		$0.25 \times p/50$
Production capacity in a period (varies from period-to-period)[c]	3	U[$d/3 - .05d$, $d/3 + .05d$]
		U[$d/2 - .1d$, $d/2 + .1d$]
		U[$d - .15d$, $d + .15d$]
Per unit order revenue (varies from order-to-order)	2	U[28,32]
		U[38,42]

[a] U[a, b] denotes a uniform distribution on the interval [a, b].
[b] p denotes the variable production cost. We assume 50 working weeks in one year.
[c] d denotes the expected per-period total demand, which equals the mean of the distribution of order sizes multiplied by the number of orders per period.

Table 2–4 indicate, we maintain a constant ratio of average production capacity per period to average total demand per period. That is, we maintain the same average order size (average of d_{mt} values) across each of these test cases, but the average capacity per period for the 200-order problem sets is four times that of the 50-order

[2] These three parameters appeared to be the most critical ones to vary widely in order to determine how robust our solution methods were to problem parameter variation.

problem sets and eight times that of the 25-order problems. Because the total number of available orders per period tends to strongly affect the relative quality of our solutions (as we later discuss), we report performance measures across all test cases and also individually within the 25, 50, and 200 order problem sets.

In order to limit the scope of our computational tests to a manageable size, we chose to limit the variation of certain parameters across all of the test instances. The per unit production cost followed a distribution of U[20,30] for all test instances (where U$[a, b]$ denotes a Uniform distribution on the interval $[a, b]$), and all problem instances used a 16-period planning horizon. We also used an order size distribution of U[10,70] for all test problems (i.e., the d_{mt} values follow a uniform distribution on [10,70]). For the OSP, the distribution used for generating fixed delivery charges was U[100,600].[3] By including a wide range of levels of production capacity, setup cost, and order volumes, we tested a set of problems which would fairly represent a variety of actual production scenarios.

Observe that the two choices for distributions used to generate per unit order revenues use relatively narrow ranges. Given that the distribution used to generate variable production cost is U[20,30], the first of these per unit revenue distributions, U[28,32], produces problem instances in which the contribution to profit (after subtracting variable production cost) is quite small—leading to fewer attractive orders after considering setup and holding costs. The second distribution, U[38,42], provides a more profitable set of orders. We chose to keep these ranges very narrow because our preliminary test results showed that a tighter range, which

[3] We performed computational tests with smaller per-order delivery charges, but the results were nearly equivalent to those presented for the OSP-NDC in Table 2.4.2, since the profitability of the orders remained essentially unchanged. As we increased the average delivery charge per order, more orders became unprofitable, creating problem instances that were quite different from the OSP-NDC case.

implies less per unit revenue differentiation among orders, produces more difficult problem instances. Those problem instances with a greater range of per unit revenue values among orders tended to be solved in CPLEX via branch-and-bound much more quickly than those with tight ranges, and we wished to ensure that our computational tests reflected more difficult problem instances.

A tighter range of unit revenues produces more difficult problem instances due to the ability to simply 'swap' orders with identical unit revenues in the branch-and-bound algorithm, leading to alternative optimal solutions at nodes in the branch-and-bound tree. For example, if an order m in period t is satisfied at the current node in the branch-and-bound tree, and some other order m' is not satisfied, but $r_{mt} = r_{m't}$ and $d_{mt} = d_{m't}$, then a solution which simply swaps orders m and m' has the same objective function as the first solution, and no improvement in the bound occurs as a result of this swap. So, we found that when the problem instance has less differentiation among orders, the branch-and-bound algorithm can take substantially longer, leading to more difficult problem instances. Barnhart et al. [7] and Balakrishnan and Geunes [6] observed similar swapping phenomena in branch-and-bound for machine scheduling and steel production planning problems, respectively.

All linear and mixed integer programming (MIP) formulations were solved using the CPLEX 6.6 solver on an RS/6000 machine with two PowerPC (300MHz) CPUs and 2GB of RAM. We will refer to the best solution provided by the CPLEX branch-and-bound algorithm as the *MIP solution*. The remaining subsections summarize our results. Section 2.4.2 reports the results of our computational experiments for the OSP-NDC and the OSP, and Section 2.4.3 presents the findings for the OSP-AND (all-or-nothing order satisfaction) problem. For the OSP-AND problem instances discussed in Section 2.4.3, we assume that the revenue parameters provided represent revenues in excess of fixed delivery charges (since we always satisfy all or none of the demand for the OSP-AND, this is without loss of generality).

2.4.2 Results for the OSP and the OSP-NDC

Recall that the OSP assumes that we have the flexibility to satisfy any proportion of an order in any period, as long as we do not exceed the production capacity in the period. Because of this, when no fixed delivery charges exist, the only binary variables in the OSP-NDC correspond to the T binary setup variables, and solving these problem instances to optimality using CPLEX's MIP solver did not prove to be very difficult. The same is not necessarily true of the OSP-AND, as we later discuss in Section 2.4.3. Surprisingly, the OSP (which includes a binary fixed delivery-charge forcing (z_{mt}) variable for each order-period combination) was not substantially computationally challenging either. All of the OSP-NDC and all but two of the OSP instances were solved optimally using branch-and-bound within the allotted branch-and-bound time limit of one hour. Even though we are able to solve the OSP and OSP-NDC problem instances using CPLEX with relative ease, we still report the upper bounds provided by the different LP relaxations for these problems in this section. This allows us to gain insight regarding the strength of these relaxations as problem parameters change, with knowledge of the optimal mixed integer programming (MIP) solution values as a benchmark.

Table 2.4.2 presents optimality gap measures based on the solution values resulting from the LP (OSP-LP) relaxation upper bound, the aggregated setup forcing (ASF) relaxation upper bound, and the disaggregated setup forcing (DASF) relaxation upper bound for the OSP-NDC and OSP problem instances. The last row of the table shows the percentage of problem instances for which CPLEX was able to find an optimal solution via branch-and-bound. As Table 2.4.2 shows, for the OSP-NDC, all three relaxations provide good upper bounds on the optimal solution value, consistently producing gaps of less than 0.25%, on average. As expected, the [ASF] formulation provides better bounds than the simple OSP-LP relaxation, and the [DASF] formulation provides the tightest bounds. We note that as the number of potential orders and the per-period production capacities increase, the relative performance of the relaxations improves, and the optimality gap decreases. Since an optimal solution exists such that at most one order per period will be partially satisfied under any relaxation, as the problem size grows, we fulfill a greater proportion of orders in their entirety. So the impact of our choice of which order to partially satisfy diminishes with larger problem sizes. Note also, however, that a small portion of this improvement is attributable to the increased optimal solution values in the 50- and 200-order cases.

For the OSP, we have non-zero fixed delivery costs and cannot therefore eliminate the binary z_{mt} variables from formulation [OSP]. In addition, since formulation [OSP] includes the continuous v_{mt} variables, it has the highest number of variables of any of the capacitated versions we consider. This does not necessarily, however, make it the most difficult problem class for solution via CPLEX, as a later comparison of the results for the OSP and OSP-AND indicates.

The upper bound optimality gap results reported in Table 2.4.2 for the OSP are significantly larger than those for the OSP-NDC.[4] This is because this formulation permits setting fractional values of the fixed delivery-charge forcing (z_{mt}) variables, and therefore does not necessarily charge the entire fixed delivery cost when meeting a fraction of some order's demand. For this problem set the [DASF] formulation provides substantial value in obtaining strong upper bounds on the optimal net profit although, as shown in Table 2–6, the size of this formulation

[4] For the two problems that could not be solved to optimality via branch-and-bound using CPLEX due to memory limitations, the MIP solution value used to compute the upper bound optimality gap is the value of the best solution found by CPLEX.

makes solution via CPLEX substantially more time consuming as the number of orders per period grows to 200.

Table 2–5: OSP-NDC and OSP problem optimality gap measures.

	OSP-NDC				OSP			
% Gap	Orders per Period			Overall	Orders per Period			Overall
(from MIP)	25	50	200	Average	25	50	200	Average
OSP-LP[a]	0.24%	0.14%	0.05%	0.14%	9.26%	6.09%	0.57%	5.31%
ASF[b]	0.18	0.12	0.04	0.11	9.21	6.07	0.56	5.28
DASF[c]	0.11	0.07	0.03	0.07	1.58	0.35	0.10	0.68
% Opt[d]	100	100	100	100	100	99.7	99.7	99.8

Note: Entries in each "orders per period" class represent an average among 360 test instances.

[a] (OSP-LP − MIP)/MIP × 100%.

[b] (ASF − MIP)/MIP × 100%.

[c] (DASF − MIP)/MIP × 100%.

[d] % of problems for which CPLEX branch-and-bound found an optimal solution.

Table 2–6 summarizes the solution times for solving the OSP-NDC and the OSP. The MIP solution times reflect the average time required to find an optimal solution for those problems that were solved to optimality in CPLEX (the two problems that CPLEX could not solve to optimality are not included in the MIP solution time statistics). We used the OSP-LP formulation as the base formulation for solving all mixed integer programs. The table also reports the times required to solve the LP relaxations for each of our LP formulations (OSP-LP, ASF, and DASF). We note that the [ASF] and [DASF] LP relaxations often take longer to solve than the mixed integer problem itself. The [DASF] formulation, despite providing the best upper bounds on solution value, quickly becomes less attractive as the problem size grows because of the size of this LP formulation. Nonetheless, the relaxations provide extremely tight bounds on the optimal solution as shown in the table. As we later show, however, solving the problem to optimality in CPLEX is not always a viable approach for the restricted OSP-AND discussed in the following section.

Table 2–6 reveals that the MIP solution times for the OSP were also much greater than for the OSP-NDC. This is due to the need to simultaneously track the binary (z_{mt}) and continuous (v_{mt}) variables for the OSP with non-zero fixed delivery costs. As expected, the average and maximum solution times for each relaxation increased with the number of orders per period. As we noted previously, the percentage optimality gaps, however, substantially decrease as we increase the number of orders per period.

Table 2–6: OSP-NDC and OSP solution time comparison.

	OSP-NDC Orders per Period			**OSP** Orders per Period		
Time Measure (CPU seconds)	25	50	200	25	50	200
Average MIP Solution Time	0.1	0.1	0.2	3.3	19.1	129.4
Maximum MIP Solution Time	0.1	0.1	0.3	44.8	541.3	3417.2
Average OSP-LP Solution Time	0.1	0.1	0.3	0.1	0.1	0.3
Maximum OSP-LP Solution Time	0.1	0.2	0.5	0.1	0.1	0.5
Average ASF Solution Time	0.5	1.5	14.0	0.4	1.0	8.3
Maximum ASF Solution Time	0.7	2.2	25.2	0.6	1.6	15.4
Average DASF Solution Time	5.3	27.3	727.2	3.3	15.7	333.8
Maximum DASF Solution Time	18.4	64.3	1686.7	12.1	47.1	1251.9

Note 1: Entries represent average/maximum among 360 test instances.
Note 2: LP relaxation solution times include time consumed applying the LP rounding heuristic to the resulting LP solution, which was negligible.

We next present the results of applying our heuristic solution approaches to obtain good solutions for the OSP and OSP-NDC. We employ the three heuristic solution methods discussed in Section 2.3.3, denoting the Lagrangian-based heuristic as LAGR, the greatest unit profit heuristic as GUP, and the LP rounding heuristic as LPR. Table 2–7 provides the average percentage deviation from the best upper bound (as a percentage of the best upper bound) for each heuristic solution method. Note that since we found an optimal solution for all but two of the OSP and OSP-NDC problem instances, the upper bound used in computing the heuristic solution gaps is nearly always the optimal mixed integer solution value. The last row in Table 2–7 shows the resulting lower bound gap from our composite solution procedure, which selects the best solution among all of the heuristic methods applied. The average lower bound percentage gap is within 0.06% of optimality for the OSP-NDC, while that for the OSP is 1.69%, indicating that overall, our heuristic solution methods are quite effective. As the table indicates, the heuristics perform much better in the absence of fixed delivery costs. For the Lagrangian-based and LP rounding heuristics, we can attribute this in part to the difficulty in obtaining good relaxation upper bounds for the OSP as compared to the OSP-NDC. Observe that as the upper bound decreases (i.e., as the number of orders per period increases), these heuristics tend to improve substantially. The GUP heuristic, on the other hand, appears to have difficulty identifying a good combination of setup periods in the presence of fixed delivery charges. Although it appears, based on average performance, that the LPR heuristic dominates the LAGR and GUP heuristics, the last row of the table reveals that this is not universally true. Each of our heuristic approaches provided the best solution value for some nontrivial subset of the problems tested.

Table 2–7: OSP and OSP-NDC heuristic solution performance measures.

	OSP-NDC				**OSP**			
% Gap	Orders per Period			Overall	Orders per Period			Overall
(from UB)	25	50	200	Average	25	50	200	Average
LAGR v. UB[a]	1.34%	0.58%	0.32%	0.75%	6.35%	4.07%	2.16%	4.19%
GUP v. UB[b]	1.00	0.69	0.44	0.71	7.27	6.91	5.39	6.52
LPR v. UB[c]	0.25	0.15	0.05	0.15	8.32	5.31	0.96	4.86
Best LB[d]	0.10	0.07	0.02	0.06	3.08	1.55	0.44	1.69

Note: Entries in each "orders per period" class represent an average among 360 test instances.

[a] (LAGR – UB)/UB × 100%.

[b] (GUP – UB)/UB × 100%.

[c] (LPR – UB)/UB × 100%.

[d] Uses the best heuristic solution value for each problem instance.

2.4.3 Results for the OSP-AND

We next provide our results for the OSP-AND where, if we choose to accept an order, we must satisfy the entire order (i.e., no partial order satisfaction is allowed).

Finding the optimal solution to the OSP-AND can be much more challenging than for the OSP, since we now face a more difficult combinatorial "packing" problem (i.e., determining the set of orders that will be produced in each period is similar to a multiple knapsack problem).

Table 2–8 provides upper bound optimality gap measures based on the solution values resulting from our different LP relaxation formulations, along with the percentage of problem instances that were solved optimally via the CPLEX branch-and-bound algorithm. Observe that the upper bound optimality gap measures are quite small and only slightly larger than those observed for the OSP-NDC. The reason for this is that the LP relaxation formulations are identical in both cases (as discussed in Section 2.2), and the optimal LP relaxation solution violates the all-or-nothing requirement for at most one order per period. Thus, even in the OSP-NDC case, almost all orders that are selected are fully satisfied in the LP relaxation solution. In contrast to the [OSP] formulation, the binary z_{mt} variables in the OSP-AND model now represent "order selection" variables rather than fixed delivery-charge forcing variables. That is, since we net any fixed delivery charge out of the net revenue parameters R_{mt}, and the total revenue for an order in a period now equals $R_{mt}z_{mt}$ in this formulation, we have strong preference for z_{mt} variable values that are either close to one or zero. In the [OSP] formulation, on the other hand, the z_{mt} variables are multiplied by the fixed delivery-charge terms (F_{mt}) in the objective function, leading to a strong preference for low values of the z_{mt} variables and, therefore, a weaker upper bound on optimal net profit. Note also that as the number of possible orders increases (from the 25-order case to the 200-order case), the influence of the single partially satisfied order in each period on the objective function value diminishes, leading to a reduced optimality gap as the number of orders per period increases. As the last row of Table 2–8 indicates, we were still quite successful in solving these problem instances to optimality in

CPLEX. The time required to do so, however, was substantially greater than that for either the OSP or OSP-NDC, because of the complexities introduced by the all-or-nothing order satisfaction requirement.

Table 2–9 summarizes the resulting solution time performance for the OSP-AND. We note here that our relaxation solution times are quite reasonable, especially as compared to the MIP solution times, indicating that quality upper bounds can be found very quickly. Again, the MIP solution times reflect the average time required to find an optimal solution for those problems that were solved to optimality in CPLEX (those problems which CPLEX could not solve to optimality are not included in the MIP solution time statistics). The table does not report the time required to solve our different LP relaxation formulations, since the OSP-AND LP relaxation is identical to the OSP-NDC LP relaxation, and these times are therefore shown in Table 2–6.

Unlike our previous computational results for the OSP and the OSP-NDC, we found several problem instances of the OSP-AND in which an optimal solution was not found either due to reaching the time limit of one hour or because of memory limitations. For the problem instances we were able to solve optimally, the MIP solution times were far longer than those for the OSP problem. This is due to the increased complexity resulting from the embedded “packing problem” in the OSP-AND problem. Interestingly, however, in contrast to our previous results for the OSP, the average and maximum MIP solution times for the OSP-AND were *smaller* for the 200-order per period problem set than for the 25 and 50-order per period problem sets. The reason for this appears to be because of the nearly non-existent integrality gaps of these problem instances, whereas these gaps increase when the number of orders per period is smaller.

Table 2–8: OSP-AND optimality gap measures.

	Orders per Period			
Gap Measurement	25	50	200	**Overall Average**
OSP-LP vs. MIP Solution[a]	0.34%	0.20%	0.06%	0.20%
ASF vs. MIP Solution[b]	0.28	0.18	0.05	0.17
DASF vs. MIP Solution[c]	0.21	0.10	0.03	0.11
% Optimal[d]	96.7	94.2	100	97

Note: Entries within each "orders per period" class represent average among 360 test instances.

[a] (OSP-LP − MIP)/MIP × 100%.

[b] (ASF − MIP)/MIP × 100%.

[c] (DASF − MIP)/MIP × 100%.

[d] % of problems for which CPLEX branch-and-bound found an optimal solution.

Table 2–9: OSP-AND solution time comparison.

	Orders per Period		
Time Measure (CPU seconds)	25	50	200
Average MIP Solution Time	42.0	67.9	21.9
Maximum MIP Solution Time	1970.1	1791.8	1078.8

Note: Entries represent average/maximum among 360 test instances.

Table 2–10 shows that once again our composite heuristic procedure performed extremely well on the problems we tested. The percentage deviation from optimality in our solutions is very close to that of the OSP-NDC, and much better than that of the OSP, with an overall average performance within 0.25% of optimality. We note, however, that the best heuristic solution performance for both the OSP-NDC and the OSP-AND occurred using the LP rounding heuristic applied to the DASF LP relaxation solution. As Table 2–6 showed, solving the DASF LP relaxation can be quite time consuming as the number of orders per period grows, due to the size of this formulation. We note, however, that for the OSP-NDC and OSP-AND, applying the LP rounding heuristic to the ASF LP relaxation solution produced results very close to those achieved using the DASF LP relaxation solution in much less computing time. Among all of the 3,240 OSP, OSP-NDC, and OSP-AND problems tests, the best heuristic solution value was within 0.67% of optimality on average, indicating that overall, the heuristic solution approaches we presented provide an extremely effective method for solving the OSP and its variants.

Table 2–10: OSP-AND heuristic solution performance measures.

	OSP-AND			
	Orders per Period			Overall
Gap Measurement	25	50	200	Average
LAGR vs. UB[a]	3.95%	3.92%	0.33%	2.73%
GUP vs. UB[b]	1.85	0.83	0.46	1.04
LPR vs. UB[c]	0.80	0.31	0.12	0.41
Best LB[d]	0.49	0.19	0.06	0.25

Note: Entries within each "orders per period" class represent average among 360 test instances

[a] (LAGR – UB)/UB × 100%.

[b] (GUP – UB)/UB × 100%.

[c] (LPR – UB)/UB × 100%.

[d] Uses the best heuristic solution value for each problem instance.

2.5 Conclusions

When a producer has discretion to accept or deny production orders, determining the best set of orders to accept based on both revenue and production/delivery cost implications can be quite challenging. For situations when no production capacities exist, we show how the order selection problem can be solved using a similar approach to the Wagner-Whitin [83] dynamic programming algorithm employed for the ELSP. When facing production capacities, several variations of the problem emerge, and we formulated and presented solution approaches to these as well.

We considered variants of the problem both with and without fixed delivery charges, as well as contexts that permit the producer to satisfy any chosen fraction of any order quantity, thus allowing the producer to ration its capacity. We provided three linear programming relaxations that produce strong upper bound values on the optimal net profit from integrated order selection and production planning decisions. We also provided a set of three effective heuristic solution methods for the OSP. Computational tests performed on a broad set of randomly generated problems demonstrated the effectiveness of our heuristic methods and upper bounding procedures. Problem instances in which the producer has the flexibility to determine any fraction of each order it will supply, and no fixed delivery charges exist, were easily solved using the MIP solver in CPLEX. When fixed delivery charges are present, however, the problem becomes more difficult, particularly as the number of available orders increases. Optimal solutions were still obtained, however, for nearly all test instances within one hour of computing time when partial order satisfaction was allowed. When the producer must take an all-or-nothing approach, satisfying the entire amount of each order it chooses to satisfy, the problem becomes substantially more challenging, and the heuristic solutions we presented become a more practical approach for solving such problems.

We expand our discussion of demand (or order) selection flexibility in a production planning context over the next two chapters. Specifically, we will introduce pricing as a decision variable in the requirements planning problem in Chapter 3. Then we will consider the role that demand uncertainty plays in demand source selection decisions in Chapter 4.

2.6 Appendix

Description of Feasible Solution Generator (FSG) Algorithm for OSP

This appendix describes the Feasible Solution Generator (FSG) algorithm, which takes as input a solution that is feasible for all OSP problem constraints except the production capacity constraints, and produces a capacity-feasible solution. Note that we present the FSG algorithm as it applies to the OSP, and that certain straightforward modifications must be made for the OSP-AND version of the problem.

Phase I: Assess attractiveness of additional setups

0) Let j denote a period index, let $p(j)$ be the most recent production period prior to and including period j, and let $s(j)$ be the next setup after period j. If no production period exists prior to and including j, set $p(j) = 0$. Set $j = T$ and

$s(j) = T + 1$ and let X_j denote the total planned production (in the current, possibly capacity-infeasible solution) for period j.

1) Determine the most recent setup $p(j)$ as described in Step 0. If $p(j) = 0$, go to Phase II. If $X_{p(j)} \leq C_{p(j)}$, set $s(p(j) - 1) = p(j)$ and $j = p(j) - 1$ and repeat Step 1 (note that we maintain $s(j) = j + 1$). Otherwise, continue.

2) Compare the desired production in period $p(j)$, $X_{p(j)}$, with actual capacities over the next $s(j) - p(j)$ periods. If $X_{p(j)} > \sum_{t=p(j)}^{s(j)-1} C_t$, and the sum of the revenues for all selected orders for period j exceed the setup cost in period j, then add a production setup in period j and transfer all selected orders in period j to the new production period j. Otherwise do not add the setup in period j. Set $s(p(j) - 1) = p(j), j = p(j) - 1$, and return to Step 1.

Phase II: Transfer/remove least profitable production orders

0) Let $d_{m,p(j),j}$ denote the amount of demand from order m in period j to be satisfied by production in period $p(j)$ in the current (possibly capacity-infeasible) plan. When reading in the problem data, all profitable order and production period combinations were determined. Based on the solution, we maintain a list of all orders that were satisfied, and this list is kept in *nondecreasing* order of per-unit profitability. Per-unit profitability is defined as follows: $\Pi_{m,p(j),j} = r_{mj} - p_{p(j)} - \sum_{t=p(j)}^{j-1} h_t - \frac{F_{mj}}{d_{mj}}$. We will use this list to determine the least desirable production orders to maintain.

1) If no periods have planned production that exceeds capacity, go to Phase III. While there are still periods in which production exceeds capacity, find the next least profitable order period combination, $(m^*, p(j^*), j^*)$, in the list.

2) If $X_{p(j^*)} > C_{p(j^*)}$, consider shifting or removing an amount equal to $d^* = \min\{d_{m^*,p(j^*),j^*}, X_{p(j^*)} - C_{p(j^*)}\}$ from production in period $p(j^*)$ (otherwise, return to Step 1). If an earlier production period $\tau < p(j^*)$ exists such that $X_\tau < C_\tau$, then move an amount equal to $\min(d^*, C_\tau - X_\tau)$ to the production in period τ; i.e., $d_{m^*,\tau,j^*} = \min\left(d^*, C_\tau - X_\tau\right)$. Otherwise, reduce the amount of production in period $p(j^*)$ by d^* and set $d_{m^*,p(j^*),j^*} = d_{m^*,p(j^*),j^*} - d^*$.

3) Update all planned production levels and order assignments and update the number of periods in which production exceeds capacity. Return to Step 1.

Phase III: Attempt to increase production in under-utilized periods

0) Create a new list for each period of all profitable orders <u>not</u> fulfilled. Each list is indexed in *nonincreasing* order of per-unit profitability, as defined earlier. Let j denote the first production period.

1) If $j = T + 1$, STOP with a feasible solution. Otherwise, continue.

2) If $C_{p(j)} > X_{p(j)}$, excess capacity exists in period $p(j)$. Choose the next most profitable order from period j, and let m^* denote the order index for this order. Let $d_{m^*,p(j),j} = \min\left\{d_{m^*,j}, C_{p(j)} - X_{p(j)}\right\}$, and assign an additional $d_{m^*,p(j),j}$ to production in period $p(j)$.

3) If there is remaining capacity and additional profitable orders exist for period j, the repeat Step 2. Otherwise, set $j = j + 1$ and return to Step 1.

CHAPTER 3
PRICING, PRODUCTION PLANNING, AND ORDER SELECTION FLEXIBILITY

3.1 Introduction

Firms that produce made-to-order goods often make pricing decisions prior to planning the production required to satisfy demands. These decisions require the firm's representatives (often sales/marketing personnel in consultation with manufacturing management) to determine prices, which imply certain demand volumes the firm will need to satisfy. Such pricing decisions are typically made prior to establishing future production plans and are in many cases made based on the collective judgment of sales and marketing personnel. This results in decisions that do not account for the interaction between pricing decisions and production requirements, and how these factors affect overall profitability. Lee [44] recently noted that one of the common pitfalls of supply chain management practice occurs when those who influence demand within the firm (e.g., marketing, sales) do not properly account for operations costs in demand planning, while supply chain managers fail to recognize that demand is not completely determined exogenously. He argues that integrating supply and demand-based management offers great opportunity for future value creation and serves as "the next competitive battleground in the 21st century."

Since production environments often involve significant fixed production costs, justifying these fixed costs requires a demand level at which revenues exceed not only variable costs, but the fixed costs incurred in production as well. Decisions on the demand volume the organization must satisfy, and the implied revenues and costs, can be a critical determinant of the firm's profitability. Past operations modeling literature has not fully addressed integrated pricing and production planning decisions in make-to-order systems with the types of nonlinear production cost structures often found in practice as a result of production economies of scale. We offer modeling and solution approaches for integrating these decisions in single-stage systems.

Most of the requirements planning literature focuses on production requirements based on pre-specified demands, with no adjustments for price flexibility. In this chapter, we introduce a requirements planning model that implicitly determines the best demand levels to satisfy in order to maximize contribution to profit when demand is a decreasing function of price. In other words, the firm will select the demand level to satisfy by setting a single price for the product.

We make several contributions to the literature through our model and solution approaches introduced in this chapter. First, our combined pricing and production planning model permits multiple price-demand curves in each period, which effectively represents the possibility of offering different prices in different markets, where each market has a unique response to market price. Moreover, this model generalizes the *order selection* approach presented in Chapter 2, where a firm faced a set of customer orders, from which it selected the most profitable subset. In the *order selection* context, we can use our requirements planning with pricing model and apply a unique price to each order, rather than a single price for all demands. Our solution approach also accommodates more general production cost functions than previously considered in the requirements planning and pricing literature, along with explicit consideration of both general concave and piecewise-linear concave revenue functions.

Given fixed plus linear production costs and piecewise-linear concave revenue functions, we also provide a ‘tight’ linear programming formulation of our model, using a dual-based solution approach to show that this formulation has zero duality gap. This result, and the formulations discovered while developing the approach, played a key role in formulating the relaxations used in solving the capacitated OSP models in Chapter 2, where production capacities varied over time. Our final major contribution also addresses a capacitated version of the model. Assuming time-invariant production capacity limits and piecewise-linear concave revenue functions in the total demand satisfied, we show that this problem can be solved in polynomial time.

Given the recent emphasis on differential pricing and demand management in manufacturing (e.g., Lee [44], Chopra and Meindl [22]), these models and associated solution approaches have the potential for broad application in practice. Analytics Operations Engineering, Inc., an operations strategy and execution consulting firm, recently cited application contexts in the specialty papers and timber industries in which integrated pricing and production planning models such as the ones we discuss can add substantial value in practice (for more details on these applications, please see Burman [17]).

Thomas [74] provided an analysis and solution algorithm for a related integrated pricing and production planning decision model. His model generalized the Wagner & Whitin [83] model by characterizing demand in each of a set of discrete time periods as a downward-sloped function of the price in each period, thus treating each period's price as a decision variable. The model proposed by Thomas [74] sets only a single price for all demands in any given period, whereas our model permits differential pricing in different markets. Moreover, we demonstrate that a 'tight' linear programming formulation exists for this problem under piecewise-linear concave revenue functions. We also extend the analysis to account for more general production cost functions in each period.

Additional contributions to the integrated pricing and production planning problem include the work of Kunreuther and Schrage [41] and Gilbert [33], who considered the problem when a single price must be used over the entire horizon. Kunreuther and Schrage [41] provided bounds on the optimal solution value under time varying production cost assumptions, while Gilbert [33] assumed time-invariant production setup and holding costs and provided an exact polynomial-time algorithm. Recall the paper by Loparic, Pochet, and Wolsey [50] that we introduced in Chapter 2. They considered a problem in which a producer wishes to maximize net profit from sales of a single item and does not have to satisfy all outstanding demand in every period. Their model contains no pricing decisions, effectively assuming that only one demand source exists in every period, and that the revenue from a single demand source is proportional to the volume of demand satisfied. In contrast, we allow revenue to be a general concave nondecreasing function of the amount of demand satisfied, which is consistent with a downward-sloped demand curve as a function of price. Also discussed in Chapter 2 was the paper by Lee, Çetinkaya, and Wagelmans [43], in which they introduce a production planning model with *demand time windows*. While their model assumes that all pre-specified demands must be filled during the planning horizon, our approach implicitly determines demand levels through pricing.

Bhattacharjee and Ramesh [13] considered the pricing problem for perishable goods using a very general function to characterize demand as a function of price. They also assumed upper and lower bounds on prices, characterized structural properties of the optimal profit function, and developed heuristic methods for solving the resulting problems. Biller, Chan, Simchi-Levi, and Swann [14] analyzed a model similar to ours under strictly linear production costs (i.e., no fixed setup costs, and assuming time-varying production capacity limits), which they solved efficiently using a greedy algorithm. While our discussion of the relevant literature has focused on deterministic approaches for integrated pricing and production planning problems, some additional work on dynamic pricing exists that addresses stochastic demand environments; for past work on integrated pricing and production/inventory planning in a stochastic demand setting, please see Thomas [75], Gallego and van Ryzin [30], and Chan, Simchi-Levi, and Swann [21].

The remainder of this chapter is organized as follows. Section 3.2 presents a formal definition and mixed integer programming formulation of the general requirements planning problem with pricing. In this section we provide our solution approaches for this problem, the first of which extends the Wagner-Whitin [83] shortest path solution method (discussed in Chapter 2) to contexts with general concave revenue functions and fixed-charge production costs. Assuming piecewise-linear concave revenue functions, we then provide a dual-based polynomial-time algorithm for solving the uncapacitated problem. This dual-based solution approach allows us to show that the problem reformulation in Section 3.2.2 has a linear programming relaxation whose optimal value equals that of the optimal mixed integer solution; i.e., the problem formulation is "tight". We also explore the generality of our solution approaches with respect to different functional forms for the production cost functions and under multiple market price-demand curves in any given period. In addition to presenting solution approaches to several uncapacitated versions of the problem, we provide an analysis of the equal-capacity version of the model under piecewise-linear concave revenue functions. Section 3.3 discusses different pricing interpretations from our models, and illustrates how our pricing model can be cast as an equivalent "order selection" problem, thus broadening its potential for application in practice.

3.2 Requirements Planning with Pricing

Consider a producer who manufactures a good to meet demand over a finite number of time periods, T. The production cost function in period t is denoted $g_t(\cdot)$, and is a nondecreasing concave function of the amount produced in period t, which we denote by x_t. Similarly, the revenue function in period t is denoted by $R_t(\cdot)$, and is a nondecreasing concave function of the *total demand satisfied* in period t, which we denote by D_t, with $R_t(0) = 0$ for all $t = 1, \ldots, T$. We assume that D_t, the total demand satisfied in any period t, is the sum of the demands satisfied from some M_t distinct markets. In each market we employ a standard assumption of a one-to-one correspondence between price and market demand volume in any period, where market demand is a downward-sloping function of price (see Gilbert [33]), and each market's revenue is a nondecreasing concave function of demand satisfied in the market. Given a total demand value of D_t in period t we solve an optimization subproblem to determine a price value in period t in every market m (equivalently, $D_t = \sum_{m=1}^{M_t} d_{mt}(\theta_{mt})$ where θ_{mt} is the price in market m in period t and $d_{mt}(\cdot)$ is the total demand in market m in period t as a function of price). Section 3.2.3.1 discusses how to determine the price in each market in period t given a demand volume of D_t; for now it is sufficient to simply consider the decision variables for the total demand in each period (i.e., the D_t variables).

Inventory costs are charged against ending inventory, where h_t denotes the unit holding cost in period t and I_t is a decision variable for the end-of-period inventory in period t. Letting C denote the production capacity limit (which does not depend on time), we formulate the *requirements planning with pricing* (RPP) problem as follows.

[RPP]

$$\text{maximize} \quad \sum_{t=1}^{T} \left(R_t(D_t) - (g_t(x_t) + h_t I_t) \right)$$

$$\text{subject to:} \quad D_t + I_t = x_t + I_{t-1} \qquad t = 1, \ldots, T, \tag{3.1}$$

$$x_t \leq C \qquad t = 1, \ldots, T, \tag{3.2}$$

$$x_t, I_t, D_t \geq 0 \qquad t = 1, \ldots, T. \tag{3.3}$$

The objective function maximizes net profit after production and holding costs; constraint set (3.1) ensures inventory balance in all periods and constraint set (3.2) enforces production capacity limits. The general RPP problem defined above maximizes the difference between concave functions and is, therefore, in general a difficult global optimization problem (see Horst and Tuy [37]). By providing certain somewhat mild restrictions on the functional forms of the revenue and production cost functions, $R_t(D_t)$ and $g_t(x_t)$, we arrive at a family of special cases of the RPP problem, several of which have broad applicability in practice.

Consistent with the vast majority of past production planning literature, except where specifically noted, we henceforth assume that production costs contain a fixed-charge structure; i.e., a fixed cost of S_t is incurred when performing a production setup in any period t, while the variable cost per unit in period t equals p_t (we later discuss in Section 3.2.3.2 the necessary extensions to handle production costs that contain a more general piecewise-linear nondecreasing concave cost structure). Under fixed plus linear production costs, unlimited production capacity, and a single price offered to all markets in each period we have the model first analyzed by Thomas [74], who proposed a dynamic programming recursion for solving the problem. The algorithm is similar to the Wagner-Whitin [83] algorithm for the ELSP, and relies on similar key structural properties of the problem. These properties include the zero inventory ordering (ZIO) property (if inventory is held at the end of period $t-1$ then we do not perform a setup in period t). The following section describes an equivalent shortest path algorithm (refer to Section 2.2.2 for a complete discussion) for this problem, along with an explicit characterization of the solution approach under concave revenue functions. While the shortest path method we present is generally equivalent to the dynamic programming method proposed by Thomas [74] when production costs contain a fixed-charge structure, we depart from this work in the following respects:

(i) we provide an exact solution approach for contexts in which total revenue is concave and nondecreasing in the amount of demand satisfied;

(ii) we show that the shortest path method generalizes to cases with multiple demand sources, each with a unique concave revenue curve; and

(iii) we show how to generalize the shortest path approach to provide an exact procedure for the case of piecewise-linear and concave production costs.

Thus, the following section lays the foundation for subsequent generalizations of our solution methodology to broader contexts.

3.2.1 Shortest Path Approach for the Uncapacitated RPP

Retaining our assumption of a fixed-charge production cost structure and assuming the revenue function $R_t(D_t)$ in every period t is a general nondecreasing concave function of D_t with $R_t(0) = 0$, we now update the Wagner-Whitin [83] shortest path approach (introduced in Chapter 2) for the uncapacitated RPP problem. Note that under these assumptions, for any fixed choice of the demand vector $(D_1, D_2, \ldots, D_T)$, the resulting problem is a simple ELSP. Now, we can decompose the T-period RPP problems into a set of smaller contiguous interval subproblems, using the shortest path graph structure previously shown in Figure 2–2. To illustrate the computation of arc length $c(t, t')$, where a setup is performed in period t and the next setup occurs in period $t' > t$, we solve the period $t, \ldots, t'-1$ subproblem of maximizing net profit in these periods. This period $t, \ldots, t'-1$ subproblem can be stated as

$$\text{maximize:} \quad \sum_{\tau=t}^{t'-1} \left(R_\tau(D_\tau) - h_\tau \textstyle\sum_{j=\tau+1}^{t'-1} D_j \right) - p_t \textstyle\sum_{j=t}^{t'-1} D_j \tag{3.4}$$

$$\text{subject to:} \quad D_\tau \geq 0 \qquad \tau = t \ldots, t'-1. \tag{3.5}$$

This decision problem separates by period, and since we are maximizing a set of nondecreasing concave functions, we arrive at the following characterization of the optimal amount of demand to satisfy in period τ, given a most recent setup in period t. For notational convenience we let $v_{t\tau} \equiv p_t + \sum_{j=t}^{\tau-1} h_j$ denote the cost per unit of demand satisfied in period τ using a setup in period $t \leq \tau$.

Theorem 1 *For the uncapacitated RPP, given a production setup in period t only, if a demand quantity D_τ exists such that $v_{t\tau}$ is in the set of subgradients of $R_\tau(\cdot)$ at D_τ, then D_τ is an optimal demand quantity for the subproblem given by (3.4) and (3.5).*

A proof of Theorem 1 can be found in Appendix A of Section 3.5. Note that if $R_\tau(\cdot)$ is everywhere differentiable with $\lim_{D\to\infty} R'_\tau(D) \leq v_{t\tau} \leq \lim_{D\downarrow 0} R'_\tau(D)$, then the optimal demand quantity as stated in the theorem can be determined by finding D_τ such that $R'_\tau(D_\tau) = v_{t\tau}$.

Given any $t \leq \tau \leq t' - 1$, if a $D_\tau > 0$ exists that satisfies the condition of Theorem 1, then the optimal value of D_τ for the subproblem, which we denote by $D^*_\tau(t)$, equals this demand value. Otherwise, assuming a finite (non-negative) value of $v_{t\tau}$, we must have either $D^*_\tau(t) = 0$ (if all subgradients at all $D_\tau > 0$ are less than $v_{t\tau}$) or $D^*_\tau(t) = \infty$ (if a subgradient exists for each $D_\tau > 0$ that is greater than $v_{t\tau}$). Then the maximum possible profit in periods $t, \ldots, t' - 1$ (assuming the only setup within these periods occurs in period t, which we denote by $\Pi(t, t')$) is given by

$$\Pi(t,t') = \sum_{\tau=t}^{t'-1} \left(R_\tau(D^*_\tau(t)) - h_\tau \sum_{j=\tau+1}^{t'-1} D^*_j(t) \right) - p_t \sum_{j=t}^{t'-1} D^*_j(t) - S_t, \tag{3.6}$$

and the arc length for arc (t, t') is therefore given by

$$c(t,t') = \max\{0, \Pi(t,t')\}. \tag{3.7}$$

With appropriate preprocessing and recursive computations of the $\Pi(t, t')$ values, we can determine all $\Pi(t, t')$ values in $\mathcal{O}(T^2)$ time. As discussed previously, the longest path on an acyclic network can be found in $\mathcal{O}(T^2)$ time in the worst case (see Lawler [42]). Therefore, the overall solution effort is no worse than $\mathcal{O}(T^2)$.

We next consider a particular special case of the concave revenue functions, which we will use for more detailed analysis in subsequent sections. Suppose that the revenue function in each period can be represented as a nondecreasing piecewise-linear concave function of demand. We assume that the revenue function in period t has $J_t + 1$ consecutive (contiguous) linear segments. The first J_t of these segments have interval *width* values $d_{1t}, d_{2t}, \ldots, d_{J_t t}$, and we let r_{jt} denote the slope (per unit revenue) within the j^{th} linear segment; the $(J_t + 1)^{\text{st}}$ segment has slope zero (i.e., the maximum possible total revenue is finite with value $\sum_{j=1}^{J_t} r_{jt} d_{jt}$ for $t = 1, \ldots, T$). This implies that we can state our revenue functions as follows:

$$R_t(D_t) = \begin{cases} \sum\limits_{j=1}^{k-1} r_{jt} d_{jt} + r_{kt} \left(D_t - \sum\limits_{j=1}^{k-1} d_{jt} \right) & \text{for } \sum\limits_{j=1}^{k-1} d_{jt} \leq D_t < \sum\limits_{j=1}^{k} d_{jt}, \\ & \quad k = 1, \ldots, J_t, \\ \sum\limits_{j=1}^{J_t} r_{jt} d_{jt} & \text{for } \sum\limits_{j=1}^{J_t} d_{jt} \leq D_t. \end{cases} \tag{3.8}$$

where $r_{1t} > r_{2t} > \cdots > r_{J_t t} > 0$. Theorem 1 implies that an optimal solution exists such that the total demand satisfied in each period t occurs at one of the breakpoint values; i.e., at $\sum_{j=1}^{k} d_{jt}$ for some k between one and J_t (note that an optimal demand value cannot exist in the $(J_t + 1)^{\text{st}}$ interval if costs are positive, which we assume throughout, since costs will increase and revenues remain constant). Denote such a value of D_τ by $D_\tau^*(t)$. Then,

$$c(t, t') = \max \left(0, \sum_{\tau=t}^{t'-1} \left(R_t(D_\tau^*(t)) - v_{t\tau} D_\tau^*(t) \right) - S_t \right).$$

The time needed to compute these values is $O(T^2)$ multiplied by the time required to find $D_\tau^*(t)$ and evaluate $R_t(D_\tau^*(t))$ for all t, τ. Note that if the functions $R_t(\cdot)$ are piecewise-linear and concave with at most $J_{\max}$ segments, the slopes at each breakpoint and the resulting $R_t(D_\tau^*(t))$ computations can be performed in $\mathcal{O}(J_{\max})$ time, for a total arc 'cost' calculation time of $\mathcal{O}(J_{\max} T^2)$. Since the acyclic longest path problem requires $\mathcal{O}(T^2)$ operations, our total solution time is no worse than $\mathcal{O}(J_{\max}T^2)$.

3.2.2 Dual-ascent Method for the Uncapacitated RPP

When the revenue functions are piecewise-linear and concave in every period, and production costs contain a fixed plus variable cost structure, we can also use a dual-based algorithm to solve the uncapacitated RPP, which we next describe. This approach requires first reformulating the RPP. As we later show, this new formulation is "tight"; i.e., its linear programming relaxation objective function value equals the optimal objective function value of RPP. We begin by providing an explicit base formulation of the uncapacitated RPP under piecewise-linear concave revenue functions and fixed plus linear production costs, using much of the notation already defined in the previous sections. We define a set of binary variables z_{jt} for $t = 1, \ldots, T$ and $j = 1, \ldots, J_t$, such that $z_{jt} = 1$ if $D_t \geq \sum_{k=1}^{j} d_{kt}$ (i.e., when the total demand satisfied in period t occurs at the j^{th} or higher breakpoint of the piecewise-linear concave revenue curve); otherwise $z_{jt} = 0$ when $D_t \leq \sum_{k=1}^{j-1} d_{kt}$.

By the definition of the z_{jt} variables and the fact that an optimal solution exists where total demand falls at an interval breakpoint in each period, we therefore have that the total demand satisfied in period t equals $D_t = \sum_{j=1}^{J_t} d_{jt}z_{jt}$, and the corresponding total revenue equals $\sum_{j=1}^{J_t} r_{jt}d_{jt}z_{jt}$. We next define a new set of binary setup variables, y_t, for $t = 1, \ldots, T$, where $y_t = 1$ if we perform a setup in period t, and $y_t = 0$ otherwise. We can thus formulate the uncapacitated RPP with piecewise-linear concave revenue functions, which we refer to as the RPP_{PLC}, as follows.

[RPP$_{PLC}$]

$$
\begin{aligned}
\text{maximize:}\quad & \textstyle\sum_{t=1}^{T}\left(\sum_{j=1}^{J_t} r_{jt}d_{jt}z_{jt} - S_t y_t - p_t x_t - h_t I_t\right) \\
\text{subject to:}\quad & I_{t-1} + x_t = \textstyle\sum_{j=1}^{J_t} d_{jt}z_{jt} + I_t && t = 1,\ldots,T, && (3.9)\\
& 0 \le x_t \le \left(\textstyle\sum_{\tau=t}^{T}\sum_{j=1}^{J_\tau} d_{j\tau}\right) y_t && t = 1,\ldots,T, && (3.10)\\
& I_0 = 0,\ I_t \ge 0, && t = 1,\ldots,T, && (3.11)\\
& 0 \le z_{jt} \le 1 && t = 1,...,T, && \\
& && j = 1,\ldots,J_t, && (3.12)\\
& y_t \in \{0,1\} && t = 1,...,T. && (3.13)
\end{aligned}
$$

In the above [RPP$_{PLC}$] formulation, the objective function provides the net revenue after subtracting production and holding costs. Constraint set (3.9) ensures inventory balance, while the setup forcing constraints (3.10) enforce setting y_t equal to one if any production occurs in period t. Note that the coefficient of y_t in these constraints equals the total demand from period t through T, thereby effectively leaving the problem uncapacitated. Constraints (3.11) through (3.13) encode our variable restrictions. Since an optimal solution exists for the uncapacitated version of the problem such that the demand satisfied in any period occurs at one of the breakpoint values of the period's revenue function, [RPP$_{PLC}$] provides the same optimal solution value as the formulation obtained by explicitly imposing the binary restriction on the z_{jt} variables. We formulate the problem with the relaxed binary restrictions, however, for later extension to the equal-capacity case in Section 3.2.4.

Note that we have not imposed any specific constraints on the relationship between z_{jt} variables corresponding to the same revenue function in a given period t. The following property allows us to consider each of the intervals of the piecewise-linear concave revenue function independently from one another in our mixed integer programming formulation (that is, we need not introduce any explicit constraints in our formulation that specify the strict ordering of the piecewise-linear segments of the revenue functions).

Property 1 ***Contiguity Property***: *For the [RPP$_{PLC}$] problem defined above, if an optimal solution exists such that $z_{j-1,t} = 0$, then $z_{kt} = 0$ for $k = j, \ldots, J_t$ in any optimal solution.*

Proof: Suppose that an optimal solution exists with objective function value Z^* with $z_{kt} = 0$ and $z_{lt} = 1$ for some $l > k$, and let period $s \leq t$ denote the setup period in which the production occurred that satisfied demand in period t. Since $z_{lt} = 1$, we must have that $r_{lt} \geq p_s + \sum_{\tau=s}^{t-1} h_\tau$; otherwise a solution exists such that $z_{lt} = 0$ with objective function value greater than Z^*, which contradicts the optimality of the solution with $z_{lt} = 1$. Since $r_{kt} > r_{lt}$ we must also have $r_{kt} > p_s + \sum_{\tau=s}^{t-1} h_\tau$, and a solution exists with $z_{kt} = 1$ and an objective function value greater than Z^*, a contradiction of the optimality of the solution with $z_{kt} = 0$ and $z_{lt} = 1$, which implies that if $z_{kt} = 0$ in an optimal solution z_{lt} must equal zero for $l = k + 1, \ldots, J_t$ in any optimal solution (i.e., the contiguity property). $\diamondsuit$

We can also use the arguments in the contiguity property proof to show that if $z_{kt} = 1$ in an optimal solution, then we must also have $z_{jt} = 1$ for $j = 1, \ldots, k - 1$. The contiguity property thus ensures that the quantities $\sum_{j=1}^{J_t} d_{jt} z_{jt}$ and $\sum_{j=1}^{J_t} r_{jt} d_{jt} z_{jt}$ correctly provide the total demand satisfied and the total revenue in period t, without the need to introduce any explicit dependencies among the z_{jt} variables in our mixed integer programming formulation.

While the [RPP$_{PLC}$] formulation correctly captures the RPP$_{PLC}$ problem we have defined, its linear programming relaxation value does not necessarily equal the optimal value of the RPP$_{PLC}$; i.e., its integrality gap is not necessarily zero. We next derive an equivalent problem formulation for which the integrality gap is indeed equal to zero. We show this by developing a dual-ascent algorithm for the dual of this formulation that provides an optimal dual solution whose complementary primal solution is feasible for all of the integer restrictions of the [RPP$_{PLC}$] formulation. We note that this approach generalizes a related approach for the ELSP developed by Wagelmans, van Hoesel, and Kolen [82]. Alternative approaches also include extending the proof techniques for the reformulated ELSP shown in Nemhauser and Wolsey [58], Barany, Van Roy, and Wolsey [10], and Barany, Edmonds, and Wolsey [9].

Starting with the [RPP$_{PLC}$] formulation, we can equivalently state the objective function as:

$$\text{minimize:}\quad \textstyle\sum_{t=1}^{T}\left(S_t y_t + p_t x_t + h_t I_t\right) - \sum_{t=1}^{T}\sum_{j=1}^{J_t} r_{jt} z_{jt} \tag{3.14}$$

Since $I_t = \sum_{\tau=1}^{t} x_\tau - \sum_{\tau=1}^{t}\sum_{j=1}^{J_\tau} d_{j\tau} z_{j\tau}$, we can eliminate the inventory variables from the formulation via substitution. We next introduce a new cost parameter, c_t, where $c_t \equiv p_t + \sum_{\tau=t}^{T} h_\tau$. The objective function of the RPP$_{PLC}$ can now be written as:

$$\text{minimize:}\quad \sum_{t=1}^{T}\left(S_t y_t + c_t x_t\right) - \sum_{t=1}^{T} h_t\left(\sum_{\tau=1}^{t}\sum_{j=1}^{J_\tau} d_{j\tau} z_{j\tau}\right) - \sum_{\tau=1}^{T}\sum_{j=1}^{J_\tau} r_{j\tau} d_{j\tau} z_{j\tau} \tag{3.15}$$

We next define ρ_{jt} as a modified revenue parameter for linear segment j in period t, where $\rho_{jt} = \sum_{\tau=t}^{T} h_\tau + r_{jt}$. The development of our dual-ascent procedure requires capturing the exact amount of production in each period that corresponds to the amount of demand satisfied within each linear segment of the piecewise-linear revenue function in the current and all future periods. We thus define $x_{jt\tau}$ as the number of units produced in period t corresponding to demand satisfaction within linear segment j in period τ, for $\tau \geq t$, and replace each x_t with $\sum_{\tau=t}^{T}\sum_{j=1}^{J_\tau} x_{jt\tau}$. We next provide a reformulation of the LP relaxation of the RPP$_{PLC}$, which we denote by [RPP$'_{PLC}$], that lends itself nicely to our dual-based approach.

[RPP$'_{PLC}$]

$$\text{minimize:} \sum_{t=1}^{T}\left(S_t y_t + c_t \sum_{\tau=t}^{T}\sum_{j=1}^{J_\tau} x_{jt\tau}\right) - \sum_{t=1}^{T}\sum_{j=1}^{J_t} \rho_{jt} d_{jt} z_{jt}$$

$$\begin{aligned}
\text{subject to:} \quad & \sum_{t=1}^{\tau} x_{jt\tau} - d_{j\tau} z_{j\tau} = 0 && \tau = 1, \ldots, T, j = 1, \ldots, J_\tau, && (3.16)\\
& d_{j\tau} y_t - x_{jt\tau} \geq 0 && t = 1, \ldots, T, && \\
& && \tau = t, ..., T, j = 1, \ldots, J_\tau, && (3.17)\\
& -z_{j\tau} \geq -1 && \tau = 1, \ldots, T, j = 1, \ldots, J_\tau, && (3.18)\\
& y_t, x_{jt\tau}, z_{jt} \geq 0 && t = 1, \ldots, T, && \\
& && \tau = t, \ldots, T, j = 1, \ldots, J_\tau. && (3.19)
\end{aligned}$$

Recall that we introduced a very similar formulation ([UOSP$'$]) in Chapter 2 for the purpose of developing heuristic solution approaches to the OSP problem. In this section, we disaggregate the setup forcing constraints (2.13) from [UOSP$'$] to arrive at the above formulation [RPP$'_{PLC}$].

Note that if $z_{jt} = 1$ in a solution we say that the *demand corresponding to segment j in period t is satisfied* in the corresponding solution. This manner of describing the solution will facilitate a clearer description of our formulation and the dual algorithm and solution that later follow. Constraints (3.16) ensure that if the demand in segment j in period τ is satisfied, then a production amount equal to this demand must occur in some period less than or equal to τ. If any production occurs in period t, constraint set (3.17) forces $y_t = 1$, thus allowing production in period t for segment j demand in period τ to equal any amount up to $d_{j\tau}$; otherwise, if $y_t = 0$, no production can be allocated to period t. Constraints (3.18) and (3.19) represent the (relaxed) variable restrictions. Note that since a positive cost exists for setups, we can show that the constraint $y_t \leq 1$ is unnecessary in the above relaxation, and so we omit this constraint from the relaxation formulation. It is straightforward to show that the [RPP'_{PLC}] formulation with the additional requirement that all y_t are binary variables is equivalent to our original RPP_{PLC}.

To formulate the dual of [RPP'_{PLC}], let $\mu_{j\tau}$, $w_{jt\tau}$, and $\pi_{j\tau}$ denote dual multipliers associated with constraints (3.16), (3.17), and (3.18), respectively. Taking the dual of [RPP'_{PLC}], we arrive at the following dual formulation [DP]:

[DP]

$$\begin{aligned}
\text{maximize:} \quad & \textstyle\sum_{\tau=1}^{T}\sum_{j=1}^{J_\tau} -\pi_{j\tau} & & \\
\text{subject to:} \quad & \sum_{\tau=t}^{T}\sum_{j=1}^{J_\tau} d_{j\tau} w_{jt\tau} \le S_t & & t = 1,\ldots,T, \qquad (3.20)\\
& \mu_{j\tau} - w_{jt\tau} \le c_t & & t = 1,\ldots,T,\\
& & & \tau = t,\ldots,T, j = 1,\ldots,J_\tau \qquad (3.21)\\
& -d_{j\tau}\mu_{j\tau} - \pi_{j\tau} \le -\rho_{j\tau} d_{j\tau} & & \tau = 1,\ldots,T, j = 1,\ldots,J_\tau, \qquad (3.22)\\
& \pi_{j\tau}, w_{jt\tau} \ge 0 \; ; \; \mu_{j\tau}\text{unrestricted} & & t = 1,\ldots,T,\\
& & & \tau = t,\ldots,T, j = 1,\ldots,J_\tau. \qquad (3.23)
\end{aligned}$$

Inspection of formulation [DP] indicates that we can set $w_{jt\tau}$ equal to the maximum between 0 and $\mu_{j\tau} - c_t$ without loss of optimality; similarly, an optimal solution exists with $-\pi_{j\tau}$ equal to the minimum between 0 and $d_{j\tau}(\mu_{j\tau} - \rho_{j\tau})$. The above formulation can, therefore, be re-written in a more compact form as:

[CDP]

$$\begin{aligned}
\text{maximize:} \quad & \sum_{\tau=1}^{T}\sum_{j=1}^{J_\tau} \min\left(0, d_{j\tau}(\mu_{j\tau} - \rho_{j\tau})\right) & \\
\text{subject to:} \quad & \sum_{\tau=t}^{T}\sum_{j=1}^{J_\tau} d_{j\tau}\left\{\max\left(0, \mu_{j\tau} - c_t\right)\right\} \le S_t \quad t = 1,\ldots,T. & (3.24)
\end{aligned}$$

We note some important properties of the [CDP] formulation. First, we have no incentive to set any $\mu_{j\tau}$ variable value in excess of $\rho_{j\tau}$, since any increase above this value does not affect the objective function value. Second, we can initially set each

$\mu_{j\tau} = \min_{t=1,\ldots,\tau}\{c_t\}$ for all $\tau = 1, \ldots, T$ and $j = 1, \ldots, J_\tau$, without utilizing any of the "capacity", S_t, in each constraint. We can also eliminate any segment-period combination (j, τ) such that $\min_{t=1,\ldots,\tau}\{c_t\} \geq \rho_{j\tau}$, since any demand satisfied within such a segment will never provide a positive contribution to profit. In describing our solution approach, we will refer to the constraint for period t in [CDP] as the t^{th} constraint (or constraint t) of the formulation. Our approach for solving [CDP] is to use a dual-ascent procedure that increases the dual variables in increasing time index order. That is, we increase the values of the μ_{j1} variables before we increase any μ_{jt} values for $t > 1$. We then focus on increasing the μ_{j2} variables, and so on. We begin by simultaneously increasing the value of all μ_{j1} variables. If for some segment l in period 1, μ_{l1} reaches a value of ρ_{l1} before constraint 1 becomes tight, we say that this segment "drops out" in period 1 and we do not further increase the value of μ_{l1} (i.e., μ_{l1} is fixed at ρ_{l1} in the solution). We then continue to increase all other μ_{j1} values until constraint 1 becomes tight. Let J_t^0 denote the set of all segments that drop out in period t, and let J_t^1 denote the set of all segments that do not drop out in period t. We define μ_1^* as the value of μ_{j1} for all segments that do not drop out in period 1, where

$$\mu_1^* = c_1 + \frac{S_1 - \sum_{j \in J_1^0} d_{j1} \max\left(0, \rho_{j1} - c_1\right)}{\sum_{j \in J_1^1} d_{j1}}.$$

Note that at this point, after determining μ_1^*, the first constraint of [CDP] is tight (assuming $J_1^1 \neq \emptyset$; we later discuss the necessary modifications if $J_1^1 = \emptyset$). We next focus on increasing the μ_{j2} variable values. When we increase the values of the μ_{j2} variables, these variables can be blocked from increase by either dropping out (i.e., when $\mu_{l2} = \rho_{l2}$ for some segment l), by tightening constraint 2, or by hitting the value c_1 (observe that no μ_{j2} value can be greater than c_1 since constraint 1 is already tight, and such a value would, therefore, violate constraint 1). Letting μ_2^*

equal the value of μ_{j2} for all $j \in J_2^1$, we have

$$\mu_2^* = \min\left\{c_1; c_2 + \frac{S_2 - \sum_{j\in J_2^0} d_{j2} \max\left(0, \rho_{j2} - c_2\right)}{\sum_{j\in J_2^1} d_{j2}}\right\}.$$

Applying this same approach in period 3 produces

$$\mu_3^* = \min\left\{\begin{array}{l} c_1;\ c_2 + \dfrac{S_2 - \sum\limits_{j\in J_2^0} d_{j2}\max(0,\rho_{j2}-c_2) - \sum\limits_{j\in J_2^1} d_{j2}\max\left(\mu_2^*-c_2,0\right) - \sum\limits_{j\in J_3^0} d_{j3}\max(0,\rho_{j3}-c_2)}{\sum\limits_{j\in J_3^1} d_{j3}}; \\ c_3 + \dfrac{S_3 - \sum\limits_{j\in J_3^0} d_{j3}\max(0,\rho_{j3}-c_3)}{\sum\limits_{j\in J_3^1} d_{j3}} \end{array}\right\},$$

or in general, for period τ:

$$\mu_\tau^* = \min_{i\leq\tau}\left\{c_i + \frac{S_i - \sum\limits_{t=i}^{\tau}\sum\limits_{j\in J_t^0} d_{jt}\max(0,\rho_{jt}-c_i) - \sum\limits_{t=i}^{\tau-1}\sum\limits_{j\in J_t^1} d_{jt}\max(\mu_t^*-c_i,0)}{\sum\limits_{j\in J_j^1} d_{j\tau}}\right\}. \tag{3.25}$$

Our final dual solution takes the form:

$$\mu_{j\tau} = \begin{cases} \rho_{j\tau}, & j \in J_\tau^0, \\ \mu_\tau^*, & j \in J_\tau^1, \end{cases} \quad \text{for } \tau = 1, \ldots, T, \text{ and } j = 1, \ldots, J_\tau.$$

Note that it is possible that the set J_τ^1 is empty for some τ after applying the algorithm, since all orders in period τ may drop out before hitting any of the constraints. In such cases μ_τ^* requires no definition. We can summarize this dual-ascent solution approach as follows:

CDP Dual-Ascent Solution Algorithm

0. Delete any segment-period combination $(j,\ \tau)$ such that $\min\limits_{t=1,\ldots,\tau}\{c_t\} \geq \rho_{j\tau}$.

1. Set $\mu_{j\tau} = \min\limits_{t=1,\ldots,\tau}\{c_t\}$ for all $\tau = 1,\ \ldots,\ T$ and $j = 1, \ldots, J_\tau$. Set iteration counter $k = 1$.

2. Let $J_k^0 = J_k^1 = \{\emptyset\}$. Simultaneously increase all μ_{jk} for $j = 1, \ldots, J_k$ from the initial value of $\min\limits_{t=1,\ldots,k}\{c_t\}$. If, while increasing the μ_{jk} values, some $\mu_{lk} = \rho_{lk}$ before the μ_{jk} values are blocked from increase by any constraint, fix μ_{lk} at ρ_{lk}, insert segment l into J_k^0, and continue to simultaneously increase μ_{jk} for all $j \notin J_k^0$ until some constraint $\gamma(k) \le k$ blocks the μ_{jk} values from further increase. When constraint $\gamma(k) \le k$ blocks the μ_{jk} values from further increase then, for all segments $j \notin J_k^0$, insert j into J_k^1 and set μ_{jk} using equation (3.25); i.e., set $\mu_{jk} = \mu_k^* = c_{\gamma(k)} + \frac{S_{\gamma(k)} - \sum\limits_{t=\gamma(k)}^{k} \sum\limits_{j \in J_t^0} d_{jt} \max\left(0, \rho_{jt} - c_{\gamma(k)}\right) - \sum\limits_{t=\gamma(k)}^{k-1} \sum\limits_{j \in J_t^1} d_{jt} \max\left(\mu_t^* - c_{\gamma(k)}, 0\right)}{\sum\limits_{j \in J_k^1} d_{jk}}$. (If all $j = 1, \ldots, J_k$ enter J_k^0 before some constraint becomes tight, then μ_k^* requires no definition.)

3. Set $k = k + 1$. If $k = T$, stop with dual feasible solution. Otherwise, repeat Step 2.

Note that in each period k we must check the value of ρ_{jk} for each segment $j = 1, \ldots, J_k$ and determine whether this value of μ_{jk} will tighten or violate any of the constraints $1, \ldots, k$. Since we need to apply this comparison for $k = 1, \ldots, T$, we can bound the complexity of this dual-ascent algorithm by $\mathcal{O}\left(J_{\max} T^2\right)$, the same as that of the shortest-path algorithm in the previous section. We next show that the dual-ascent solution procedure outlined above not only solves [CDP], but also leads to a primal complementary solution in which all of the binary restrictions in formulation [RPP$_{PLC}$] are satisfied; i.e., the dual-ascent procedure solves the RPP$_{PLC}$. Before showing this, we first need the following lemma.

Lemma 1 *For any pair of positive integers τ and l such that $\tau + l \le T$ and μ_τ^* and $\mu_{\tau+l}^*$ are defined as in the dual-ascent algorithm, we necessarily have $\mu_\tau^* \ge \mu_{\tau+l}^*$.*

Proof: Let $k < \tau$ be such that

$$\mu_\tau^* = c_k + \frac{S_k - \sum_{t=k}^{\tau} \sum_{j \in J_t^0} d_{jl} \max\left(0, \rho_{jt} - c_k\right) - \sum_{t=k}^{\tau-1} \sum_{j \in J_t^1} d_{jl} \max\left(\mu_t^* - c_k, 0\right)}{\sum_{J \in J_\tau^1} d_{j\tau}},$$

from which we can conclude that $\mu_\tau^* \ge c_k$ (since the numerator on the right hand side is the slack of constraint k, which must be nonnegative, since we maintain dual feasibility at all times). Next consider $\mu^*_{\tau+l}$:

$$\mu^*_{\tau+l} \leq c_k + \frac{S_k - \sum_{t=k}^{\tau+l} \sum_{j \in J_t^0} d_{jl} \max\left(0, \rho_{jt} - c_k\right) - \sum_{t=k}^{\tau+l-1} \sum_{j \in J_t^1} d_{jt} \max\left(\mu^*_t - c_k, 0\right)}{\sum_{j \in J^1_{\tau+l}} d_{j,\tau+l}}$$

Since μ^*_τ tightens constraint k, the quantity in the numerator above must be zero and we therefore have $\mu^*_{\tau+l} \leq c_k \leq \mu^*_\tau$ for all $\tau = 1, \ldots, T$ and $\tau + l \leq T$, since τ was chosen arbitrarily. $\diamondsuit$

Lemma 1 is required for proving the following result, the proof of which can be found in Appendix B of Section 3.5.

Theorem 2 *The dual-ascent algorithm presented above solves [CDP]. Moreover, the complementary primal solution to the dual solution produced by the algorithm satisfies the integrality restrictions of the RPP_{PLC} and therefore provides an optimal solution for the RPP_{PLC}.*

Theorem 2 implies that formulation [RPP_{PLC}] is tight, and we can easily find the solution value for the RPP_{PLC} using a linear programming solver. The algorithms we have developed, however, have better worst case complexity ($\mathcal{O}\left(J_{\max} T^2\right)$) than solution via linear programming. To provide some insight on the structure of the primal solution, given the dual solutions, we can show that the tight constraints in the dual solution correspond to periods in which we setup in the complementary primal solution. Further, if $\mu_{jk} = \mu^*_k$, then the demand in segment j in period k is satisfied using the setup corresponding to the constraint that blocked μ^*_k from further increase (i.e., period $\gamma(k)$ from Step 2 of the dual-ascent algorithm).

As was shown in Chapter 2, we cannot reduce this bound to $\mathcal{O}(T \log T)$, as Federgruen and Tzur [24] and Wagelmans, van Hoesel, and Kolen [82] do for the ELSP, since we cannot ensure that cumulative demand satisfied as we increase the number of periods in a problem instance is nondecreasing. See Section 2.2.2 for a presentation of a counterexample.

3.2.3 Polynomial Solvability for Other Production Costs and Price-Demand Curves

To this point we have made two sets of key assumptions that have facilitated providing polynomial-time solution methods for the uncapacitated RPP. The first of these assumptions relies on the production cost function taking a fixed-charge structure in each period, while the second assumes that a single price-demand curve exists in each period. We next explore the degree to which we can relax these assumptions, while retaining our ability to apply the polynomial-time solution methods we have presented. First we consider contexts in which multiple price-demand response curves exist in each period; this would correspond to contexts in which the producer has multiple available markets in which to sell its output, with each market having a unique response to price. We then consider the impacts of a piecewise-linear concave production cost structure (which may include a fixed setup cost) in each period.

3.2.3.1 Multiple price-demand curves

In this section we show that any uncapacitated RPP with multiple demand curves in a period can be reformulated as an RPP with only a single demand curve per period. We will show that this holds for general concave revenue functions and for piecewise-linear concave functions in particular. This implies that the piecewise-linear concavity of the revenue functions is preserved under the transformation from a multiple demand curve per period problem to a single demand curve per period problem. Suppose we now have M_t distinct revenue functions in period t, each corresponding to a distinct revenue *source*, and that D_{mt} is now the decision variable for the amount of demand we satisfy for revenue source m in period t; $R_{mt}(D_{mt})$ is the revenue function associated with revenue source m in period t (a revenue source may be an individual market or customer). We can rewrite the uncapacitated RPP as

$$\begin{array}{lll} \text{maximize} & \sum_{t=1}^{T}\sum_{m=1}^{M_t} R_{mt}(D_{mt}) - \sum_{t=1}^{T}(g_t(x_t) + h_t I_t) & \\ \text{subject to:} & \sum_{m=1}^{M_t} D_{mt} = D_t & t = 1, \ldots, T, \\ & D_t + I_t = x_t + I_{t-1} & t = 1, \ldots, T, \\ & x_t, I_t, D_t \geq 0 & t = 1, \ldots, T, \\ & D_{mt} \geq 0 & m = 1, \ldots, M_t, t = 1, \ldots, T. \end{array}$$

Now observe that, for a given choice of D_t, we will choose the demand quantities for each revenue source that yield the maximum profit. So the uncapacitated RPP is equivalent to

$$\begin{array}{lll} \text{maximize} & \sum_{t=1}^{T} \tilde{R}_t(D_t) - \sum_{t=1}^{T}(g_t(x_t) + h_t I_t) & \\ \text{subject to:} & D_t + I_t = x_t + I_{t-1} & t = 1, \ldots, T, \\ & x_t, I_t, D_t \geq 0 & t = 1, \ldots, T. \end{array}$$

where the *aggregate revenue function* for period t, $\tilde{R}_t(D_t)$, is defined through the following subproblem (SP) as

[**SP**]

$$\tilde{R}_t(D_t) \equiv \max \left\{ \sum_{m=1}^{M_t} R_{mt}(D_{mt}) : \sum_{m=1}^{M_t} D_{mt} = D_t; D_{mt} \geq 0, m = 1, \ldots, M_t \right\}.$$

The function $\tilde{R}_t(D_t)$ is concave (see Rockafellar [66] Theorem 5.4), and clearly $\tilde{R}_t(0) = 0$. It now also easily follows that if $R_{mt}(\cdot)$ is piecewise-linear and concave (and $R_{mt}(0) = 0$) for all m and t, $\tilde{R}_t(\cdot)$ is piecewise-linear and concave for all t (and $\tilde{R}_t(0) = 0$). This can be shown by ordering the slopes of all segments in a given period in decreasing order, and noting that the function $\tilde{R}_t(D_t)$ will "use" these segments in nondecreasing index order (or nonincreasing value order).

Observe that if the $R_{mt}(\cdot)$ functions are all everywhere differentiable, then the demand values selected for each revenue source in a given period t as a result of solving subproblem [SP] will be such that $R'_{1t}(D_{1t}) = R'_{2t}(D_{2t}) = \cdots = R'_{M_t t}(D_{M_t t})$. In other words, at the optimal demand level, the marginal revenue for each revenue source will be equal. Thus, if the revenue sources are distinct but have identical revenue functions, we will of course charge the same price to every revenue source.

3.2.3.2 Piecewise-linear concave production costs

We next consider the case in which the production cost function in each period is piecewise-linear concave and nondecreasing in the production volume in the period. Note that any nondecreasing piecewise-linear concave function can be viewed as the minimum of a number of fixed-charge functions. Therefore, if the production functions are piecewise-linear and concave with a finite number of segments, we can view this as a choice between a finite number of alternative production modes. It is easy to see that, in any period, we will of course only use a single production mode without loss of optimality.

We can write such a production cost function in the following form:

$$g_t(x) = \begin{cases} 0 & \text{if } x = 0, \\ \min_{k=1,\ldots,\ell_t} \{S_{kt} + p_{kt}x\} & \text{if } x > 0, \end{cases}$$

where k denotes an index for different production “modes”. Given a sequence of periods $t, \ldots, t'-1$ and positive production in period t, we now essentially also need to choose which of the ℓ_t cost functions (or production *modes*) to use. Given a production setup in period t only, the unit production plus holding cost associated with period τ ($\tau = t, \ldots, t'-1$) under production mode k equals $v_{kt\tau} \equiv p_{kt} + \sum_{s=t}^{\tau-1} h_s$. As with our previous analysis and development of the shortest path algorithm (see Theorem 1), the optimal quantity of demand satisfied in period

τ under production *mode* k using a setup in period t, which we denote by $D^*_{k\tau}(t)$, is then equal to any value of D such that $v_{kt\tau}$ is in the set of subgradients of $R_\tau(D)$ at D. Let

$$\Pi_k(t,t') = \sum_{\tau=t}^{t'-1} (R_t(D^*_{k\tau}(t)) - v_{kt\tau}D^*_{k\tau}(t)) - S_{kt}$$

and

$$c(t,t') = \max\left\{0, \max_{k=1,\ldots,\ell_t} \Pi_k(t,t')\right\}.$$

The value of $\Pi_k(t,t')$ provides the maximum profit possible in periods $t,\ldots,t'-1$ under production mode k assuming we satisfy demand amounts of $D^*_{k\tau}(t)$ for $\tau = t,\ldots,t'-1$. As a result, $c(t,t')$, as before, provides the maximum possible profit in periods $t,\ldots,t'-1$ assuming the only setup that can satisfy demand in these periods must occur in period t (if at all). We can therefore use the same shortest path graph structure as before (shown in Figure 2–2) with these modified arc length computations to determine an optimal solution. Note that due to the concavity of the production cost function, automatically, the production quantity corresponding to the best production mode k lies in the correct segment; i.e., the production costs have been computed correctly. The time required to find all arc profits is $O(LT^2)$ multiplied by the time required to find $D^*_{k\tau}(t)$ for some k,t,τ, where $L = \max_{t=1,\ldots,T} \ell_t$ is the maximum number of linear segments for any of the T piecewise-linear concave production cost functions. As this analysis shows, the case of piecewise-linear concave production cost functions can be handled in a straightforward manner, even under general concave revenue functions, without a substantial increase in problem complexity.

3.2.4 Production Capacities

This section considers a capacitated version of the RPP_{PLC} where production capacities are equal in all periods. In Chapter 2, we showed that RPP_{PLC} with time-varying finite production capacities is NP-Hard by demonstrating that it generalizes the capacitated lot sizing problem (CLSP). The special case of the

CLSP where production capacities are equal in every period, however, can be solved in polynomial time (see Florian and Klein [28]) with a complexity of $\mathcal{O}(T^4)$. Because of this, we next investigate whether the equal-capacity version of the RPP_{PLC} contains a similar special structure that we might exploit to solve this problem in polynomial time.

The polynomial solvability of the equal-capacity CLSP relies on characterizing so-called *regeneration intervals* (Florian and Klein [28]). A regeneration interval is characterized by a pair of periods, τ and τ' (with $\tau < \tau'$) such that $I_\tau = I_{\tau'} = 0$, and $I_{\tau+1}$, $I_{\tau+2}$, $\ldots$, $I_{\tau'-1} > 0$ in an optimal solution. An optimal solution therefore consists of a sequence of regeneration intervals (including the possibility of a single regeneration interval $(0, T)$). A *capacity constrained sequence* between periods $\tau+1$ and τ' is one in which $x_t = 0$ or C for all periods between (and including) $\tau+1$ and τ' except for at most one. For the equal-capacity CLSP, an optimal solution exists consisting of a capacity constrained sequence within each regeneration interval (see Florian and Klein [28]). Given any choice of demands in every period for the equal-capacity RPP_{PLC} problem, the resulting problem is an equal-capacity CLSP; thus, an optimal solution exists for the equal-capacity RPP_{PLC} problem that consists of capacity constrained production sequences within each of a set of consecutive regeneration intervals.

Let $D_{\tau\tau'} = \sum_{t=\tau+1}^{\tau'} d_t$ denote the total demand satisfied between periods $\tau+1$ and τ', where d_t is the demand satisfied in period t. If (τ, τ') comprises a regeneration interval, we know that total production in periods $\tau+1, \ldots, \tau'$ must equal $D_{\tau\tau'}$ (since $I_\tau = I_{\tau'} = 0$ and $D_{\tau\tau'}$ is the demand satisfied in periods $\tau+1, \ldots, \tau'$). Since at most one period contains production at a value other than 0 or C in a capacity-constrained sequence, we must have $D_{\tau\tau'} = kC + \epsilon$, where k is some nonnegative integer, and ϵ is the amount produced in the period in which we do not produce at 0 or C (assuming $D_{\tau\tau'}$ is not evenly divisible by C, in which case ϵ equals zero). So, given $D_{\tau\tau'}$, in each of the periods $\tau+1,\ldots,\tau'$, we either produce 0, ϵ, or C, with a production amount of ϵ in only one of the periods. We can easily determine both k and ϵ given $D_{\tau\tau'}$ and C; i.e., $\epsilon = D_{\tau\tau'} (\text{mod } C)$, and $k = \lfloor D_{\tau\tau'}/C \rfloor$. We then construct a shortest-path graph that contains a path for every feasible capacity-constrained production sequence between periods $\tau+1$ and τ'. Solving this shortest-path problem provides the minimum cost capacity constrained sequence for every (τ, τ') pair (with $\tau' > \tau$). Given a value of $D_{\tau\tau'}$ for every possible (τ, τ') pair, we can use this $\mathcal{O}(T^4)$ CLSP solution approach to solve the equal-capacity RPP$_{PLC}$. The challenge then lies in determining appropriate $D_{\tau\tau'}$ values for each possible (τ, τ') pair. To address this issue, we next show that the candidate set of $D_{\tau\tau'}$ values for each (τ, τ') pair can be limited to a manageable number of choices. Note that Loparic [49] provides a similar analysis for a lot-sizing model in which total revenue is linear in the amount of demand satisfied.[1]

Consider a regeneration interval (τ, τ'), and recall that by definition we must have $I_\tau = 0, I_j > 0$ for $j = \tau+1,\ldots,\tau'-1$, and $I_{\tau'} = 0$. The adjusted revenue parameter that we introduced in Section 3.2 (i.e., $\rho_{jt} = r_{jt} + \sum_{s=t}^{T} h_s$ for $\tau < t \leq \tau'$) will play an important role in the analysis that follows. We also let δ_{jt} denote the decision variable for the amount of demand within segment j in period t that we satisfy, and recall that at most d_{jt} units of demand exist within segment j in period t. The following lemma is important in developing a useful solution algorithm.

Lemma 2 *Suppose an optimal solution for RPP$_{PLC}$ contains a regeneration interval (τ, τ'), and suppose $\rho_{jt} \geq \rho_{it'}$ with $\tau < t, t' \leq \tau'$. If an optimal solution exists with $\delta_{jt} < d_{jt}$, an optimal solution also exists with $\delta_{it'} = 0$. Equivalently, if an*

[1] I would like to gratefully acknowledge the insightful comments and direction provided by Yves Pochet at the International Workshop on Optimization in Supply Chain Planning in Maastricht, The Netherlands June 2001, that significantly strengthened the material in this section.

optimal solution exists such that $\delta_{it'} > 0$, then an optimal solution also exists with $\delta_{jt} = d_{jt}$.

Proof: Consider the regeneration interval (τ, τ') and consider some $\rho_{jt} \geq \rho_{it'}$ with $\tau < t, t' \leq \tau'$. Suppose we have an optimal solution with $\delta_{it'} > 0$, and $\delta_{jt} < d_{jt}$. Since $I_t > 0$ for $t = \tau + 1, \ldots, \tau' - 1$, we can increase δ_{jt} by some $\epsilon > 0$ and decrease $\delta_{it'}$ by ϵ without changing the amount produced in each of the periods $\tau + 1, \ldots, \tau'$. In particular, if $t \leq t'$ we can set $\epsilon = \min\{d_{jt} - \delta_{jt}; \delta_{it'}; \min\{I_t, \ldots, I_{t'-1}\}\}$. The resulting change in objective function value equals $(r_{jt} - (r_{it'} - \sum_{s=t}^{t'-1} h_s))\epsilon = (\rho_{jt} - \rho_{it'})\epsilon \geq 0$, and we either have $\delta_{jt} = d_{jt}$, $\delta_{it'} = 0$, or $I_s = 0$ for some $s = t, \ldots, t'-1$ (in the later case, t and t' no longer belong to the same regeneration interval). Similarly, if $t' < t$, we take $\epsilon = \min\{v_{it'}; d_{jt} - \delta_{jt}\}$, and the resulting change in objective function value equals $((r_{jt} - \sum_{s=t'}^{t-1} h_s) - r_{it'})\epsilon = (\rho_{jt} - \rho_{it'})\epsilon \geq 0$, with either $\delta_{it'} = 0$ or $\delta_{jt} = d_{jt}$. $\diamondsuit$

Lemma 2 ensures that an optimal solution exists such that, within each regeneration interval (τ, τ'), there is at most one period from $\tau + 1$ to τ' in which demand will not be satisfied at a value equal to one of the breakpoints of the revenue function. The following additional lemma allows us to further reduce the number of potential values of $D_{\tau\tau'}$ that we must consider for a given regeneration interval.

Lemma 3 *An optimal solution exists for RPP_{PLC} containing consecutive regeneration intervals (τ, τ') where the production plan in periods $\tau + 1, \ldots, \tau'$ is one of the following types:*

(i) We produce 0 or C in every production period in the interval $\tau + 1, \ldots, \tau'$ with at most one $0 < \delta_{jt} < d_{jt}$ in the interval; or

(ii) We produce at a value of ϵ, with $0 < \epsilon < C$, in at most one production period in the interval $\tau + 1, \ldots, \tau'$ (and all other production levels are either 0 or C in this interval), with all δ_{jt} values equal to either 0 or d_{jt} within the interval.

Appendix C in Section 3.5 contains a proof of Lemma 3. Lemmas 2 and 3 taken together imply that a limited number of candidate optimal solutions must be considered for each possible regeneration interval (note that the number of possible regeneration intervals is bounded by $\mathcal{O}(T^2)$). Letting $J_{\max}$ denote the maximum number of linear segments of the revenue functions among all periods (i.e., $J_{\max} = \max_{s=1,\ldots,T}\{J_s\}$), Lemmas 2 and 3 lead to the following theorem:

Theorem 3 *The equal-capacity RPP_{PLC} problem can be solved in $\mathcal{O}(J_{\max}T^7)$ time.*

Proof: Consider a potential regeneration interval (τ, τ') containing n periods, and let $J(\tau, \tau')$ denote the total number of linear segments in periods $\tau + 1, \ldots, \tau'$. For potential regeneration interval (τ, τ') we sort $J(\tau, \tau')$ values of ρ_{jt}. Let this index sequence of sorted values be denoted by $\left(\tau_1, \tau_2, \ldots, \tau_{J(\tau,\tau')}\right)$ (i.e., $\rho_{\tau_1} \geq \rho_{\tau_2} \geq \ldots \geq \rho_{\tau_{J(\tau,\tau')}}$), where each index τ_i identifies a unique segment-period pair within the regeneration interval. For potential regeneration interval (τ, τ'), note that Lemma 2 implies that if δ_{τ_i} takes a value strictly between 0 and d_{τ_i} we must have $\delta_{\tau_{i+k}} = 0$ for $k = 1, \ldots, J(\tau, \tau') - i$.

Lemma 3 implies that within each potential regeneration interval (τ, τ') of length n we need to consider two types of solutions. The first type of solution produces a quantity of zero or C in each of the n periods. For this type of solution we will have at most one $\delta_{\tau_i} < d_{\tau_i}$ for $1 \leq \tau_i \leq J(\tau, \tau')$, with $\delta_{\tau_{i-k}} = d_{\tau_{i-k}}$ for $k = 1, \ldots, i - 1$ and $\delta_{\tau_{i+k}} = 0$ for $k = 1, \ldots, J(\tau, \tau') - i$. The choice of the segment τ_i such that $\delta_{\tau_i} < d_{\tau_i}$ (of which there are $J(\tau, \tau') + 1$ possible choices, including the choice to produce zero for all periods) uniquely determines the number of periods in which we must produce at full capacity and, therefore, the values of δ_{τ_i} for $i = 1, \ldots, J(\tau, \tau')$. This in turn determines fixed demand levels that must be satisfied in an equal-capacity lot sizing problem for the regeneration interval (τ, τ'), which is solvable in $\mathcal{O}(n^4)$ using Florian and Klein's [28] algorithm.

The second type of solution we must consider sets each δ_{τ_i} equal to zero or d_{τ_i} for all $i = 1, \ldots, J(\tau, \tau')$ and produces at a value strictly between zero and C in at most one period in the regeneration interval. The choice of the index τ_i such that $\delta_{\tau_{i-k}} = d_{j_{i-k}}$ for $k = 0, \ldots, i-1$ and $\delta_{\tau_{i+k}} = 0$ for $k = 1, \ldots, J(\tau, \tau') - i$ uniquely determines the number of periods in which production at full capacity is required, and the value of production required in the single period such that $x_t < C$. Again, there are $J(\tau, \tau') + 1$ possible choices, including the choice to produce zero for all periods. In total we must consider $2J(\tau, \tau') + 1$ potential values of the demand vector $\left(\delta_{\tau_1}, \delta_{\tau_2}, \ldots, \delta_{\tau_{J(\tau,\tau')}}\right)$ for each regeneration interval of length n, which implies that the number of potential demand vectors for an interval of length n is bounded by $\mathcal{O}(J_{\max} n)$. For each of these vector values, we solve an $\mathcal{O}\left(n^4\right)$ equal-capacity lot-sizing problem. So the time required to evaluate a regeneration interval of length n is $\mathcal{O}\left(J_{\max} n^5\right)$, which is clearly bounded by $\mathcal{O}\left(J_{\max} T^5\right)$. Since we have at most $\mathcal{O}\left(T^2\right)$ potential regeneration intervals, the total effort required is $\mathcal{O}\left(J_{\max} T^7\right)$. $\diamondsuit$

Note that, unlike the uncapacitated case, in the capacitated case it is now possible to choose an optimal demand level in any period that is strictly between breakpoint values of the revenue function. While this is not a critical distinction in the pricing setting, we must explicitly consider this factor in the order selection problem setting discussed in the following Section.

3.3 Pricing and Order Selection Interpretations

Although our discussion has centered on the concept of pricing, up to this point we have said little about the actual pricing decisions that result from our models. That is, since our model assumes a one-to-one correspondence between price and demand within a market in every period, we have worked solely with demand levels as decision variables. Recall that we assumed that this price-demand relationship in each period is represented by the function $D_t = \sum_{m=1}^{M_t} d_{mt}(\theta_{mt})$, or equivalently $\theta_{mt} = d_{mt}^{-1}(D_t)$, where θ_{mt} denotes the price offered to market m in period t, and $d_{mt}^{-1}(D_t)$ is determined by solving an optimization subproblem, as discussed in Section 3.2.3.1. Given a total demand satisfied of D_t in period t, we also assumed that a total revenue of $R_t(D_t)$ is realized, where $R_t(D_t)$ is a nondecreasing concave function of demand D_t.

Given this relationship between demand and revenue, we can interpret the actual prices paid for the units sold in at least two ways, depending on the model's intended application context. We use Figure 3–1 to illustrate two such interpretations. Figures 3–1(a) and 3–1(b) show identical piecewise-linear revenue curves with three segments and segment slopes $r_1 > r_2 > r_3$. In both cases, the total revenue achieved at the demand level D' equals $R(D') = r_1 d_1 + r_2 d_2 + r_3 d_3$. In Figure 3–1(a) we assume that a market exists with a total of d_1 units of demand, each of which is willing to pay an amount of r_1 per unit of the good, while a second market with a total of d_2 units of demand provides a revenue of r_2 per unit, and a third market contains d_3 units of demand with a revenue of r_3 per unit. In this case, the price paid for units of demand falling within a segment corresponds to the slope of the segment. This interpretation might apply when different market segments (e.g., geographical segments) actually pay different prices, and each of the d_j values corresponds to a given market m's total available demand value, d_{mt}, in period t; i.e., d_j represents a market "size" in Figure 3–1(a).

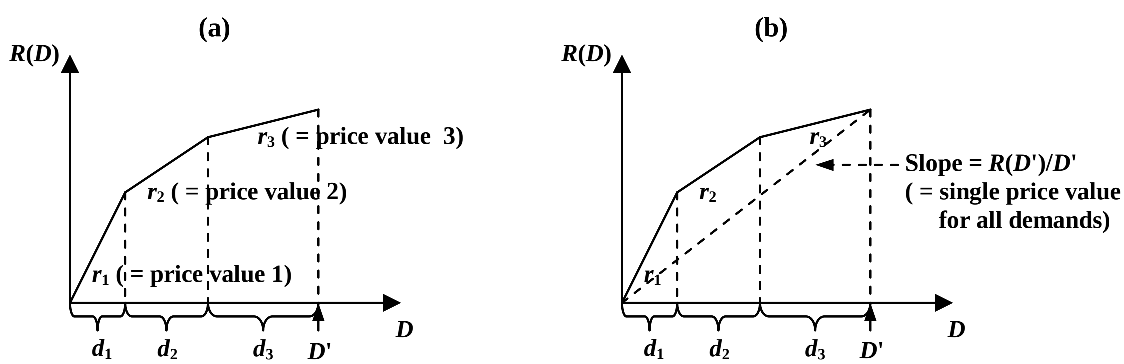


Figure 3–1: Pricing interpretations based on total revenue and demand.

In Figure 3–1(b), on the other hand, we assume that we have only a single market available, and all satisfied demands provide the same per-unit revenue (price), which at a demand level of D' is given by $\theta(D') = R(D')/D' = \sum_{j=1}^{3} r_j d_j / \sum_{j=1}^{3} d_j$. This interpretation implies that a total of D' demands exist that are willing to pay $\theta(D')$, which equals the slope of the line connecting the origin to $R(D')$. This interpretation applies to cases in which the supplier must charge a single price to all customers in the market. In either case, the models are completely the same, but the pricing interpretations and the contexts to which these interpretations apply are quite different. Note that when the revenue function is characterized by a differentiable concave function, the only practical interpretation is one in which the price paid for each unit equals the slope of the line connecting the origin to $R(D')$, which is $R(D')/D'$ ($R'(D')$ of course indicates the *marginal total* revenue at D').

Given our interpretation of Figure 3–1(a), we can now view the individual segments of the piecewise-linear revenue curve in a different light. That is, each linear segment may not only correspond to separate units of demand from an individual market, but might alternatively be associated with an aggregate order from an individual customer, where individual customers are willing to pay different unit prices for the item (or, alternatively, different customers have a different unit cost associated with fulfilling their orders). Given this interpretation, the RPP$_{PLC}$ model can be utilized in a broader set of contexts, where the firm does not set prices, but can select from a number of customer orders, each of which offers a particular net revenue per unit ordered.

Recall that, in the "order acceptance and denial" environment introduced in Chapter 2, firms can either commit to fulfilling an order or decline the order based on several factors, including the capacity to meet the order and the economic attractiveness of the order. The RPP$_{PLC}$ model can also be applied within such contexts. In this order selection setting, we now assume that a set of orders for the supplier's good exists in each of the T periods of the planning horizon, and redefine J_t as the number of orders that request fulfillment in period t. The index j now corresponds to individual order indices, and we let r_{jt} denote the revenue per unit provided by order j in period t, while d_{jt} is the order quantity associated with order j in period t. We index all orders within a period in nonincreasing order of unit revenues (i.e., $r_{1t} \geq r_{2t} \geq \cdots \geq r_{J_t t}$). We redefine the binary z_{jt} variables previously used in the [RPP$_{PLC}$] formulation as follows: $z_{jt} = 1$ if we accept order j in period t, and $z_{jt} = 0$ otherwise. These variables can now be interpreted as order selection variables. The remaining production quantity (x_t), production setup (y_t), and inventory (I_t) variables in the [RPP$_{PLC}$] formulation retain their original definition.

Since the formulation is completely unchanged except for our interpretation of the meaning of certain parameters and decision variables, we can use the same shortest path and dual-ascent methods we presented to solve this *order selection* problem. In the uncapacitated production setting, recall that an optimal solution exists for the RPP$_{PLC}$ problem such that the amount of demand selected in each period falls at one of the breakpoints of the revenue function. Under the order selection interpretation, this implies that an optimal solution exists in which every order will be either fully accepted (and fulfilled in its entirety), or will be declined.

We next briefly discuss the implications of finite capacity limits within the order selection context, again restricting our discussion to the equal-capacity case. Since the [RPP$_{PLC}$] formulation served as our starting point for the analysis of the equal-capacity case, and the uncapacitated order selection problem is formulated exactly the same as the [RPP$_{PLC}$] formulation, we can essentially follow the discussion in Section 3.2.4 with our new order selection interpretation. This approach assumes, however, that customers will permit partial order satisfaction;

i.e., for order j in period t we are free to satisfy any amount of the order quantity between zero and d_{jt}. For contexts in which such partial order satisfaction is allowed, we can therefore apply the same approach discussed in Section 3.2.4 to solve the equal-capacity version of the order selection problem. If, however, customers do not permit partial order satisfaction, the problem is NP-Hard.

To demonstrate the difficulty of the problem when partial order satisfaction is not allowed, we next briefly consider the single-period special case of this problem, where $T = 1$. Note that we now explicitly require the binary restrictions on the z_{j1} variables for this problem. For this single-period special case we can write the inventory balance constraints as $x_1 = \sum_{j=1}^{J_1} d_{j1} z_{j1}$, which simply implies that the production in the only period must equal the demand we choose to satisfy. Given that we have only a single-period problem, we will either perform a setup or not. If we do not perform a setup, then the objective function equals zero. If we do setup, then we need to solve the following problem to determine the optimal solution:

$$
\begin{aligned}
\text{maximize:} \quad & \sum_{j=1}^{J_1} (R_{j1} - p_1 d_{j1}) z_{j1} \\
\text{subject to:} \quad & \sum_{j=1}^{J_1} d_{j1} z_{j1} \leq C, \\
& z_{j1} \in \{0, 1\}, \qquad j = 1, \ldots, J_1.
\end{aligned}
$$

The above problem is a knapsack problem in its most general form (since the R_{j1} and d_{j1} parameters can take arbitrary nonnegative values), indicating that the all-or-nothing order satisfaction version of the capacitated problem problem is NP-Hard, even in the single-period special case (although the single-period version is not strongly NP-Hard). This problem is therefore clearly NP-Hard for the multiple-period case with or without equal capacities in all periods.

3.4 Conclusions

Allocating appropriate amounts of resources to anticipated demand sources has been a well-researched problem in revenue management, although the work has primarily focused on service industry applications (e.g., airline and hospitality industry applications; see, for example, van Ryzin and McGill [79]). As we have discussed, an increasing amount of attention is being placed on revenue management, through pricing models, in manufacturing contexts. We contribute to this effort by providing models and efficient solution methods for a general set of pricing problems in manufacturing settings where fixed setup costs comprise a substantial part of operations costs. In addition to pricing applications, we showed that our modeling approach also applies to order selection problems, the focus of Chapter 2, in which a supplier must choose from a set of outstanding orders to maximize its contribution to profit after production costs. As we have shown, our models and methods also allow for efficiently solving problems in which time-invariant finite production capacities exist.

Most revenue management literature addresses anticipated demand that is stochastic in nature, which is why selecting the best utilization of resources to achieve maximum profit is such a difficult problem. In the next chapter, we also consider the effects of stochastic demand on our demand selection decisions.

3.5 Appendix

Appendix A

Theorem 1 *For the uncapacitated RPP, given a production setup in period t only, if a demand quantity D_τ exists such that $v_{t\tau}$ is in the set of subgradients of $R_\tau(\cdot)$ at D_τ, then D_τ is an optimal demand quantity for the subproblem given by (3.4) and (3.5).*

Proof: Given a period $\tau \geq t$, and assuming a setup in period t only, we need to choose the demand quantity d in period t such that the total revenue at d in period t minus the total cost incurred in satisfying the quantity d in period t is maximized. By the definition of $v_{t\tau}$, the total cost (excluding the setup cost, which has already been incurred) for satisfying d units in period t using production in period t equals $v_{t\tau}d$. We need to therefore solve the following problem to determine the optimal demand value to satisfy in period τ:

$$\begin{aligned} \text{maximize:} \quad & R_\tau(d) - v_{t\tau}d \\ \text{subject to:} \quad & d \geq 0. \end{aligned}$$

Consider a value D such that $v_{t\tau}$ is in the set of subgradients of $R_\tau(d)$ at $d = D$. This implies by the definition of a subgradient of a concave function that $R_\tau(d) \leq R_\tau(D) + v_{t\tau}(d - D)$ for all $d \geq 0$ (the domain of $R_\tau(\cdot)$). This implies that $R_\tau(d) - v_{t\tau}d \leq R_\tau(D) - v_{t\tau}D$ for all $d \geq 0$, which implies the result. $\diamondsuit$

Appendix B

Theorem 2 *The dual-ascent algorithm presented in Section 3.2.2 solves [CDP]. Moreover, the complementary primal solution to the dual solution produced by the algorithm satisfies the integrality restrictions of formulation [RPP$_{PLC}$] and therefore provides an optimal solution for the RPP$_{PLC}$.*

Proof: Let $F(\tau)$ denote the optimal value of a problem consisting of periods 1, …, τ. As we have demonstrated through our shortest-path approach, the following recursive relationship holds for the RPP$_{PLC}$:

$$F(\tau) = \min_{i \leq \tau} \left\{ F(i-1) + \min \left\{ S_i + c_i \bar{d}_{i\tau} - \bar{\rho}_{i\tau} \right\} \right\},$$

where $\bar{d}_{i\tau} = \sum_{t=i}^{\tau} \sum_{j \in J^*(t,i)} d_{jt}$, $\bar{\rho}_{i\tau} = \sum_{t=i}^{\tau} \sum_{j \in J^*(t,i)} \rho_{jt} d_{jt}$, and $J^*(t,i) = \{j : \rho_{jt} \geq c_i\}$.

In our dual problem, the only variables contributing value to the objective function are those contained in the sets, J^1_τ for $\tau = 1, \ldots, T$. In other words, letting Z^T_D denote the objective function value of our dual solution for a T-period problem, we have

$$Z_D^T = \sum_{t=1}^{T} \sum_{j \in J_t^1} \min\left(0, d_{jt}(\mu_t^* - \rho_{jt})\right) = \sum_{t=1}^{T} \sum_{j \in J_t^1} d_{jt}\left(\mu_t^* - \rho_{jt}\right),$$

since $\rho_{jt} \geq \mu_t^*$ for all $j \in J_t^1$ by definition. To show the optimality of our dual-ascent procedure, we need to show that

$$F(\tau) = \sum_{t=1}^{\tau} \sum_{j \in J_t^1} d_{jt}\left(\mu_t^* - \rho_{jt}\right),$$

where we have $F(\tau) \geq \sum_{t=1}^{\tau} \sum_{j \in J_t^1} d_{jt}\left(\mu_t^* - \rho_{jt}\right)$ for all feasible μ_t^* by weak duality. We first show that $F(1) = \sum_{j \in J_1^1} d_{j1}\left(\mu_1^* - \rho_{j1}\right)$ and $F(2) = \sum_{t=1}^{2} \sum_{j \in J_t^1} d_{jt}\left(\mu_t^* - \rho_{jt}\right)$ directly, and then use induction to show the general result. For $\tau = 1$, the result is straightforward, since the final objective function after implementing the dual procedure is equal to $\sum_{j \in J_1^1} d_{j1}\left(\mu_1^* - \rho_{j1}\right)$. If J_1^1 is empty, then the objective function equals zero, which implies we do no setup and satisfy no demand. Otherwise,

$$\begin{aligned}
Z_D^1 &= \sum_{j \in J_1^1} d_{j1}\left(\mu_1^* - \rho_{j1}\right) \\
&= \sum_{j \in J_1^1} d_{j1}\left[c_1 + \frac{S_1 - \sum_{j \in J_1^0} d_{j1} \max\left(0, \rho_{j1} - c_1\right)}{\sum_{j \in J_1^1} d_{j1}}\right] - \sum_{j \in J_1^1} \rho_{j1} d_{j1} \\
&= S_1 + \sum_{j \in J_1^1} d_{j1}\left(c_1 - \rho_{j1}\right) + \sum_{j \in J_1^0} d_{j1} \max\left(0, \rho_{j1} - c_1\right) \\
&= S_1 + \sum_{j \in J_1^1} d_{j1}\left(c_1 - \rho_{j1}\right) + \sum_{j \in \bar{J}_1^0} d_{j1}\left(c_1 - \rho_{j1}\right) \\
&= S_1 + \sum_{j \in J_1^1 \cup \bar{J}_1^0} d_{j1}\left(c_1 - \rho_{j1}\right),
\end{aligned}$$

where $\bar{J}_1^0$ is the set of all $j \in J_1^0$ such that $\rho_{j1} \geq c_1$. We now have constructed a dual feasible solution with an objective function value equal to that of a primal feasible solution that sets up in period 1 and satisfies all demand for segments j such that $j \in J_1^1 \cup \bar{J}_1^0$, implying that this solution is optimal for the primal problem.

We next consider the case of $\tau = 2$. In this case we have $Z_D^2 = \sum_{t=1}^{2} \sum_{j \in J_t^1} d_{jt} \left(\mu_t^* - \rho_{jt} \right)$. If J_1^1 is empty, then we have a single-period problem (for period 2) and we can refer to the proof above for the case of $\tau = 1$. Suppose then that neither J_1^1 nor J_2^1 is empty. In the process of applying our dual-ascent algorithm, we encounter one of the two cases below:

Case I: $\mu_2^* = c_1$. This implies that constraint 2 does not become tight and further increases in μ_2^* are blocked by the first constraint. In this case the dual objective equals

$$Z_D^2 = \sum_{t=1}^{2} \sum_{j \in J_t^1} d_{jt} \left(\mu_t^* - \rho_{jt} \right) = S_1 + \sum_{j \in J_1^1 \cup \bar{J}_1^0} d_{j1} \left(c_1 - \rho_{j1} \right) + \sum_{j \in J_2^1} d_{j2} \left(c_1 - \rho_{j2} \right),$$

which equals the primal objective function value of a primal feasible solution that sets up in period 1 only and uses this setup to satisfy segments j in period 1 such that $j \in J_1^1 \cup \bar{J}_1^0$ and in period 2 such that $j \in J_2^1$.

Case II: $\mu_2^* = c_2 + \frac{S_2 - \sum_{j \in J_2^0} d_{j2} \max(0, \rho_{j2} - c_2)}{\sum_{j \in J_2^1} d_{j2}}$. This implies that constraint 2 becomes tight before μ_2^* reaches c_1 and further increases in μ_2^* are blocked by the second constraint. In this case the dual objective equals

$$Z_D^2 = \sum_{t=1}^{2} \sum_{j \in J_t^1} d_{jt} \left(\mu_t^* - \rho_{jt} \right) = S_1 + S_2 + \sum_{j \in J_1^1 \cup \bar{J}_1^0} d_{j1} \left(c_1 - \rho_{j1} \right) + \sum_{j \in J_2^1 \cup \bar{J}_2^0} d_{j2} \left(c_2 - \rho_{j2} \right),$$

where $\bar{J}_2^0$ is the set of all $j \in J_2^0$ such that $\rho_{j2} \geq c_2$. This value of Z_D^2 is equal to the primal objective function value of a primal feasible solution that sets up in periods 1 and 2 and satisfies demand in all segments j in period 1 such that $j \in J_1^1 \cup \bar{J}_1^0$ using the setup in period 1 and satisfies demand in all segments j in period 2 such that $j \in J_2^1 \cup \bar{J}_2^0$ using the setup in period 2. We have so far shown that $Z_D^\tau = F(\tau)$ for $\tau = 1$ and 2. We next use induction to show that this holds for all $\tau > 2$. Assume there is at least one attractive segment in some period τ; i.e., O_τ^1 exists for some $\tau \in \{1, \ldots, T\}$ (otherwise the optimal dual solution value equals zero and no demand is satisfied). For some $k \le \tau$ we must have

$$\mu_\tau^* = c_k + \frac{S_k - \sum_{t=k}^{\tau} \sum_{j \in J_t^0} d_{jl} \max(0, \rho_{jt} - c_k) - \sum_{t=k}^{\tau-1} \sum_{j \in J_t^1} d_{jl} \max(\mu_t^* - c_k, 0)}{\sum_{j \in J_\tau^1} d_{j\tau}}.$$

The τ-period objective function then becomes

$$\begin{aligned} Z_D^\tau &= \sum_{t=1}^{\tau} \sum_{j \in J_t^1} d_{jt}(\mu_t^* - \rho_{jt}) \\ &= \sum_{t=1}^{k-1} \sum_{j \in J_t^1} d_{jt}(\mu_t^* - \rho_{jt}) + \sum_{t=k}^{\tau-1} \sum_{j \in J_t^1} d_{jt}(\mu_t^* - \rho_{jt}) + \sum_{j \in J_\tau^1} d_{j\tau}(\mu_\tau^* - \rho_{j\tau}) \end{aligned}$$

Substituting for μ_τ^*, the last expression can be written as:

$$\begin{aligned} &\sum_{j \in J_\tau^1} d_{j\tau}(\mu_\tau^* - \rho_{j\tau}) \\ &\quad = \sum_{j \in J_\tau^1} d_{j\tau} c_k + S_k - \sum_{t=k}^{\tau} \sum_{j \in J_t^0} d_{jl} \max(0, \rho_{jt} - c_k) - \sum_{t=k}^{\tau-1} \sum_{j \in J_t^1} d_{jl} \max(\mu_t^* - c_k, 0) - \sum_{j \in J_\tau^1} \rho_{j\tau} d_{j\tau} \end{aligned}$$

Returning to the τ-period objective function, and using our induction hypothesis, we now have

$$\begin{aligned} Z_D^\tau = F(k-1) + \sum_{j \in J_\tau^1} d_{j\tau} c_k + S_k + \sum_{t=k}^{\tau-1} \sum_{j \in J_t^1} d_{jt}(\mu_t^* - \rho_{jt} - \max(\mu_t^* - c_k, 0)) \\ - \sum_{t=k}^{\tau} \sum_{j \in J_t^0} d_{jt} \max(0, \rho_{jt} - c_k) - \sum_{j \in J_\tau^1} \rho_{j\tau} d_{j\tau}. \end{aligned}$$

Since for $t < \tau$, if μ_t^* is defined (i.e., $J_t^1 \neq \{\emptyset\}$), we have $\mu_t^* \ge \mu_\tau^* \ge c_k$ (from Lemma 1), the above can be rewritten as

$$\begin{aligned} Z_D^\tau &= F(k-1) + \sum_{j \in J_\tau^1} d_{j\tau}(c_k - \rho_{j\tau}) + S_k \\ &\quad + \sum_{t=k}^{\tau-1} \sum_{j \in J_t^1} d_{jt}(c_k - \rho_{jt}) - \sum_{t=k}^{\tau} \sum_{j \in J_t^0} d_{jt} \max(0, \rho_{jt} - c_k) \\ &= F(k-1) + S_k + \sum_{t=k}^{\tau} \sum_{j \in J_t^1} d_{jt}(c_k - \rho_{jt}) + \sum_{t=k}^{\tau} \sum_{j \in \bar{J}_t^0} d_{jt}(c_k - \rho_{jt}), \end{aligned}$$

where $\bar{J}_t^0 = \{j \in J_t^0 : \rho_{jt} \geq c_k\}$.

From our previous definitions, we can now simplify Z_D^τ to

$$Z_D^\tau = F(k-1) + S_k + c_k \bar{d}_{k\tau} - \bar{\rho}_{k\tau},$$

which corresponds to the objective function value of a primal feasible solution, implying that $Z_D^\tau \geq F(\tau)$. But by weak duality we have $Z_D^\tau \leq F(\tau)$, and so we must have $Z_D^\tau = F(\tau)$, the optimal solution value of the primal. Moreover, the complementary primal solution is also feasible for the binary restrictions of [RPP$_{PLC}$]. $\diamondsuit$

Appendix C

Lemma 3 *An optimal solution exists for RPP$_{PLC}$ containing consecutive regeneration intervals (τ, τ') where the production sub-plan in periods $\tau+1, \ldots, \tau'$ is of one of the following types:*

(i) We produce 0 or C in every production period in the interval $\tau+1, \ldots, \tau'$ with at most one $0 < \delta_{jt} < d_{jt}$ in the interval; or

(ii) We produce $0 < \epsilon < C$ in at most one production period in the interval $\tau+1, \ldots, \tau'$ (and all other production levels are either 0 or C in this interval), with all δ_{jt} values equal to either 0 or d_{jt} within the interval.

Proof: We have shown that an optimal solution exists containing a sequence of regeneration intervals, and that at most one δ_{jt} value exists in a regeneration interval with $0 < \delta_{jt} < d_{jt}$ (Lemma 2); we also know that a capacity-constrained production sequence exists. We therefore need to show that given an optimal solution satisfying these properties with a production quantity x_s within a regeneration interval such that $0 < x_s < C$, and with a δ_{jt} for some period t in the regeneration interval such that $0 < \delta_{jt} < d_{jt}$, an optimal solution also exists satisfying the conditions stated in the lemma. Suppose that we do have such an optimal solution, and that the production period s occurs prior to or including the demand period

t. Since inventory in each period in the regeneration interval is positive, a feasible solution exists for the regeneration interval that uses the same setup periods and reduces δ_{jt} by one unit, along with inventory in periods $s, \ldots, t-1$, and production in period s. Since this solution does not improve over our optimal solution (and given the linearity of costs), this implies that at least as good a solution exists that increases δ_{jt} by one unit, along with inventory in periods $s, \ldots, t-1$ and production in period s. Repeating this argument until either $x_s = C$ or $\delta_{jt} = d_{jt}$ implies the result of the lemma. Similarly, if period t is before period s, a feasible solution exists for the regeneration interval that uses the same setup periods, increases δ_{jt} by one unit, reduces inventory in periods $t, \ldots, s-1$ by one unit, and increases x_s by one unit. Since this solution does not improve over our optimal solution, this implies that at least as good a solution exists that reduces δ_{jt} by one unit, increases inventory in periods $t, \ldots, s-1$ by one unit, and reduces x_s by one unit. Repeating this argument until either $\delta_{jt} = 0$ or $x_s = 0$ proves the result. $\diamondsuit$

CHAPTER 4
SELECTING MARKETS UNDER DEMAND UNCERTAINTY

4.1 Introduction

Thus far, our approach to selecting the best demand sources (orders, markets, etc.) to satisfy relied on deterministic information concerning the size and timing of each demand. While we usually have some data for planning purposes, typically via scheduled orders or demand forecasts, the exact amounts are often inaccurate. Therefore, it is extremely important for a firm making such product ordering (or manufacturing) decisions to account for the stochastic nature of demand. As demand becomes less predictable, our selection decisions will surely be influenced. We study the market selection problem with demand uncertainty in order to develop a robust modeling approach that addresses such types of demand.

The classic newsvendor problem has been studied extensively in research literature due in large part to its industry applications. The retail and airline industries have shown that operating with a perishable good (e.g., seasonable fashion items, airline seats or flights) requires the attention of a single selling season model, which is addressed through the newsvendor model. In a similar vein, manufacturing firms are producing items with ever-decreasing product lives, in an effort to stay competitive with the latest offering of other firms. This is especially true in the technology sector where, by the time a firm starts to realize demand during the selling season, it is often too late to place a second order with a supplier due to long lead times. In other words, the firm must live with its previous order quantity decision and now possibly pay a premium for expediting additional product to capture any additional unforeseen demand.

No matter how much effort is spent on trying to reduce product and process lead times, certain industries will likely exist where obtaining materials or more product at a reasonable unit cost will require a substantial amount of time. Even if the firm operates in a so-called Quick Response (QR) mode with its suppliers, the lead times may still be long relative to the selling season (see Iyer and Bergen [39] for a discussion of QR in the apparel industry). This leads us to study questions concerning integrated order quantity and market selection decisions under uncertain demand.

We consider a firm that offers a product for a single selling season. The firm uses an overseas or "long lead time" supplier as the primary source for its product, and thus the order quantity must be decided far in advance of actual sales. The firm has the flexibility to select which market demand sources to satisfy, where each demand source is a random variable. In the classic newsvendor model, the preferred order quantity is dependent on the distribution of total demand. However, in our context, the demand distribution is dependent on the markets the firm selects. Thus, the market selection decision must be made prior to ordering from the firm's supplier so that an appropriate order can be received in time for the selling season. In addition to each market's demand distribution being random, we assume that this distribution can be influenced by the level of advertising effort used within each market. By expending more effort in a market, the firm can increase the demand for its product. We address appropriate advertising response functions which measure marketing effectiveness based on the level of advertising spending (see Vakratsas, Feinberg, Bass, and Kalyanaram [78]). We also examine the effect of budgetary constraints. The marketing budget could prevent the firm from capturing additional expected market demand, and, thus, additional profits, regardless of the firm's ordering or production capacity.

As product life cycles continue to decrease and assessing demand risk for market entry becomes increasing critical, many companies find themselves faced with similar issues that we address here. Claritas, a market research and strategic planning company, has cited several clients, including Eddie Bauer, that wanted better knowledge of their customers in order to minimize demand risk. Claritas has had many success stories in identifying profitable customers, assessing potential markets, and ranking opportunities. Recently, Fisher, Raman, and McClelland [27] studied how 32 leading retailers, all of which offer short-life-cycle products (some with a single selling season) with unpredictable demand, to determine how effectively each company used available data sources to understand their customers. In the present marketplace, these retailers are saying that they must make better use of demand information if they want to make profitable market selections. Finally, Carr and Lovejoy [20] also discuss this problem's motivation from an inverse newsvendor point-of-view. They cite a client firm making industrial products, and this firm desires a marketing strategy that selects appropriate demands or markets to enter while working within a fixed production level.

In contrast to the approach developed by Carr and Lovejoy [20], we do not assume a predetermined capacity limit that the firm must obey when selecting markets. Rather, our model jointly determines the capacity acquired and the markets selected in order to maximize a firm's profit. Moreover, the firm can influence market demands through judicious use of advertising resources. The resulting models lead to interesting new nonlinear and integer optimization problems, for which we develop tailored solution methods. These models also allow us to develop insights regarding the parameters and tradeoffs that are influential in integrated market selection and capacity acquisition decisions for items with a single selling season. Thus, this work provides new contributions to the operations modeling and management literature as well as the literature on operations research methodologies.

Many researchers have contributed to the wide range of literature that exists on stochastic inventory control, for which Porteus [62] provides a nice overview. Particularly relevant to our work is the literature that focuses on the newsvendor problem. In addition to the work by Porteus, reviews by Tsay, Nahmias, and Agrawal [76] and Cachon [18] provide more recent research directions concerning supply chain contracts and competitive inventory management in the context of a single-period "newsvendor" setting.

Additional literature considers the multi-item newsvendor problem as well, for which we can draw some similarities to our "multiple-market" setting. In our problem we have one cost for producing a single product, and the individual market delivery costs, sales and advertising costs, and revenues provide differentiation among markets. Each market has a certain amount of random demand, and we attempt to satisfy the demand from the markets we select to maximize overall profit. In the typical multiple-product setting, each product has unique production, salvage, ordering and perhaps distribution costs. However, there is no differentiation between the demand sources for a particular item. Moon and Silver [55] present heuristic approaches for solving the multi-item newsvendor problem with a budget constraint.

Other researchers have investigated how production capacity can be adjusted within the framework of a newsvendor-type problem. Fine and Freund [25] consider cost-flexibility tradeoffs in investing in product-flexible manufacturing capacity. They formulate the capacity investment decision as a two-stage stochastic program, where all future production and inventory decisions are rolled into one future period. This is notably different from our approach in that they consider production capacity constrained problems as opposed to budget constrained problems. They improve profitability by increasing capacity and then allocating it appropriately, while we improve profitability by working within the advertising budget, selecting specific markets to serve, and expending advertising effort in these markets to achieve the appropriate level of expected demand.

Yield management, which has been an active field of research for a long time, is also closely related to our study of market selection decisions. Given a set of random demands (or demand distributions) and selling prices within each demand segment, a firm will use yield management techniques to determine the amount of product to offer at each selling price so as to maximize overall profit. Some of the most notable yield (or revenue) management research focuses on pricing and seat allocation decisions within the airline industry. Belobaba [12] and Williamson [84] provide nice reviews of this topic. Other excellent papers on the effects of pricing and yield management include Gallego and van Ryzin [30], [31], Biller, Chan, Simchi-Levi and Swann [14], Petruzzi and Dada [59], and Monahan, Petruzzi, and Zhao [54]. The more recent work by Monahan et al. [54] focuses on the parallel between the dynamic pricing problem and the dynamic inventory problem. They offer a reinterpretation of the dynamic pricing problem as a price-setting newsvendor problem with recourse, which leads to insights into the actions and profits of a price-setting newsvendor.

Additional demand or market selection decisions can be found in the game theory literature concerning market entry, as discussed in Rhim, Ho, and Karmarkar [63]. When faced with competition for any market demand segment, a firm must make selections concerning production sites, capacities and quantities. They focus on a multi-firm approach, where each competitor's decisions will be affected by the timing of entry of the other competitors, and each competitor's level of entry. Our current research does not consider multi-firm decisions, and this could be an interesting extension to the single-firm decisions we present in this dissertation.

Our work is based largely on the relationship between expected revenue and demand uncertainty, which is directly correlated with the area of portfolio optimization. The seminal work by Markowitz [53] over 50 years ago, followed by countless articles in this stream of research, introduced the concept of mean-variance optimization. The mean-variance approach attempts to achieve a desired rate of return while minimizing the risk involved with obtaining that return. We determine an optimal set of markets based on their expected revenues (or returns) and the associated demand uncertainties (return risk). Our modeling approach differs in that we do not place a minimum level on expected profit, nor do we focus on risk minimization while meeting a desired profit level. For a more recent review of portfolio optimization and risk aversion, see Brealy and Myers [16].

Some of the more closely-related stochastic work on demand and order selection are Petruzzi and Monahan [60], Carr and Duenyas [19], and Carr and Lovejoy [20]. Petruzzi and Monahan [60] address selecting between two sources of demands, the primary market and the secondary (or outlet store) market, for which to supply product. While these demands might occur simultaneously, the firm must decide the preferred time to move the product to the outlet store market. Carr and Duenyas [19] consider a sequential production system that receives demand for both make-to-stock and make-to-order products. A contractual obligation exists to produce the make-to-stock demand, and the firm can supplement its production by accepting additional make-to-order demand sources. They approach this problem using queueing theory, in an effort to provide an optimal admission of make-to-order demand and overall sequencing of production jobs. Beyond the Carr and Duenyas [19] paper on joint admission control and production decisions, the most relevant work is found in Ha [35], which presents a queueing theory approach to stock rationing across several demand classes for a single-item, make-to-stock production system.

Carr and Lovejoy [20] examine an inverse newsvendor problem, which optimally chooses a demand distribution based on some pre-defined order quantity or capacity set by a supplier. They select a demand distribution from a set of feasible demand portfolios, which may include several customer classes. To create a demand portfolio, they select these customer classes, each of which has random demand that follows a normal demand distribution, and determine the amount of demand to satisfy within each class while not exceeding the pre-defined capacity. Because they consider an inverse newsvendor, there is no decision to make in setting the order quantity. Furthermore, they assume that all customer classes within each portfolio have already been ranked by some exogenous criteria, and demand is allocated in such a way that higher priority classes are filled completely before lower priority classes are considered. Our decision process is different for several key reasons. We simultaneously select the customer demands to satisfy and determine an appropriate order quantity to request from the supplier, making the order quantity a decision variable. Since the ranking of all demands may change based on the available funds for marketing, we cannot provide an a priori ranking of demands, but allow the model to implicitly determine the most attractive set of markets. We also require that all unmet demand from these sources will be expedited, ensuring that the demand of all "selected" customers is fulfilled.

First, we define and formulate our selective newsvendor problem in Section 4.2. Given a set of demand distributions and no budgeting or capacity constraints in place, we provide a demand selection algorithm that determines the markets to penetrate to maximize profit. Several managerial insights are also provided. In Section 4.3, we evaluate the effect that marketing plays in increasing the expected value and variance of demand within any individual demand source. We address several functional forms of the advertising response function, assuming an unlimited advertising budget. We then present the general selective newsvendor problem with limited market resources in Section 4.4, as well as a solution approach based on the problem's KKT conditions. In Section 4.5, we first provide computational results that illustrate the benefits of using the basic selective newsvendor problem. Then, we examine the solution to the limited resources problem given several forms of marketing effort. After presenting these results, we finally offer some insight into a few other modeling considerations presented in Section 4.6.

4.2 The Selective Newsvendor Problem

4.2.1 Problem Formulation and Solution Approach

Assume we have a set of n potential markets that a supplier can serve. Denote r_i as the per unit revenue of the item in market i, where $i = 1, \ldots, n$. Let D_i denote the random variable for demand from market i, where D_i has pdf and cdf $f_i(D_i)$ and $F_i(D_i)$ with mean μ_i and variance σ_i^2. We also assume that market demands are statistically independent. The firm must decide far in advance of the selling season both the actual markets it will serve (and thus in which markets it will apply marketing effort prior to the selling season[1]) and the total quantity Q it will procure from the overseas supplier at a per unit cost of c. The fixed cost of entering market i is S_i.

We assume without loss of generality that r_i is net of any market-specific unit variable cost due to production or delivery of the item. We can then assume, again without loss of generality, that $r_i > c$, otherwise it would be unprofitable to enter market i, and we could immediately eliminate the market from consideration. We assume that, given a set of selected markets, the firm must ultimately satisfy all

[1] Note that we initially assume that the firm will apply a fixed amount of selling effort in each market selected, and that this amount of selling effort determines the market's demand distribution. That is, initially, marketing effort is not a decision variable. These fixed amounts of marketing effort are also independent across all markets.

realized demand in these markets, and that a high-cost domestic supplier exists from which the firm can expedite units of the good (after observing demand) at a per unit cost of e, where $e > c$. Any unsold items remaining at the end of the selling season will be sold at a salvage value of v per unit, where $c > v$. The firm wishes to maximize net profit from its market selection decision for the single selling season.

In addition to the order quantity decision variable Q, the firm must decide (before placing the order for Q units) the markets it will satisfy. Let $y_i = 1$ if the firm decides to satisfy demand in market i, and 0 otherwise. Given a binary vector of market selection variables y, let $D^y = \sum_{i=1}^n D_i y_i$ denote the total demand of the selected markets, and denote its pdf by f_y and its cdf by F_y. It is easy to see that the total selected demand has mean $E(D^y) = \sum_{i=1}^n \mu_i y_i = \mu_y$ and variance $\mathrm{Var}(D^y) = \sum_{i=1}^n \sigma_i^2 y_i = \sigma_y^2$. We can then express the firm's expected profit as a function $G(Q, y)$ of the order quantity Q and the binary vector of market selection variables y:

$$G(Q, y) = \sum_{i=1}^n (r_i \mu_i - S_i) y_i - cQ + v \int_0^Q (Q - x) f_y(x) \mathrm{d}x - e \int_Q^\infty (x - Q) f_y(x) \mathrm{d}x.$$

For a given vector y, the profit function $G(Q, y)$ is concave, and maximizing the profit is equivalent to minimizing the cost in the associated newsvendor problem. This then yields an optimal order quantity as a function of y, say Q_y^*, given by

$$F_y(Q_y^*) = \frac{e - c}{e - v}.$$

Assuming that $F_y(Q_y^*)$ is invertible, we have

$$Q_y^* = F_y^{-1}(\rho), \tag{4.1}$$

where $\rho = \frac{e-c}{e-v}$.

To facilitate our later analysis of the optimal market selection decisions, we define $\Lambda_y(Q)$ as the *loss function* for a given order quantity Q and market selection vector y; i.e., $\Lambda_y(Q) = \int_Q^\infty (x-Q) f_y(x)\mathrm{d}x$. Using this notation, the firm's expected profit can be written as

$$G(Q_y^*, y) = \sum\nolimits_{i=1}^n (r_i\mu_i - S_i)y_i - (c-v)Q_y^* - v\mu_y - (e-v)\Lambda_y(Q_y^*).$$

The form of the loss function $\Lambda_y(Q_y^*)$ depends on the distribution of D^y, and in general, can be quite difficult to characterize. If each market's demand is normally distributed, we can easily characterize the distribution of D^y (it is also normally distributed), and we can also employ the *standard normal loss function*, $L(z) = \int_z^\infty (u-z)\phi(u)du$ (where $\phi(u)$ is the p.d.f. of the standard normal distribution, with c.d.f. $\Phi(u)$) to simplify our analysis. That is, if D^y is normally distributed, then we can write $\Lambda_y(Q_y^*) = \sigma_y L(z(\rho))$, where $z(\rho) = \frac{Q_y^* - \mu_y}{\sigma_y} = \Phi^{-1}(\rho)$ is the standard normal variate value associated with the fractile ρ. Moreover, assuming normally distributed demand, we can rewrite our optimal order quantity equation (4.1) as

$$Q_y^* = \sum\nolimits_{i=1}^n \mu_i y_i + z(\rho)\sqrt{\sum\nolimits_{i=1}^n \sigma_i^2 y_i}. \tag{4.2}$$

Using equation (4.2) and the identity $\Lambda_y(Q) = \sigma_y L(z)$, under normally distributed demand, we can write the firm's expected profit as

$$G(Q_y^*, y) = \sum\nolimits_{i=1}^n \left[(r_i - c)\mu_i - S_i\right] y_i - \{(c-v)z(\rho) + (e-v)L(z(\rho))\}\sqrt{\sum\nolimits_{i=1}^n \sigma_i^2 y_i}.$$

Note that given the cost parameters c, v, and e, the coefficient of the square root term is a nonnegative constant. Letting $K(c,v,e)$ equal this coefficient (i.e., $K(c,v,e) = \{(c-v)z(\rho) + (e-v)L(z(\rho))\}$), we can rewrite the expected profit equation as

$$G(Q_y^*, y) = \sum\nolimits_{i=1}^{n} ((r_i - c)\mu_i - S_i)y_i - K(c, v, e)\sqrt{\sum\nolimits_{i=1}^{n} \sigma_i^2 y_i}.$$

Now define $\bar{r}_i = (r_i - c)\mu_i - S_i$ as the total expected net revenue from serving market i with the overseas supplier, after including market entry costs. To maximize the firm's expected profit, we must solve the following selective newsvendor problem (SNP):

[SNP]

$$\text{maximize} \quad \sum\nolimits_{i=1}^{n} \bar{r}_i y_i - K(c, v, e)\sqrt{\sum\nolimits_{i=1}^{n} \sigma_i^2 y_i} \tag{4.3}$$

$$\text{subject to:} \quad y_i \in \{0, 1\} \qquad i = 1, \ldots, n. \tag{4.4}$$

Based on a similar approach first defined in Shen, Coullard, and Daskin [71], we can solve this problem using a simple sorting scheme and a selection algorithm that we next describe. We first sort markets in *nonincreasing order* of the ratio of the coefficient of y_i in the first (linear) term in the objective function, to the coefficient of y_i in the second (square root) term in the objective function. This results in indexing the markets such that

$$\frac{\bar{r}_1}{\sigma_1^2} \geq \frac{\bar{r}_2}{\sigma_2^2} \geq \cdots \geq \frac{\bar{r}_n}{\sigma_n^2}.$$

Note that the numerator of the ratio $\bar{r}_i/\sigma_i^2$ is the expected net revenue, while the denominator reflects the uncertainty in market demand. We will therefore refer to this ratio generically as the expected revenue to uncertainty ratio for a market i. The following property allows us to apply an efficient market selection algorithm.

Property 2 ***Decreasing Expected Revenue to Uncertainty (DERU) Ratio Property****: After indexing markets in decreasing order of expected net revenue to uncertainty, an optimal solution to [SNP] exists such that if we select customer l, we also select customers $1, 2, \ldots, l-1$.*

This Decreasing Expected Revenue to Uncertainty (DERU) Ratio property indicates that there are n candidate solutions from which we can find an optimal solution. Therefore, the dominant computational effort involves sorting markets according to this ratio, which can be done in $\mathcal{O}(n \log n)$ time. Once the markets are sorted, starting with market 1, we can determine the firm's expected profit by adding one market at a time to our candidate solution. Let $z(i)$ represent the expected profit when markets $1, \ldots, i$ are selected. When $z(i+1) > z(i)$, update the optimal solution to select markets 1 through i. Continue this procedure until $i = n$. It is possible for total profit to decrease after adding some market i, but then to increase and achieve a maximum total profit value based on selecting all markets up to and including $i+1$. Thus, we must enumerate all candidate solutions provided by the DERU Ratio property to find the optimal solution.

This ratio ordering is intuitively appealing, as a higher net revenue makes a market more attractive, while increases in the market's uncertainty leads to a less attractive market. It is important to point out that a market will not necessarily be attractive even though, by assumption, each market's net revenue is positive. Also, when two markets have the same ratio, they are equally attractive. If we have markets j and k such that $\bar{R}_j/\sigma_j^2 = \bar{R}_k/\sigma_k^2 = \gamma$, these markets can be treated as a single "aggregate market" with demand $D_{j+k} = D_j + D_k$, since the resulting ratio of this new market $j+k$ equals $(\bar{R}_j + \bar{R}_k)/(\sigma_j^2 + \sigma_k^2) = \gamma$. We can therefore use the term "decreasing" in place of "nonincreasing" in our description of this property without ambiguity.

By examining Equation (4.3), we see that the firm's expected profit is limited by the demand accuracy estimate within each market i. By reducing this variance estimate (through improving forecasts, reducing supplier lead times, or other measures) in each market, we could increase the expected profit and possibly include a greater number of markets in the optimal solution. Note, however, that the fixed cost S_i might increase to include some market research effort to improve on existing forecasting techniques. Or, the firm's suppliers may want to share in the increased profits to offset their costs of offering shorter lead times. In such cases, the firm would need to weigh the cost of improved forecasts against the benefit of lower demand uncertainty.

4.2.2 Managerial Insights for the SNP

Section 4.2.1 provides a nice solution approach for the SNP, given any set of markets and each market's per-unit revenue (selling price − production cost), expected demand, and standard deviation of demand. In this section, we provide insights and observations to assist a supply manager in determining the influence that each of these factors has on the acceptance or rejection of a particular market. We can also show how a market's profitability will change based on the set of markets known to exist in the optimal solution. Furthermore, we can use this information to examine the sensitivity of a particular market to changes in selling price, expected demand, or other market parameters. We assume that all markets not pre-selected for entry have been sorted according to the DERU Ratio property. Thus, any candidate solution containing market $k+1$ must also contain markets $1, \ldots, k$.

First, we assume that no markets have been selected. In order for a firm to profit from entering any one market k, we must have

$$G(Q^*_{y^k}, y^k) = \bar{r}_k - K(c, v, e)\sqrt{\sigma_k^2} \geq 0$$

where y^k represents the solution vector in which $y_k = 1$ and $y_i = 0$ for all $i \neq k$. Let $\bar{r}_k^0$ be the minimum net revenue required to achieve a profit in this single market k. Then,

$$\bar{r}_k^0 = K(c, v, e)\sigma_k. \tag{4.5}$$

The single market k will be attractive so long as the unit selling price can be set such that $\bar{r}_k \geq \bar{r}_k^0$. Since $\bar{r}_k = (r_k - c)\mu_k - S_k$, we can state this selling price as

$$r_k^0 = \frac{K(c,v,e)\sigma_k}{\mu_k} + \frac{S_k}{\mu_k} + c = K(c,v,e)\mathrm{CV}(k) + \frac{S_k}{\mu_k} + c,$$

where $\mathrm{CV}(k)$ represents the coefficient of variation for market k. The above equation directly shows the effect that variability has on the required selling price.

Now suppose we are given a set of markets $1, \ldots, k$ known to be in an optimal solution. We would include market $k+1$ if its incremental net revenue exceeds its incremental uncertainty cost. Since there is no limitation on the amount of product supplied from the overseas supplier, we would increase the optimal quantity Q^* to

$$Q^* = \sum\nolimits_{i=1}^{k+1} \mu_i + z(\rho)\sqrt{\sum\nolimits_{i=1}^{k+1} \sigma_i^2}.$$

This implies that the change in total cost will be

$$K(c,v,e)\left(\sqrt{\sum\nolimits_{i=1}^{k+1} \sigma_i^2} - \sqrt{\sum\nolimits_{i=1}^{k} \sigma_i^2}\right).$$

Let $\beta_k = \sum_{i=1}^{k} \sigma_i^2$ represent the total variance of all selected markets $1, \ldots, k$. Then the expected incremental uncertainty cost $(EIUC)$ of including market $k+1$ is

$$\begin{aligned} EUIC_{k+1} &= K(c,v,e)\left[\sqrt{\beta_k + \sigma_{k+1}^2} - \sqrt{\beta_k}\right] \\ &= K(c,v,e)\sigma_{k+1}\left[\sqrt{\tfrac{\beta_k + \sigma_{k+1}^2}{\sigma_{k+1}^2}} - \sqrt{\tfrac{\beta_k}{\sigma_{k+1}^2}}\right] \\ &= \bar{r}_{k+1}^0\left(\sqrt{1 + \tfrac{\beta_k}{\sigma_{k+1}^2}} - \sqrt{\tfrac{\beta_k}{\sigma_{k+1}^2}}\right) \end{aligned} \tag{4.6}$$

Note that as k increases, and likewise, β_k increases, the incremental uncertainty cost is dominated by the uncertainty of market k, or σ_k. We can now provide a necessary condition for including or not including market $k+1$ in an optimal solution.

Property 3 *If an optimal solution for the SNP contains only markets* $1, 2, \ldots, k$, *where* $k < n$, *then* $\bar{r}_{k+1} < \bar{r}_{k+1}^0\left(\sqrt{1 + \frac{\beta_k}{\sigma_{k+1}^2}} - \sqrt{\frac{\beta_k}{\sigma_{k+1}^2}}\right)$. *Equivalently, if an* *optimal SNP solution exists that selects markets* $1, \ldots, k$*, and* $\bar{r}_{k+1}$ *satisfies* $\bar{r}_{k+1} \geq \bar{r}^0_{k+1}\left(\sqrt{1+\frac{\beta_k}{\sigma^2_{k+1}}} - \sqrt{\frac{\beta_k}{\sigma^2_{k+1}}}\right)$*, then an optimal solution exists that selects markets* $1, \ldots, k+1$.

Notice that we have generalized the single-market condition that determines whether selecting market k by itself is worthwhile. Now we are testing if we should accept or not accept an additional market into an existing optimal solution. We can calculate the incremental cost of adding market $k+1$ to be $\bar{r}^0_{k+1}\left(\sqrt{1+\frac{\beta_k}{\sigma^2_{k+1}}} - \sqrt{\frac{\beta_k}{\sigma^2_{k+1}}}\right)$. If the incremental net revenue from market $k+1$, or $\bar{r}_{k+1}$, is greater than this incremental cost, then it would be profitable to enter market $k+1$ as well. If there are additional markets beyond $k+1$, say $k+2, \ldots, n$, and the above condition is not satisfied, then we cannot say whether market $k+1$ will ultimately be included in the optimal solution. However, if $k+1$ is the only new or additional market, then we can use Property 2 as a sufficient condition for selecting or not selecting the market. Now let's assume that $k+1$ is the only additional market, and it is not profitable to include. Then, if additional markets also become available to penetrate, we should consider all markets not yet selected (which includes $k+1$) to determine an updated optimal selection. This approach leads directly to the development of sufficient conditions for selecting or not selecting a group of additional markets, which will be discussed shortly. It is also worth noting the following special case. Consider a scenario in which the market entry cost S_{k+1} is either negligible or nonincreasing based on the market index. This implies that the required per-unit net revenue, $r_{k+1} - c$, to include market $k+1$ will naturally decrease as additional markets are included in the solution and the value of β_k increases.

We may also be able to set the price or influence the amount of demand in market $k+1$ such that there exists some marginal profit and the condition for selecting market $k+1$ is satisfied. This allows a supply manager greater flexibility when considering expanding on a current set of markets being served. If a firm has contracted with k markets and has an option to add the $k+1^{\text{st}}$ market, this marginal profit check can be used as a benchmark for expanding its operation. We also observe that as a firm serves more markets, the variability that exists in the additional market becomes less important; i.e., the variability from the markets already being served will provide an existing buffer in safety stock that can accommodate the variability brought in by the additional market.

As previously stated, Property 3 provides a necessary condition for selecting or not selecting a market. The next property, based on the DERU Ratio property, provides a sufficient condition for not selecting any additional markets beyond some current best selection of $1, \ldots, k$ markets.

Property 4 *If* $\sum_{i=k+1}^{k+j} \bar{r}_i < K(c,v,e)\left(\sqrt{\beta_{k+j}} - \sqrt{\beta_k}\right)$*, for* $j = 1, \ldots, n-k$*, then an optimal solution exists that does not include markets* $k+1, \ldots, n$*.*

We iteratively check every valid subset of markets beyond market k and up to market n to search for any incremental profit. Starting with market $k+1$ and adding the next market (in the decreasing order defined by the DERU Ratio Property) to the subset, we compare the total incremental net revenue to the total incremental uncertainty cost. If the condition is satisfied by this search, then we know that markets $k+1, \ldots, n$ will not be selected in the optimal solution.

We can also strengthen Property 3 and present a sufficient condition for selecting additional markets given a portion of the optimal solution.

Property 5 *If an optimal solution exists to SNP that selects markets* $1, \ldots, k$*, and* $\sum_{i=k+l+1}^{k+j} \bar{r}_i \geq K(c,v,e)\left(\sqrt{\beta_{k+j}} - \sqrt{\beta_{k+l}}\right)$*, for some* $j = 1, \ldots, n-k$*, and for all* $l = 0, \ldots, j-1$*, then an optimal solution exists that selects markets* $1, \ldots, k+j$*.*

Similar to Property 4, this condition requires that we evaluate all valid market subsets between $k+1$ and j based on the ordering defined in the DERU Ratio property. Given a k-market solution, we must ensure that there is incremental profit achieved by adding markets $k+1,\dots,k+j$ for some $j=k+1,\dots,n$ that is greater than the incremental profit achieved by including markets $k+1,\dots,k+l$ for all $l=1,\dots,j-1$.

The firm may also like to set a single price across all new markets entered. This is a straightforward adjustment to the condition stated in Property 5. Let $r_{k+1,k+j}$ represent the single price for markets $k+1,\dots,k+j$; i.e., $r_{k+1}=\dots=r_{k+j}=r_{k+1,k+j}$. Also, by assigning $\beta_{k+1,k+j}=\sqrt{\sum_{i=1}^{j}\sigma_{k+i}^2}$, $(S/\mu)_{k+1,k+j}=\frac{\sum_{i=1}^{j}S_{k+i}}{\sum_{i=1}^{j}\mu_{k+i}}$, and $\mathrm{CV}(k+1,k+j)=\frac{\sqrt{\beta_{k+1,k+j}}}{\sum_{i=1}^{j}\mu_{k+i}}$, we can express the minimum price across all new selected markets as

$$r_{k+1,k+j}=K(c,v,e)\mathrm{CV}(k{+}1,k{+}j)\left(\sqrt{1+\frac{\beta_k}{\beta_{k+1,k+j}}}-\sqrt{\frac{\beta_k}{\beta_{k+1,k+j}}}\right)+(S/\mu)_{k+1,k+j}+c.$$

We note again that reducing the coefficient of variation will reduce the minimum selling price required. One can see that by increasing the expected demand in any of the new markets (without increasing the respective market's demand variance), we can reduce the coefficient of variation. On the other hand, the price would have to be set higher as the variability of the new markets increases. In fact, as the demand variability within markets $k+1,\dots,k+j$ increases such that $\beta_{k+1,k+j} >> \beta_k$, we lose the benefit of having already entered into markets $1,\dots,k$. As $\beta_{k+1,k+j}$ grows large, we can approximate the minimum required selling price to offer across all additional markets $k+1,\dots,k+j$ as $K(c,v,e)\mathrm{CV}(k+1,k+j)+(S/\mu)_{k+1,k+j}+c$, which actually represents the cost of *only* entering markets $k+1,\dots,k+j$.

Our last property introduced in this section addresses how a firm can accommodate emerging markets. Initially, a firm may be faced with an n-market selection problem. After sorting based on the DERU Ratio property, we can determine which of these n markets to accept and which not to accept. As time passes, and additional markets emerge, the firm must re-evaluate the overall selection decision. Not only should the new market or markets be considered, but all markets originally rejected may also be reconsidered for selection. Assuming we have an emerging market m, then we can use Property 5 to decide if the inclusion of market m makes some additional set of markets attractive. We will include additional markets only if the firm achieves an incremental profit for doing so. Regardless, it should be clear that the new optimal solution will contain at least the original set of markets. The following property addresses this point.

Property 6 *We are given an optimal solution that selects markets* $1, \ldots, k$, *and a single new market* m *emerges. For the new* $(n+1)$*-market selection problem, if market* m *is not chosen, then the optimal solution is the same for both the* $(n+1)$*-market and* n*-market problems.*

First, assume that the optimal solution includes markets $1, \ldots, k$. Since the original n markets are sorted by the DERU property, we should place market m in its correct DERU ratio order position. Let m represent the indexed position within the ordering, which now contains $n+1$ markets. If market m has a higher DERU ratio than market k, then market m will be immediately added to the solution. Moreover, we can use Property 5 to determine if some previously unprofitable markets should now be included in the optimal solution. If market m has a lower DERU ratio than market k, then we must evaluate the inclusion of markets $k+1, \ldots, m$, and possibly more markets, into the solution based on Property 5 (i.e., set $j = m, \ldots, n+1-k$ in Property 5, and test the condition). If the condition is met for some j, then markets $k+1, \ldots, j$ will be added to the optimal selection. Otherwise, the original solution remains unchanged.

It is clear that a firm would like the flexibility of either setting the price or influencing the product demand to ensure market $k+1$ is profitable. Property 3 allows us to check profitability based on this market-specific information.

Let's assume that we hold the selling price constant. We can then derive the minimum required expected demand in order for the firm to break even by setting $\bar{r}_{k+1} = K(c,v,e)\left(\sqrt{\beta_k + \sigma_{k+1}^2} - \sqrt{\beta_k}\right)$, given this particular selling price. Letting $\mu_{k+1}^{\min}$ represent the minimum required expected demand, and assuming σ_{k+1} can be estimated and remains constant for different levels of expected demand, we have

$$\mu_{k+1}^{\min} = \frac{K(c,v,e)}{r_{k+1} - c}\left(\sqrt{\beta_k + \sigma_{k+1}^2} - \sqrt{\beta_k}\right) + \frac{S_{k+1}}{r_{k+1} - c}. \tag{4.7}$$

This could prove to be quite valuable to the firm. For example, consider that the firm would like to enter market j, which has an expected demand of μ_j, based on a fixed or pre-determined level of marketing. But, in order to add market j and be profitable, the firm must have a minimum expected demand of $\mu_j^{\min} > \mu_j$. Through additional sales and advertising, we could increase the expected demand up to the desired level, $\mu_j^{\min}$. Of course, this comes at a cost, and the additional marketing expense term would also need to be considered in equation (4.7). We address this in Section 4.3 by allowing sales and advertising effort in each market to be a decision variable; i.e., the marketing effort will no longer be fixed for each market.

4.3 SNP and the Role of Advertising

In the previous section, we considered a problem in which the demand within any market i follows a distribution with a known mean μ_i and standard deviation σ_i. Both μ_i and σ_i implicitly assumed that the sales and advertising efforts were fixed for all markets. In this section, we generalize the model to allow each market's demand distribution to be a function of marketing effort expended, which implies that μ_i and σ_i are not necessarily fixed values. Moreover, we examine contexts in which demand is highly dependent on advertising, and a market is not profitable without some level of advertising. This implies the expected demand in market i with no advertising effort (which we denote by $\underline{\mu}_i$) provides insufficient net revenue to cover the fixed market cost; i.e., $S_i > (r_i - c)\underline{\mu}_i$.

As mentioned earlier, the expected market demand resulting from a given level of advertising is usually defined through an advertising response function. Vakratsas et al. [78], Lilien et al. [48], Mahajan and Muller [52], and Johansson [40] discuss many forms that the advertising response function can take. Common functional forms used for the advertising response function are concave, linear, or S-shaped. For most industries or products, one of these response functions will approximate the behavior of demand increases with respect to the advertising level. In recent years, researchers have focused more on S-curved demand functions, which are believed to be more broadly applicable in industry. In contrast, Simon [73] shows that certain proprietary brands actually behave in an asymmetric fashion to advertising; i.e., demand peaks immediately after the advertising increase, but the long-term demand level is much lower than the initial peak. Since we focus on a single time period, this time-based effect does not apply in our case.

4.3.1 Selective Newsvendor with Marketing Effort

In order to formulate a more general model that includes sales and advertising, we introduce some additional notation. Let a_i denote the number of units (e.g., hours, days, employees) of marketing effort expended in market i, and let $t_i > 0$ be the per-unit cost of this effort. With a slight abuse of notation, let $\mu_i(a_i)$ and $\sigma_i^2(a_i)$ denote the mean and variance of expected demand in market i as a function of marketing effort a_i. We assume that the function $\mu_i(a_i)$ is nonnegative, nondecreasing, continuous, and bounded. We also assume that some marketing level b_i exists for market i, such that $a_i > b_i$ provides no additional expected demand. In particular, let $\mu_i(a) = \bar{\mu}_i$ for all $a \geq b_i$ and $\mu_i(a) < \bar{\mu}_i$ for all $a < b_i$. This means that the maximum expected demand in market i is $\bar{\mu}_i$, attained for all $a_i \geq b_i$. We also assume that the function $\sigma_i^2(a_i)$ is nonnegative, nondecreasing, continuous, and bounded, and it has similar structural properties to the mean. Denote $\sigma_i^2(0) = \underline{\sigma}_i^2 \geq 0$ as the demand variance in market i without any marketing effort. There is also a marketing effort w_i beyond which demand variance in market i is effectively constant, at a value of $\bar{\sigma}_i^2$.

We next formulate a model for profit maximization in the presence of market choice and marketing effort flexibility, or the so-called selective newsvendor problem with marketing effort (SNP-M).

[SNP-M]

$$\begin{array}{lcr} \text{maximize} & \sum_{i=1}^{n}\left((r_i-c)\mu_i(a_i)-t_ia_i-S_i\right)y_i-K\sqrt{\sum_{i=1}^{n}\sigma_i^2(a_i)y_i} & \\ \text{subject to:} & a_i \geq 0 & i=1,\ldots,n \\ & y_i \in \{0,1\} & i=1,\ldots,n. \end{array}$$

Note that [SNP-M] is a nonlinear, integer optimization problem, which initially appears to be quite difficult to solve. We first examine several forms of the advertising response function where demand variance is independent of the marketing effort. We then present a selective newsvendor model in which expected demand and demand variance both depend on the marketing effort.

4.3.2 Case I: Demand Variance Independent of Marketing Effort

In this section, we analyze the case where demand variance is independent of marketing effort (i.e., $\sigma_i^2(a_i) = \sigma_i^2$). Recall that in the basic (SNP), the total expected net revenue from serving market i was defined as $\bar{r}_i = (r_i - c)\mu_i - S_i$. Similarly, we now define market i's expected net revenue as a function of the marketing effort a_i spent in market i. We represent this as $\bar{r}_i(a_i) = (r_i - c)\mu_i(a_i) - t_ia_i - S_i$. Now note that this optimization problem is equivalent to

$$\begin{array}{lcr} \text{maximize} & \sum_{i=1}^{n}\left(\max_{a_i\geq 0}\bar{r}_i(a_i)\right)y_i-K(c,v,e)\sqrt{\sum_{i=1}^{n}\sigma_i^2y_i} & \\ \text{subject to:} & y_i \in \{0,1\} & i=1,\ldots,n, \end{array}$$

which means that the optimum level of marketing effort to exert in each market is *independent* of the market selection decision (where, of course, marketing efforts are only exerted in selected markets). That is, we may, for each market $i = 1, \ldots, n$, find the optimal marketing effort $\hat{a}_i$ in market i that will be exerted *if* market i is selected by solving the optimization problem

$$\text{maximize}_{a_i \geq 0}\ (r_i - c)\mu_i(a_i) - t_i a_i. \tag{4.8}$$

An optimal level of marketing effort $\hat{a}_i$ can be found among all values $0 \leq a_i \leq b_i$ for which the first order conditions are satisfied; i.e., $t_i/(r_i - c) \in \partial\mu_i(a_i)$, where $\partial\mu_i(a_i)$ denotes the set of subgradients at a_i (see Bazaraa, Sherali, and Shetty [11] for a discussion on necessary conditions for optimality).

Assuming we can find an $\hat{a}_i$ that solves (4.8) for all i, we then fix the marketing level in each market at this optimum value, which reduces the selected newsvendor problem to (SNP-$\hat{\text{D}}$):

[SNP-$\hat{\mathbf{D}}$]

$$\begin{array}{lll} \text{maximize} & \sum_{i=1}^{n} \bar{r}_i(\hat{a}_i)y_i - K(c,v,e)\sqrt{\sum_{i=1}^{n}\sigma_i^2 y_i} & \\ \text{subject to:} & y_i \in \{0,1\} & i = 1,\ldots,n. \end{array}$$

When formulated in this way, we immediately find the optimal solution to this problem by using the DERU property with the ranking ratio defined by $\bar{r}_i(\hat{a}_i)/\sigma_i^2$, where $\bar{r}_i(\hat{a}_i)$ replaces $\bar{r}_i$ from the original ratio presented in Section 4.2.1.

4.3.2.1 Concave Demand

If, in addition to the assumptions mentioned above, the demand function $\mu_i(a_i)$ is a concave function of a_i, then $\hat{a}_i$ can easily be found. In particular, we may efficiently find the optimal value of a_i using binary search on the interval $a_i \in [0, b_i]$. See Figure 4–1 for two examples of concave expected demand. Notice that if we assume $\mu(a_i)$ is differentiable everywhere, as in Figure 4–1(a), then we simply find a value $\hat{a}_i$ for which $\mu_i'(\hat{a}_i) = \frac{t_i}{(r_i - c)}$ (if it exists), otherwise $\hat{a}_i = 0$. If, in addition, $\mu(a_i)$ is strictly concave, then the optimal level of marketing effort $\hat{a}_i$ is unique.

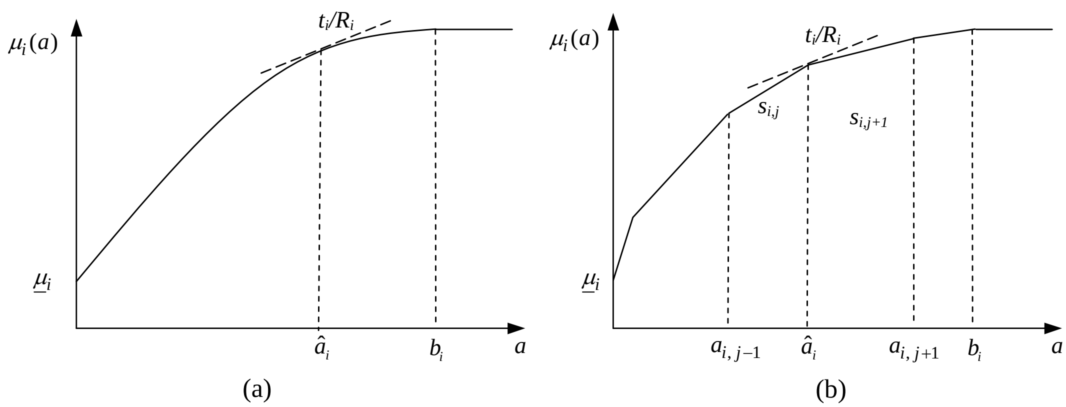


Figure 4–1: Optimal marketing effort for concave expected demand functions.

We now present the special case in which $\mu(a_i)$ is not differentiable everywhere. We first consider that expected demand in market i increases as a piecewise-linear function with decreasing slopes $s_{i1} > s_{i1} > \cdots > s_{i,J_i} > s_{iJ_i+1} = 0$ (where there are $J_i + 1$ consecutive segments) and corresponding breakpoints $a_i = a_{i0} < a_{i1} < \cdots < a_{iJ_i} = b_i$:

$$\mu_i(a) = \begin{cases} \underline{\mu}_i + \sum_{j=1}^{k-1} s_{ij}(a_{ij} - a_{i,j-1}) + s_{ik}(a - a_{i,k-1}) & \text{for } a_{i,k-1} \leq a < a_{ik}, \\ & \qquad k = 1, \ldots, J_i, \\ \underline{\mu}_i + \sum_{j=1}^{J_i} s_{ij}(a_{ij} - a_{i,j-1}) & \text{for } a \geq a_{iJ_i} = b_i. \end{cases}$$

We can then directly determine $\hat{a}_i$ for every market i as follows. Since $\hat{a}_i$ is such that $t_i/(r_i - c) \in \partial\mu_i(a_i)$ at $\hat{a}_i$, this implies that the optimal marketing effort will occur at one of the breakpoint values of the profit function. We illustrate this result in Figure 4–1(b).

We next consider the case in which expected demand in market i increases linearly with slope $s_i > 0$ for marketing effort in the interval $[0, b_i]$:

$$\mu_i(a) = \begin{cases} \underline{\mu}_i + s_i a & \text{for } 0 \leq a \leq b_i, \\ \underline{\mu}_i + s_i b_i & \text{for } a > b_i. \end{cases}$$

In fact, this is just a special case of piecewise-linear concave demand, in which there is only one slope representing the rate of increase in demand. Recalling the optimization problem stated in (4.8), $\hat{a}_i$ will be of the following form:

$$\hat{a}_i = \begin{cases} 0 & \text{for } (r_i - c)s_i - t_i \leq 0, \\ b_i & \text{for } (r_i - c)s_i - t_i > 0. \end{cases}$$

A key result of this special case is that the optimal marketing effort will always reside at either the minimum level or maximum level of advertising effort allowed (i.e., $\hat{a}_i \in \{0, b_i\}$ for this case). We will exploit a similar property for other functions in subsequent sections of the paper.

4.3.2.2 S-curved Demand

Suppose that the expected demand as a function of market effort follows an S-curve. Such a curve can be represented by a continuous function that is convex and nondecreasing up to some marketing effort level and concave and nondecreasing beyond that level. That is, the function μ_i is given by

$$\mu_i(a_i) = \begin{cases} \mu_i^{(1)}(a_i) & \text{for } 0 \leq a_i \leq \alpha_i \\ \mu_i^{(2)}(a_i) & \text{for } a_i \geq \alpha_i \end{cases}$$

where $\mu_i^{(1)}(\alpha_i) = \mu_i^{(2)}(\alpha_i)$ and $\mu_i^{(2)}(a_i) = \bar{\mu}_i$ for $a_i \geq b_i$. An example of an S-curve is given in Figure 4–2.

We are interested in finding an optimal marketing effort level $\hat{a}_i$ for this demand function. To this end, we examine the two different components $\mu_i^{(1)}$ and

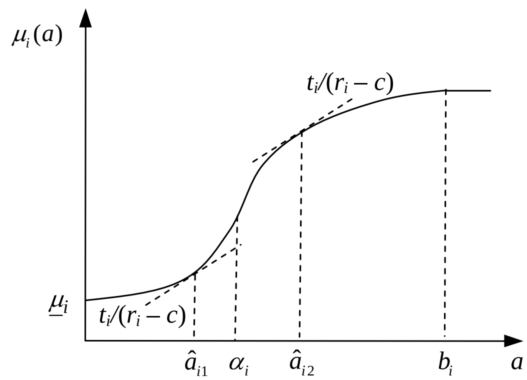


Figure 4–2: Optimal marketing effort for S-curved demand response functions.

$\mu_i^{(2)}$ independently. For $k = 1, 2$, let $\hat{a}_{ik}$ denote a value of a_i at which $t_i/(r_i - c) \in \partial\mu_i^{(k)}(a_i)$. Then, note that

(i) $\hat{a}_{i1}$ and $\hat{a}_{i2}$ correspond to a local minimum and a local maximum, respectively, of subproblem (4.8) for finding the optimal market effort, unless $\hat{a}_{i1} = \hat{a}_{i2} = \alpha_i$;

(ii) subproblem (4.8) has a local maximum at $a_i = 0$ unless $\hat{a}_{i1} = 0$ (in which case observation (i) applies);

(iii) subproblem (4.8) has a local minimum at $a_i = b_i$ unless $\hat{a}_{i2} = b_i$ (in which case observation (i) applies).

Combining observations (i)–(iii), we can now immediately conclude that that the only candidates for $\hat{a}_i$ that we need to consider are $\hat{a}_i = 0$ and $\hat{a}_i = \hat{a}_{i2}$.

4.3.3 Case II: Demand Variance Dependent on Marketing Effort

Up to this point, we have assumed that the standard deviation of demand is independent of any marketing effort exerted. In this section, we generalize the effect of marketing on demand by allowing a market's distribution (both mean and variance) of demand to be a function of the marketing effort. To address this case, we will adopt an approximation of the S-shaped curve, a broadly applicable advertising response function for expected demand.

Assume $\mu_i(a_i)$ is a convex increasing function for $0 \leq a_i \leq b_i$, and let $\mu_i(0) = \underline{\mu}_i$ and $\mu_i(a) = \bar{\mu}_i$ for all $a \geq b_i$. This describes the S-curve function for expected demand shown in Figure 4–3 below. In addition, assume the demand variance in market i is a concave and nondecreasing function, $\sigma_i^2(a_i)$, for $0 \leq a_i \leq w_i$, where $\sigma_i^2(0) = \underline{\sigma}_i^2$ and $\sigma_i^2(a) = \bar{\sigma}_i^2$ for all $a \geq w_i$.

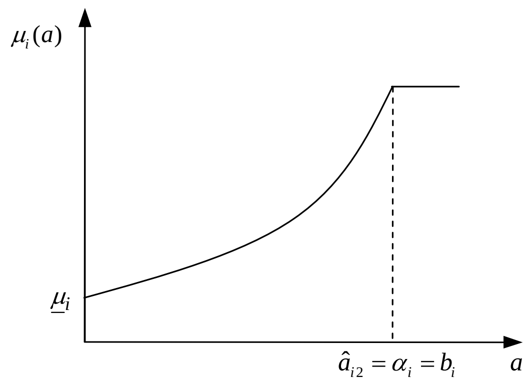


Figure 4–3: Approximation of the S-curved demand response function.

The following theorem shows that, under these assumptions, we only need to consider two distinct advertising levels in each market when solving [SNP-M], our selective newsvendor problem with marketing.

Theorem 4 *The optimal marketing effort in market i is either $\hat{a}_i = 0$ or $\hat{a}_i = b_i$ $(i = 1, \ldots, n)$.*

Proof: Fix the marketing effort levels in all markets except one, as well as the market selection variables. Without loss of generality, we may let the unrestricted market be market 1. Furthermore, let $a_i = \overline{a}_i$ for $i = 2, \ldots, n$ and $y_i = \overline{y}_i$ for $i = 1, \ldots, m$. Since the expected profit is independent of a_1 if $\overline{y}_1 = 0$, we only need to consider the case $\overline{y}_1 = 1$. Finally, define for convenience

$$V = \sum_{i=2}^{n} \sigma_i^2(\overline{a}_i)\overline{y}_i.$$

Then letting $G_1(a_1)$ denote the expected profit as a function of marketing effort in market 1 alone, and ignoring constant terms:

$$G_1(a_1) = (r_1 - c)\mu_1(a_1) - t_1 a_1 - K(c, v, e)\sqrt{\sigma_1^2(a_1) + V}.$$

The square root function is a concave increasing function, and we also know that $\sigma_i^2(\cdot)$ is concave. It then follows that the entire square root term is concave from Theorem 5.1 of Rockafellar [66], which states that an increasing concave function of a concave function is itself concave. From this result, we can now say that G_1 is convex.

This implies that the optimum marketing effort in the interval $0 \leq a_1 \leq b_1$ is at one of the two bounds: $a_1 = 0$ or $a_1 = b_1$. Furthermore, for $a_1 > b_1$ we know that

$$G_1(a_1) = (r_1 - c)\mu_1(b_1) - t_1 a_1 - K(c, v, e)\sqrt{\sigma_1^2(b_1) + V}$$

is decreasing, which means that we do not need to consider marketing effort levels in excess of b_1. This proves the desired result. $\diamondsuit$

Note that if $y_i = 1$, we will set $a_i = b_i$, since $S_i > (r_i - c)\mu_i(0) = (r_i - c)\underline{\mu}_i$. If $y_i = 0$, we can still replace a_i with b_i in the formulation without affecting the objective function value. Therefore, we set $a_i = b_i$ for $i = 1, \ldots, n$. This leads to the following formulation of the selective newsvendor problem where demand variance is dependent on marketing effort (SNP-DV):

[**SNP-DV**]

$$\begin{array}{lll} \text{maximize} & \sum_{i=1}^n \bar{r}_i(b_i)y_i - K(c, v, e)\sqrt{\sum_{i=1}^n \sigma_i^2(b_i)y_i} & \\ \text{subject to:} & y_i \in \{0, 1\} & i = 1, \ldots, n, \end{array}$$

where $\bar{r}_i(b_i) = (r_i - c)\mu_i(b_i) - t_i b_i - S_i$.

To solve for the optimal selection of markets, we can use the same DERU property introduced in Section 4.2.1, where the ratio for each market i is now defined by $\bar{r}_i(b_i)/\sigma_i^2(b_i)$.

4.3.4 Marketing Insights

In Section 4.2.2, we presented several properties to aid the firm in making market selection decisions. These properties can be updated to include the effects of advertising in a straightforward manner, based on a suitable redefinition of the value of $\bar{R}_k^0$. We also provided expressions for the minimum selling price and minimum expected demand requirements in order to achieve a profit. Here, we expand this discussion to include minimum marketing effort requirements.

Since marketing effort is a decision variable, the firm can increase the expected demand and, thus, change the expected net revenue in a market, at a cost equal to the amount of marketing effort expended. Whereas we previously defined the minimum required selling price in order to achieve a profit in market $k+1$, we will now address the minimum requirement in terms of marketing effort. There certainly are situations when the firm would prefer not to change the unit selling price, r_{k+1}. By fixing this amount, we can state a condition for selecting or not selecting market $k+1$ as

$$(r_{k+1} - c)\mu_{k+1}(\hat{a}_{k+1}) - t_{k+1}\hat{a}_{k+1} - S_{k+1} \geq K(c, v, e)\left(\sqrt{\beta_k + \sigma_{k+1}^2} - \sqrt{\beta_k}\right).$$

It would be desirable to isolate the marketing effort to determine the minimum effort required to satisfy the above condition. We address the case where expected demand is a linearly increasing function of the expended marketing effort.

Denote $a_{k+1}^{\min}$ to be the required marketing effort to select market $k+1$. Then, in the linearly increasing case, we can define this minimum marketing effort as

$$a_{k+1}^{\min} = \frac{K(c, v, e)(\sqrt{\beta_k + \sigma_{k+1}^2} - \sqrt{\beta_k}) + S_{k+1} - \underline{\mu}_{k+1}}{(r_{k+1} - c)s_{k+1} - t_{k+1}}.$$

If $a_{k+1}^{\min} > b_{k+1}$, adding solely market $k+1$ will not be profitable. Otherwise, market $k+1$ can be selected with a marketing effort of $a_{k+1}^{\min} \leq b_{k+1}$. Of course, any additional marketing up to b_{k+1} will only provide additional profit, so the firm would simply choose b_{k+1} as the appropriate marketing level. However, if marketing resources are constrained, choosing the specific amount of effort to use beyond $a_{k+1}^{\min}$ is not so clear, unless market $k+1$ is the only additional market under consideration. We will examine the limited marketing resources problem in the following section.

Now consider a scenario where both the advertising level and the selling price are flexible; i.e., a_l and r_l are both decision variables. We would then need to find an appropriate setting for both variables that equals the uncertainty cost. First, restate equation (4.5) as

$$(r_l - c)\mu_l(a_l) - t_l a_l - S_l = K(c, v, e)\sigma_l,$$

where $\mu_l(a_l)$ is defined as any advertising response function whose expected demand is dependent on the level of marketing exerted, and a_l is the marketing effort used in market l. (Note that demand variance is independent of the marketing effort in this section.)

Heuristic: Determining Valid Price/Advertising Settings: The recommended approach to solving this problem is as follows. Begin by setting r_l such that a very low profit margin is obtained; i.e., $r_l - c$ is very small. Next, search for a feasible solution to the equation where the only decision variable is a_l. Call these values r_l^1 and a_l^1. Increase the selling price by some unit amount. With the new setting of r_l^2, solve for a_l^2. Continue until sufficient data points have been collected or no advertising is required to meet the minimum net revenue (i.e., $a_l^0 = 0$). We now have a range of valid price and advertising settings from which the firm could operate.

We can assign any marketing effort when resources are unlimited, so any value for a_l can be calculated. However, if we find a solution such that $a_l > b_l$, this does not necessarily mean that the market would not be chosen. The market could be made more attractive by increasing r_l, which, in turn, would decrease the requirement for a large value of a_l. This means that this process will not result in an optimal price and advertising level setting. Instead, it offers the firm a range of values from which to consider operating with this market.

4.4 Operating with Limited Marketing Resources

In Section 4.3, we presented several solution approaches to the SNP depending on the marketing effort's influence on expected net revenue. In each case, we assumed that there were unlimited marketing resources available. Since suppliers and producers typically operate within an annual sales and advertising budget, there is most likely an upper limit on the effort that can be expended. However, the firm may not be able to spend this desired amount of marketing effort when faced with limited resources. Moreover, the DERU property no longer necessarily holds under the budget constraint. In this section, we will present the limited resources problem and discuss methods for obtaining the optimal solution.

4.4.1 Formulation of the Limited Resources Problem

We will examine the limited resources problem using the relationship between marketing effort and demand distributions previously introduced for Cases I and II in Section 4.3. We first detail the solution approach using Case II, the more general case that allows for both expected demand and demand variance to be functions of marketing effort. A similar approach can also be applied to the Case I problem, in which demand variance is fixed.

In Case II, we are given that $\mu_i(a_i)$ is a convex and nondecreasing function for $0 \leq a_i \leq b_i$, and let $\mu_i(0) = \underline{\mu}_i$ and $\mu_i(a) = \bar{\mu}_i$ for all $a \geq b_i$; i.e., the expected demand function follows the S-curve approximation function shown in Figure 4–3. In addition, denote the demand variance in market i as a general concave and nondecreasing function $\sigma_i^2(a_i)$, for $0 \leq a_i \leq w_i$, where $\sigma_i^2(0) = \underline{\sigma}_i^2$ and $\sigma_i^2(a) = \bar{\sigma}_i^2$ for all $a \geq w_i$. Furthermore, recall from Section 4.3 that market entry without any advertising is assumed to be unprofitable (i.e., $S_i > (r_i - c)\underline{\mu}_i$ for market i). Now, in this section, we impose a capacity constraint on the amount of marketing effort expended.

By isolating the fixed portion of expected demand and demand uncertainty from the components that depend on market effort, we redefine each function as

$$\begin{aligned} \mu_i(a_i) = &\quad \underline{\mu}_i + \tilde{\mu}_i(a_i), \qquad \underline{\mu}_i \geq 0,\ \tilde{\mu}_i(0) = 0, \\ \sigma_i^2(a_i) = &\quad \underline{\sigma}_i^2 + \tilde{\sigma}_i^2(a_i), \qquad \underline{\sigma}_i^2 \geq 0,\ \tilde{\sigma}_i^2(0) = 0, \end{aligned}$$

where $\tilde{\mu}_i(a_i) = \mu_i(a_i) - \underline{\mu}_i$ and $\tilde{\sigma}_i^2(a_i) = \sigma_i^2(a_i) - \underline{\sigma}_i$ for $a_i > 0$. The formulation of the selective newsvendor with limited marketing resources (LM) is closely linked to [SNP-M], except that the firm now has a maximum of B units of marketing resources available. In addition, we introduce the notation $S_i' = S_i - (r_i - c)\underline{\mu}_i > 0$, and formulation [LM] becomes

[**LM**]

$$\text{maximize} \quad \textstyle\sum_{i=1}^n \left[(r_i - c)\tilde{\mu}_i(a_i) - t_i a_i - S_i'\right] y_i - K(c, v, e)\sqrt{\textstyle\sum_{i=1}^n \left[\underline{\sigma}_i^2 + \tilde{\sigma}_i^2(a_i)\right] y_i}$$

$$\begin{aligned} \text{subject to:} \qquad & \textstyle\sum_{i=1}^n a_i \leq B, & \\ & 0 \leq a_i \leq b_i & i = 1, \ldots, n, \\ & y_i \in \{0, 1\} & i = 1, \ldots, n. \end{aligned} \tag{4.9}$$

In Section 4.3, we presented several cases in which we could fix the marketing effort variables (i.e., a_i's), reducing the problem into a form such that the optimal selection of markets can be found using the DERU property. The same approach cannot work here due to the marketing budget constraint, since we can not necessarily set each a_i to a value that achieves maximum net revenue in market i. The discussion that follows will illustrate an appropriate solution approach for market entry decisions with budgetary considerations. Though the range of potential values for a_i is implicitly defined within the demand function $\mu_i(a_i)$, we include constraint set (4.9) for a_i, whose purpose will be clear once we introduce our solution strategy for the problem.

4.4.2 Solution Approach to the Limited Resources Problem

We will implement a branch-and-bound (B&B) procedure to solve our problem. Branching will be done by fixing market selection variables y_i to appropriate values. Let I_0 denote the set of markets that are not selected (i.e., $y_i = 0$ for $i \in I_0$) and I_1 denote the set of markets that are selected (i.e., $y_i = 1$ for $i \in I_1$). The remaining markets are in I_2. A node in the B&B tree can thus be viewed as characterized by sets I_1 and I_2, and the corresponding subproblem of [LM], say [LM(I_1, I_2)], is

[LM(I_1, I_2)]

$$\text{maximize} \quad \sum_{i\in I_1} \left[(r_i - c)\tilde{\mu}_i(a_i) - t_i a_i - S_i'\right] + \sum_{i\in I_2} \left[(r_i - c)\tilde{\mu}_i(a_i) - t_i a_i - S_i'\right] y_i$$
$$-K(c,v,e)\sqrt{\sum_{i\in I_1} \left[\underline{\sigma}_i^2 + \tilde{\sigma}_i^2(a_i)\right] + \sum_{i\in I_2} \left[\underline{\sigma}_i^2 + \tilde{\sigma}_i^2(a_i)\right] y_i}$$

$$\begin{aligned} \text{subject to:} \quad & \sum_{i\in I_1\cup I_2} a_i \le B, & \\ & 0 \le a_i \le b_i & i \in I_1 \cup I_2, \qquad (4.10) \\ & y_i \in \{0,1\} & i \in I_2. \end{aligned}$$

At a given node, we will find an upper bound on the optimal value of the subproblem in that node by solving a relaxation of this problem as described below. First, observe that we may write constraint set (4.10) for $i \in I_2$ as

$$0 \le a_i \le b_i y_i \qquad i \in I_2.$$

(4.10') This now enforces that $a_i = 0$ whenever $y_i = 0$. Next, note that for markets $i \in I_1$ we can introduce an artificial continuous variable z_i that measures the fraction of the maximum marketing effort that is exerted in that market; i.e., we make the substitution $a_i = b_i z_i$ for $i \in I_1$. We can then rewrite [LM(I_1, I_2)] as

$$\begin{aligned} \text{maximize} \quad & \textstyle\sum_{i\in I_1} [(r_i - c)\tilde{\mu}_i(b_i z_i) - t_i b_i z_i - S'_i] + \sum_{i\in I_2} [(r_i - c)\tilde{\mu}_i(a_i) - t_i a_i - S'_i y_i] \\ & -K(c,v,e)\sqrt{\textstyle\sum_{i\in I_1} [\underline{\sigma}_i^2 + \tilde{\sigma}_i^2(b_i z_i)] + \sum_{i\in I_2} [\underline{\sigma}_i^2 y_i + \tilde{\sigma}_i^2(a_i)]} \end{aligned}$$

$$\begin{aligned} \text{subject to:} \quad & \textstyle\sum_{i\in I_1} b_i z_i + \sum_{i\in I_2} a_i \leq B, & & \\ & 0 \leq a_i \leq b_i y_i & i \in I_2, & \quad (4.11) \\ & 0 \leq z_i \leq 1 & i \in I_1, & \\ & y_i \in \{0, 1\} & i \in I_2. & \quad (4.12) \end{aligned}$$

Note that this problem is *equivalent* to [LM(I_1, I_2)]. We obtain an upper bound to the optimal value of this problem by relaxing constraint set (4.12) and using the following theorem.

Theorem 5 *There exists an optimal solution to the linear relaxation of [LM(I_1, I_2)] such that the upper bounding constraint set in (4.11) will be tight.*

Proof: Observe that if we reduce the value of y_i by an amount $\delta_i > 0$ such that $a_i = b_i y_i$, we do not violate any constraint. Moreover, this implies a change in objective function value equal to $\delta_i S'_i - K(c,v,e)\left[\sqrt{C - \delta_i \underline{\sigma}_i^2} - \sqrt{C}\right] > 0$, where $C = \sum_{i=1}^n [\underline{\sigma}_i^2 y_i + \tilde{\sigma}_i^2(a_i)]$, $K(c,v,e)$ is known to be nonnegative, and $S'_i > 0$. Therefore, we have increased the objective function value, which implies that $a_i = b_i y_i$ for an optimal solution. $\diamondsuit$

So we can, without the loss of optimality, assume that $a_i = b_i y_i$ in the following relaxed problem:

$$\begin{aligned} \text{maximize} \quad & \textstyle\sum_{i\in I_1} [(r_i - c)\tilde{\mu}_i(b_i z_i) - t_i b_i z_i - S'_i] + \sum_{i\in I_2} [(r_i - c)\tilde{\mu}_i(b_i y_i) - t_i b_i y_i - S'_i y_i] \\ & -K(c,v,e)\sqrt{\textstyle\sum_{i\in I_1} [\underline{\sigma}_i^2 + \tilde{\sigma}_i^2(b_i z_i)] + \sum_{i\in I_2} [\underline{\sigma}_i^2 y_i + \tilde{\sigma}_i^2(b_i y_i)]} \end{aligned}$$

$$\text{subject to:} \qquad \sum_{i\in I_1} b_i z_i + \sum_{i\in I_2} b_i y_i \le B,$$
$$0 \le z_i \le 1 \qquad i \in I_1,$$
$$0 \le y_i \le 1 \qquad i \in I_2.$$

Finally, note that by the convexity of the functions $\tilde{\mu}_i$ and the concavity of the functions $\tilde{\sigma}_i^2$ we can further relax this problem by noting that $\tilde{\mu}_i(b_i y_i) \le \tilde{\mu}_i(b_i) y_i$ and $\tilde{\sigma}_i^2(b_i y_i) \ge \tilde{\sigma}_i^2(b_i) y_i$. The relaxation to our subproblem, [R-LM(I_1, I_2)], is stated as follows:

$$\text{maximize} \quad \sum_{i\in I_1} \left[((r_i - c)\tilde{\mu}_i(b_i) - t_i b_i) z_i - S_i'\right] + \sum_{i\in I_2} \left[(r_i - c)\tilde{\mu}_i(b_i) - t_i b_i - S_i'\right] y_i$$
$$-K(c,v,e)\sqrt{\sum_{i\in I_1} \left[\underline{\sigma}_i^2 + \tilde{\sigma}_i^2(b_i) z_i\right] + \sum_{i\in I_2} \left[\underline{\sigma}_i^2 + \tilde{\sigma}_i^2(b_i)\right] y_i}$$

$$\text{subject to:} \qquad \sum_{i\in I_1} b_i z_i + \sum_{i\in I_2} b_i y_i \le B,$$
$$0 \le z_i \le 1 \qquad i \in I_1,$$
$$0 \le y_i \le 1 \qquad i \in I_2.$$

By substituting $y_i' = z_i$ for $i \in I_1$ and $y_i' = y_i$ for $i \in I_2$, we obtain a more compact formulation of the relaxation of the subproblem:

[R-LM(I_1, I_2)]

$$\text{maximize} \quad \sum_{i\in I_1} -S_i' + \sum_{i\in I_1\cup I_2} R_i y_i' - K(c,v,e)\sqrt{\sum_{i\in I_1} \underline{\sigma}_i^2 + \sum_{i\in I_1\cup I_2} C_i y_i'}$$

$$\text{subject to:} \qquad \sum_{i\in I_1\cup I_2} b_i y_i' \le B,$$
$$0 \le y_i' \le 1 \qquad i \in I_1 \cup I_2.$$

where

$$R_i = \begin{cases} (r_i - c)\tilde{\mu}_i(b_i) - t_i b_i & i \in I_1 \\ (r_i - c)\tilde{\mu}_i(b_i) - t_i b_i - S_i' & i \in I_2 \end{cases}, \text{ and}$$

$$C_i = \begin{cases} \tilde{\sigma}_i^2(b_i) & i \in I_1 \\ \sigma_i^2(b_i) & i \in I_2 \end{cases}.$$

The above form serves as our upper bounding problem at a node. We describe the solution approach to this problem, as well as the B&B implementation, in the next section.

4.4.3 Subproblem Solution and B&B Implementation

Problem [R-LM(I_1, I_2)] can be written as a special case of a more general problem discussed in Romeijn, Geunes, and Taaffe [69]. Their strategy utilizes the KKT optimality conditions and a preferential ordering of selection variables to find an optimal solution. In fact, we can use this solution approach to solve [R-LM(I_1, I_2)] in polynomial time, based on the structure of our problem. We develop the solution approach for [R-LM(I_1, I_2)] as follows.

Introducing the nonnegative dual variables λ, μ_i, ν_i, we present the KKT conditions for [R-LM(I_1, I_2)] as follows:

$$\begin{aligned} R_i - \frac{K(c,v,e)C_i}{2\sqrt{\sum_{j=1}^n C_j y_j}} - \lambda b_i - \mu_i + \nu_i &= 0 && i = 1, \ldots, n, && (4.13) \\ \lambda(\sum\nolimits_{i=1}^n b_i y_i - B) &= 0, && && (4.14) \\ \mu_i(1 - y_i) &= 0 && i = 1, \ldots, n, && (4.15) \\ -\nu_i y_i &= 0 && i = 1, \ldots, n && (4.16) \\ \sum\nolimits_{i=1}^n b_i y_i - B &\leq 0, && && \\ 0 \leq y_i &\leq 1 && i = 1, \ldots, n. && \end{aligned}$$

Note that for [R-LM(I_1, I_2)], the KKT conditions are necessary but not sufficient for optimality. Thus, our approach proceeds by enumerating all candidate KKT points. To construct candidate KKT points. We first assume that we have some candidate value of the KKT multiplier λ; we will later discuss how to determine appropriate candidate λ values. Defining $\rho_i = \mu_i - \nu_i$, we can rewrite KKT condition (4.13) as

$$\rho_i = \left[\frac{(R_i - \lambda b_i)}{C_i} - \frac{K(c, v, e)}{2\sqrt{\sum_{j=1}^{n} C_j y_j}} \right] C_i \quad i = 1, \ldots, n. \tag{4.17}$$

Observing the KKT conditions, if $\rho_i \geq 0$ we set $\mu_i = \rho_i$ and $\nu_i = 0$; otherwise set $\nu_i = \rho_i$ and $\mu_i = 0$. In terms of the primal solution, we have

$$\begin{array}{lclcl} \rho_i > 0 & \Rightarrow & \mu_i > 0 & \Rightarrow & y_i = 1 \\ \rho_i = 0 & \Rightarrow & \mu_i = \nu_i = 0 & \Rightarrow & 0 \leq y_i \leq 1 \\ \rho_i < 0 & \Rightarrow & \nu_i > 0 & \Rightarrow & y_i = 0. \end{array}$$

As indicated in Equation (4.17), we need to know the values of the y variables in order to determine each ρ_i given an appropriate value of λ. It actually turns out, however, that we do not need to know the specific values of the ρ_i variables in order to evaluate primal solutions corresponding to candidate KKT solutions. To show this, note that the second term in the equation for ρ_i is the same for all i, and the value of the ratio

$$\frac{R_i - \lambda b_i}{C_i} \tag{4.18}$$

completely determines the sign of ρ_i for each market[2] . If we rank markets in nonincreasing order of (4.18), we can be certain that if some market k has $\rho_k \geq 0$ then for all markets 1, ..., $k-1$, $\rho_i \geq 0$. Similarly, if some market l has $\rho_l < 0$, then $\rho_i < 0$ for all $i > l$. Then, for any KKT point, we must have some k_1 such that $\rho_i > 0$ for $i \leq k_1$ and some k_2 such that $\rho_i = 0$ for $k_1 < i \leq k_2$, and $\rho_i < 0$ for $i > k_2$, where $0 \leq k_1 \leq k_2 \leq n$. We therefore need to evaluate a limited number of possible k_1 and k_2 values for any given value of λ, where we set each y_i according to (4.18).

---

[2] This is the ratio obtained by applying Lagrangian relaxation to [R-LM(I_1, I_2)] when we relax the budget constraint and use Lagrangian multiplier λ.

To further reduce the values to consider for k_1 and k_2 (and the corresponding solutions in the y variables), we note that, based on a result shown in Huang et al. [38], an optimal solution exists for [R-LM(I_1, I_2)] such that at most one y_i variable is fractional. Moreover, a fractional value of y_i can only occur when the budget constraint is tight. As a result we have that, after an appropriate secondary ranking scheme (which we later discuss), a single index k exists such that $y_i = 1$ for $i = 1, \ldots, k-1$, $y_k \in [0, 1]$, and $y_i = 0$ for $i = k+1, \ldots, n$, where $k_1 + 1 \leq k \leq k_2$. When the budget constraint is tight, y_k is of the form $0 \leq y_k \leq 1$. When the budget constraint is not tight, then using the KKT conditions, we must have $\lambda = 0$, which provides one choice of ordering according to the ratio (4.18). We next consider two types of KKT solutions, those with $\lambda = 0$ and those with $\lambda > 0$.

Type I: $\lambda = 0$

If $\lambda = 0$, using ratio (4.18) we rank markets in nondecreasing order of R_i/C_i. This ratio ordering corresponds to our DERU ratio property discussed previously in the absence of a budget constraint, and this ratio also determines the signs of the ρ_i variables. Now we must ensure not only that a candidate solution obeys the DERU property, but also that the budget constraint is satisfied. Assume all markets are sorted in DERU order. Given a DERU solution that is capacity feasible containing up to market k (i.e., $\sum_{i=1}^{k} b_i \leq B$), this implies

$$\begin{aligned} y_i &= 1 && i = 1, \ldots, k-1, \\ y_k &= \min\left\{ \frac{B - \sum_{i=1}^{k-1} b_i}{b_k}, 1 \right\}, \\ y_i &= 0 && i = k+1, \ldots, n. \end{aligned}$$

If the associated solution results in

$$\begin{aligned} \rho_i &> 0 \quad \text{for } i = 1, \ldots, k-1, \\ \rho_i &< 0 \quad \text{for } i = k+1, \ldots, n, \\ \rho_k &\geq 0 \quad \text{for } y_k = 1, \\ \rho_k &= 0 \quad \text{for } 0 < y_k < 1. \end{aligned} \tag{4.19}$$

then we have a KKT point. At most n such potential solutions must be evaluated. This consideration of capacity feasible solutions resulting from the DERU sorting scheme ensures that if a KKT point exists with $\lambda = 0$, we will have considered that point.

The above approach assumes unique values of the ratio R_i/C_i. If multiple markets exist with identical ratios, we need a secondary ordering scheme for this approach to work properly. Let markets $k+1, \ldots, k_2$ have the same value of R_i/C_i. Then we can write the objective function value for [R-LM(I_1, I_2)] as

$$\begin{aligned} &\sum_{i=1}^{k} R_i + \sum_{i=k+1}^{k_2} R_i y_i + K(c,v,e)\sqrt{\sum_{i=1}^{k} C_i + \sum_{i=k+1}^{k_2} C_i y_i} \\ = \; &\sum_{i=1}^{k} R_i + \sum_{i=k+1}^{k_2} \frac{R_i}{C_i} C_i y_i \\ &+K(c,v,e)\sqrt{\sum_{i=1}^{k} C_i + \sum_{i=k+1}^{k_2} C_i y_i} \\ = \; &\sum_{i=1}^{k} R_i + \frac{R_{k+1}}{C_{k+1}} \sum_{i=k+1}^{k_2} C_i y_i \\ &+K(c,v,e)\sqrt{\sum_{i=1}^{k} C_i + \sum_{i=k+1}^{k_2} C_i y_i}. \end{aligned}$$

Increasing any y_i by ε/C_i ($\varepsilon > 0$) will have the same effect on the objective function value. However, each increase of ε/C_i consumes $\varepsilon b_i/C_i$ of the advertising budget. Therefore, we should adopt a secondary ordering scheme (when ties are present in the primary ordering) that selects markets based on increasing order of b_i/C_i. If there is a tie in both the primary and secondary ordering, this implies that we have completely identical markets. These markets could be treated as one larger market such that there still is only one fractional y_i.

Type II: $\lambda > 0$

Here we observe from the KKT conditions that when $\lambda > 0$, we must have a tight budget constraint (i.e., $\sum_{i=1}^{k} b_i y_i = B$). Given some value of λ, we can again sort markets in nonincreasing order of the ratio (4.18) and use the secondary rank ordering previously described to break any ties in the ratio. Since we must have a tight budget constraint and based on condition (4.18) (and the fact that the ratio (4.18) along with the appropriate choice of the index $k$determines the signs of the ρ_i variables) we include the first k markets in the solution such that $\sum_{i=1}^{k} b_i \geq B$ and $\sum_{i=1}^{k-1} b_i < B$, with the associated solution

$$\begin{aligned} y_i &= 1 \qquad\qquad i = 1, \ldots, k-1, \\ y_k &= \tfrac{B - \sum_{i=1}^{k-1} b_i}{b_k}, \\ y_i &= 0 \qquad\qquad i = k+1, \ldots, n. \end{aligned} \tag{4.20}$$

If the solution results in satisfying the conditions in (4.19), we have a KKT point.

Unfortunately, however, we need an appropriate value of λ in order for this procedure to work. That is, we can only be guaranteed to identify a KKT point in this way if we begin with a value of λ that corresponds to a KKT solution. Note, however, that a given ordering produced by ranking the ratios (4.18) in nondecreasing order remains the same for a range of λ values. That is, as we increase the value of λ from 0, the rank ordering only changes at specific values of λ. Moreover, we can show that as we increase λ, any pair of markets switches their relative ordering of ratio values at most once. The critical value of λ at which two markets i and j switch order in the sequence of ratio values as λ is increased from zero occurs at the point where their ratios are equal; i.e., when

$$\frac{R_i - \lambda b_i}{C_i} = \frac{R_j - \lambda b_j}{C_j}.$$

This implies that the critical value of λ, which we denote by λ_{ij}, is given by

$$\lambda_{ij} = \frac{C_i R_j - C_j R_i}{C_i b_j - C_j b_i}.$$

In other words, if $R_i/C_i > R_j/C_j$, then we have that $(R_i - \lambda b_i)/C_i > (R_j - \lambda b_j)/C_j$ for all $\lambda < \lambda_{ij}$ and $(R_j - \lambda b_j)/C_j > (R_i - \lambda b_i)/C_i$ for all $\lambda > \lambda_{ij}$. Therefore the rank ordering of ratios can only change at $(n(n-1))/2$ possible discrete values of λ_{ij}. Let p index these critical breakpoints in increasing order, and let $\bar{\lambda}_p$ denote the midpoint of the interval between λ_{p-1} and λ_p. This is equivalent to $\bar{\lambda}_p = (\lambda_p - \lambda_{p-1})/2$ for $p = 1, \ldots, (n(n-1))/2$.

We construct the index ordering for each of these values of $\bar{\lambda}_p$ using a non-increasing ordering of the ratio $(R_i - \bar{\lambda}_p b_i)/C_i$, again breaking any ties using the secondary ordering described for Type I solutions. We then construct the solution in the y variables according to (4.20), where k is such that $\sum_{i=1}^{k} b_i \geq B$ and $\sum_{i=1}^{k-1} b_i < B$. Given this solution in the y variables, we can then check if some λ exists such that $\lambda_{p-1} \leq \lambda \leq \lambda_p$ and the conditions in (4.19) are satisfied. Existence of such a λ indicates that we have found a KKT point.

Note, however, that we do not need to explicitly find the appropriate value of λ, only that we evaluate the solution implied by the ordering associated with each range of λ values such that the ordering remains constant. By evaluating all solutions such that the capacity constraint is tight and the ranking associated with each interval of λ's is obeyed, we ensure that we consider all possible KKT points such that $\lambda > 0$. The computational effort required to evaluate all Type I and Type II solutions and find the optimal market selection is $\mathcal{O}(n^3)$.

As stated previously, this solution procedure finds an optimal solution to the relaxed subproblem in which at most one market has a fractional value for y_i'. This implies that if y_i' is fractional for some market $i \in I_1$, then the corresponding market effort a_i is less than b_i. In fact, since the branching strategy does not guarantee that all markets in I_1 will actually be served in the optimal solution to the original problem [LM], it may be optimal in the subproblem [R-LM(I_1, I_2)] to perform no marketing for some $i \in I_1$.

We are now prepared to introduce the B&B procedure for solving problem [LM]. Assume that we are given the markets contained in sets I_0, I_1, and I_2. At the root node of the tree, we then solve the relaxed problem [R-LM$(\emptyset, I)$], where I represents the set of all potential markets. Based on some branching strategy, either branching on the fractional market selection variable or following a predetermined branching order, we fix one market selection variable to 0 or 1 and solve a new relaxed problem, where the one market selection variable has been placed either in set I_0 or I_1. As we move further down the tree, additional markets are added to sets I_0 and I_1 (and subsequently removed from set I_2), and we solve a problem of the form [R-LM(I_1, I_2)] at every node. Since at most one y_i' will be fractional in any subproblem solution, we can quickly construct a feasible integer solution simply by rounding the fractional y_i' to zero. This heuristic can provide a quick method for tightening the lower bound. Of course, if at any node $y_i' \in \{0, 1\}$ for $i \in I_2$, we can fathom and check if this current integer solution is an improved lower bound. We formally present the B&B procedure below, and we provide computational results using this B&B scheme in the following section.

We also note that our modeling approach can handle problems where the firm has contractual obligations in certain markets or a priori knowledge of unprofitable markets. In such situations, sets I_0 and I_1 would not be empty at the root node of the tree.

Branch-and-Bound Solution to [LM]

0) Assume we are given sets I_0, I_1, and I_2. Set the lower bound (LB) to 0, and set the upper bound (UB) to infinity. Solve the LP relaxation problem [R-LM(I_1, I_2)] associated with these set assignments. Denote Z^* as the optimal solution value. Let λ^* represent the value for which we determined the optimal ranking of markets. Define LIST as the list of all markets in I_2 ranked according to the ratio $\frac{R_i - \lambda^* b_i}{C_i}$.

1) If (z, y) is integral, STOP with the optimal solution. The marketing effort associated for each $z_i = 1$ and each $y_i = 1$ is b_i. Otherwise, set $UB = Z^*$ and continue.

2) (Branching on New Variable) Denote the first market in LIST as k. Update $I_1 = I_1 \cup \{k\}$ and $I_2 = I_2 \setminus \{k\}$.

3) If $k \in I_1$ and $y'_k = 1$ (or $k \in I_0$ and $y'_k = 0$) in the last subproblem, retain the solution (z, y), solution value Z^*, ratio ranking, and λ^* from the last subproblem, and go to Step 4. Otherwise, solve a new subproblem of the form [R-LM(I_1, I_2)]. Update LIST to include all markets in I_2 ranked according to the ratio $\frac{R_i - \lambda^* b_i}{C_i}$, where the optimal solution to this subproblem was found using a market ranking based on λ^*. Record the new solution as (z, y), with a solution value of Z^*.

4) If (z, y) is feasible and not integral, and $Z^* > LB$, go to Step 2. If (z, y) is feasible and integral, and $Z^* > LB$, set $LB = Z^*$. Continue to fathoming step.

5) (Fathoming Step) For the current node with branching market variable k, check whether this parent's other child node has already been enumerated. If so, then remove k from the appropriate set (I_0 or I_1), let $I_2 = I_2 \cup \{k\}$, update k to the parent node's branching variable and repeat Step 5. Otherwise, continue.

6) (Branching on Same Variable) If $k \in I_1$, now set $y'_k = 0$, $I_1 = I_1 \setminus \{k\}$, and $I_0 = I_0 \cup \{k\}$. Otherwise, update $I_1 = I_1 \cup \{k\}$, and $I_0 = I_0 \setminus \{k\}$. Return to Step 3.

4.5 Computational Results

We now discuss computational tests for several variants of the SNP model. First, we address the importance of the basic SNP, introduced in Section 4.2, for which we are given fixed levels of marketing to apply in each market. Then, we evaluate our solution approach for solving the SNP with limited resources, where marketing effort is a decision variable.

4.5.1 SNP Value: Minimum Market Requirement

At this stage, it is important to quantify the value to a firm when using the selective newsvendor approach. Given a set of potential customers or markets, and forecast estimates for expected demand in each market, we should be able to discern whether using the base SNP model with fixed advertising will provide any profit improvement. We have previously discussed the benefit of risk or uncertainty pooling provided by the selection of additional markets. To gain further insight into the role that the number of markets plays, consider an operation that is unprofitable with a small market set but becomes profitable with the addition of new markets. This implies that there is a minimum number of markets required before achieving a profit. The question then becomes, “Which parameters have the most influence on the minimum number of required markets?”

Consider a set of n identical markets; i.e., the expected demand (μ), demand variance (σ^2), unit revenue (r), and entry cost (S) is the same for each market. Assuming all markets will be entered (since they are identical), the resulting expected profit equation is

$$G(Q^*) = n(r-c)\mu - nS - K(c, v, e)\sigma\sqrt{n}. \qquad (4.21)$$

In order to achieve profitability, we require $G(Q^*) \geq 0$. Solving for n in equation (4.21), we have

$$n^* \geq \left(\frac{K(c, v, e)\sigma}{(r - c)\mu - S} \right)^2, \tag{4.22}$$

where n^* is the minimum number of markets required to obtain a profit. We can use equation (4.22) to draw several conclusions about the influence that each parameter has on the minimum market requirements. If the firm is faced with a low profit margin $(r - c)$, a large coefficient of variation $(CV = \mu/\sigma)$, or a high market entry cost (S), they should expect a higher minimum market requirement. Recall that $K(c, v, e) = \{(c - v)z(\rho) + (e - v)L(z(\rho))\}$. To determine the effect that $K(c, v, e)$ has on the minimum market requirement, we proceed as follows. Holding all other parameters constant in (4.22), we can plot the value of n^* as c, v, or e is increased. Figure 4–4 illustrates the effect that each parameter (c, v, e, and the market entry cost S) has on the minimum market requirement.

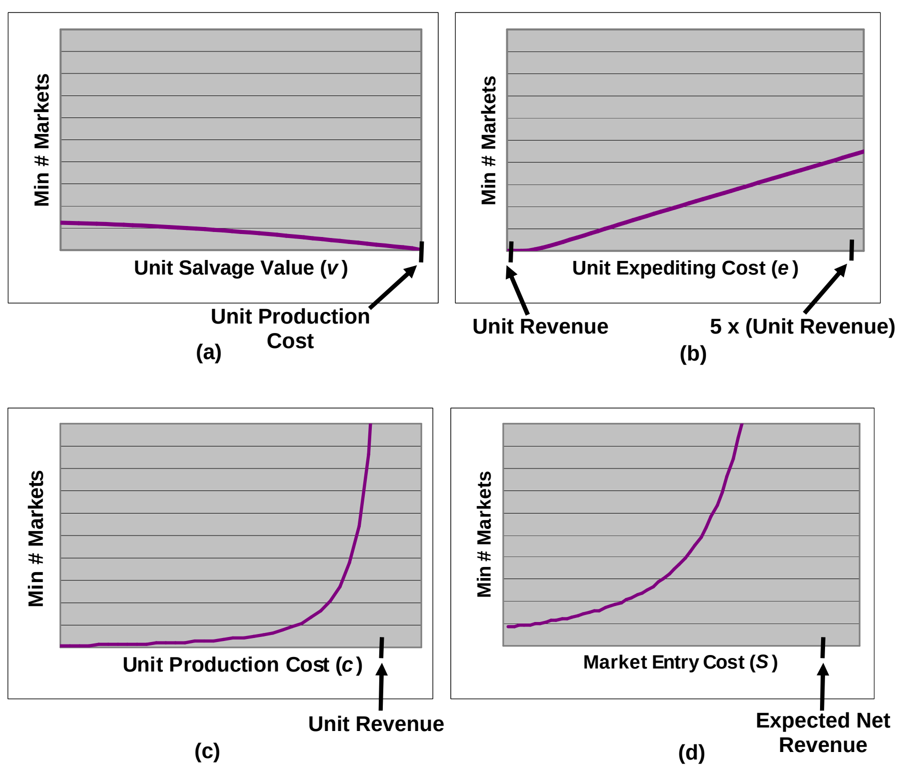


Figure 4–4: Minimum market requirement based on individual cost parameters.

Consider Figure 4–4(a), which depicts unit salvage value against the minimum number of markets required. As the unit salvage value approaches the unit production cost, the minimum market requirement decreases. This implies that if most or all of the production cost can be salvaged on unused items, the risk of entering a market is much lower, and the resulting minimum market requirement is also lower. In Figure 4–4(b), we observe that the minimum market requirement increases linearly with increases in the unit expediting cost. As would be expected, higher expediting costs will result in a larger minimum market requirement. Figure 4–4(c) compares the unit production cost against the minimum number of required markets. As the unit cost approaches the unit revenue setting, the required number of markets to obtain an expected profit increases exponentially. This result reiterates the previous conclusion that items with low profit margins will have a significant impact on the minimum market requirement. Similarly, Figure 4–4(d) further supports this argument by illustrating that the minimum market requirement also increases exponentially as market entry costs increase.

So what will happen when there are not enough candidate markets in which a firm can operate? First, we must remember that the markets in the previous example were assumed to be identical, and this is not likely to occur in an actual operation. This means that the firm can expect some differentiation between markets. If we can identify and select only those markets that will add benefit to the firm's operation, then we may achieve profitability even without having enough candidate markets. We may not need to include ALL markets in the plan, and this is where the selective newsvendor approach becomes important.

4.5.2 SNP Value: Profit Improvement

We return to our discussion of the typical market selection problem in which each market contains unique per-unit revenues, market entry cost, expected demand, and variance data. Thus far, we have shown that specific market data, the firm's cost parameters, along with the total number of possible markets in the operation, will each influence the market selection and the overall expected profit. In order to demonstrate the value of the SNP approach, we present a comparison of the optimal order quantity decisions based on the following two methods of operation:

- Method 1: Selecting a subset of markets to enter (Selective Newsvendor Approach)
- Method 2: Selecting the order quantity using all markets in which unit revenue exceeds unit cost (Maximum Market Share Approach)

After all costs and revenues have been determined, we can calculate the firm's profit for each modeling approach. Any profit improvement resulting with the SNP approach will then be recorded.

We use the following test data for the comparison. Every market has unit revenue in the range U[\$200,\$240], while the unit production cost is set at \$200. Expected demand and demand variance for each market are distributed according to U[500,1000] units and U[50000,100000], respectively. The fixed cost for market entry are drawn from U[\$2500,\$7500]. Finally, the salvage value is \$50 per unit, and we use three settings for expediting cost: \$350, \$425, and \$500 per unit, respectively. Even at the highest expediting cost of \$500 per unit, we still only obtain a critical fractile of $\rho = (e - c)/(e - v) = 0.67$. This implies that the firm is willing to accept product expediting one-third of the time. Note that it is possible for the expected net revenue in market i to be negative (i.e., $(r_i - c)\mu_i - S_i < 0$). All such markets will be removed from consideration.

Figure 4–5 presents the percent improvement in profit when implementing the SNP approach over the maximum market share approach, based on the total number of potential markets. Data are shown for each setting of the expediting cost. There is a noticeable profit improvement when the firm has 20 or fewer candidate markets to consider, which suggests that modeling this operation with a selective newsvendor approach is quite appealing. For higher levels of expediting costs, the improvement in expected profit when using the SNP approach is even more dramatic. As the number of candidate markets approaches 50, the profit gained from implementing this approach becomes minimal, which is an illustration of the effect of uncertainty or risk pooling. One can clearly see the benefit of having a large set of candidate markets from which to choose.

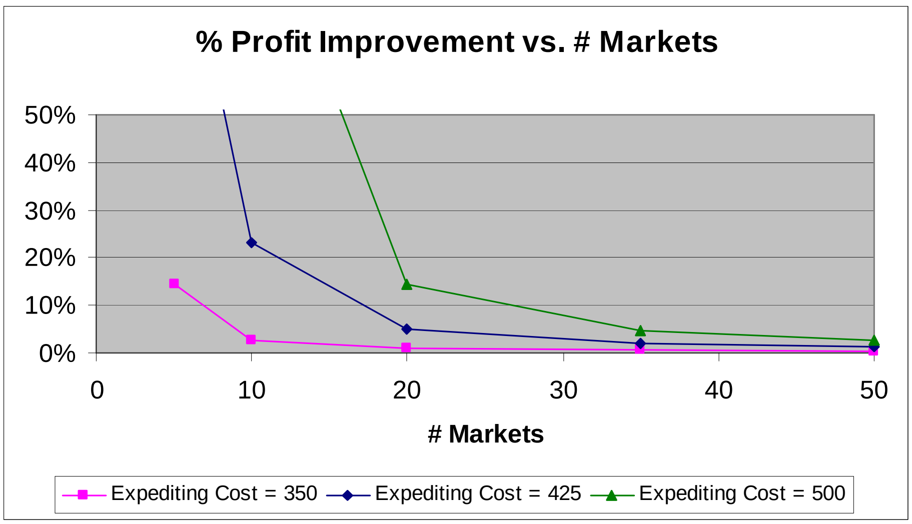


Figure 4–5: Profit improvement using SNP based on total markets available.

Figure 4–6 presents the percent improvement in profit obtained with the SNP approach, based on various levels of demand variance. For this comparison, we used a single expediting cost of $500 per unit. We tested the following ranges for demand variance within a market: U[100,2000], U[1000,10000], U[2500,25000], U[5000,50000], U[5000,75000], and U[5000,100000].

When market demand is quite predictable, there is of course very little benefit of using a selective newsvendor approach. Little or no demand uncertainty implies that there is not much difference in the candidate markets, and selecting all markets becomes the most profitable approach. Moreover, when demand variance within each market is small, the minimum market requirement is also very low.

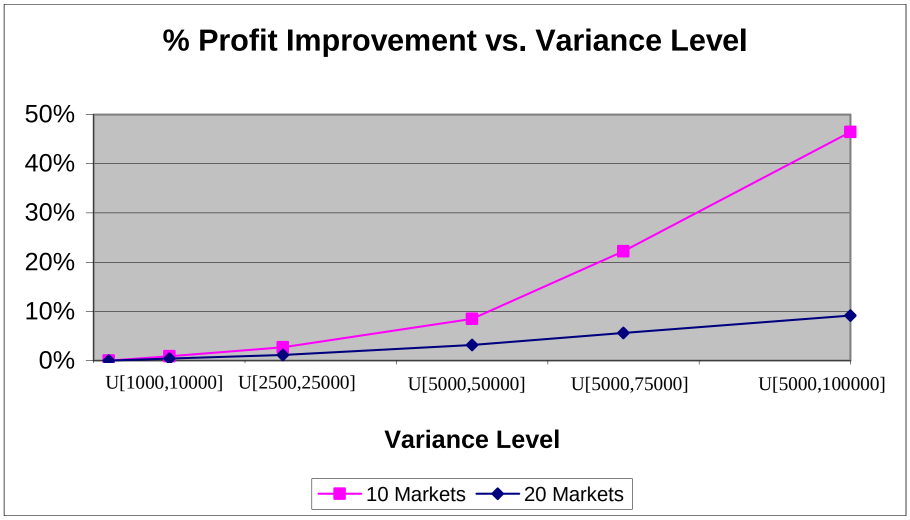


Figure 4–6: Profit improvement using SNP based on demand variance.

This is illustrated by the nearly identical results presented for the 10- and 20-market cases at low demand variance settings. As we increase the average demand variance beyond a level of 25,000 units per market, however, the improvement in profit is significant. The demand variance increase and the smaller candidate market set both contribute to the profit improvement shown on the graph.

Based on the examples of presented in this section, we can conclude that in certain contexts, firms may have the opportunity to achieve substantial profit improvements by using a selective newsvendor approach.

4.5.3 Solving the Limited Resources Problem

In this section, we examine the effectiveness of the B&B approach in solving the selective newsvendor problem with limited resources. We also perform computational tests that show the integrality gap that results from solving a relaxation of the original problem [LM]. Consider the relaxation subproblem [R-LM(I_1, I_2)] at the root node, where we assume that we are given the markets contained in sets I_0, I_1, and I_2. We will refer to this relaxation problem as [R-LM(I_1, I_2)]. Recall that we have presented solution approaches to [LM] for demand functions that follow either Case I or Case II assumptions. Case II assumes expected demand follows an S-shaped approximation that is convex and nondecreasing, and demand variance follows a concave nondecreasing function. First, we present computational results for the special case of Case II, when both expected demand and demand variance increase linearly with marketing effort. We then provide an analysis of Case I, where demand variance is assumed fixed. Again, we use a linearly increasing expected demand function in place of the more general convex nondecreasing function.

We organize the computational tests as follows. We consider three quantities for the size of the potential market pool: 10, 20, and 50. Within each market pool scenario, we set the advertising budget at each of the following levels: 25%, 50%, 75%, 100%, and 200% of the total expected demand across all available markets. As previously introduced in Section 4.5.2, unit revenue and entry cost for each market were drawn from U[$200,$240] and U[$2500,$7500], respectively, as well as assuming the production cost, expediting cost, and salvage values are $200, $500, and $50, respectively. Total expected demand depends on the following marketing effort parameters. The demand per advertising unit is distributed according to U[10,20], the unit advertising cost is distributed according to U[$30,$50], while the marketing level beyond which no additional demand can be generated is drawn from U[75,125]. For the computational tests, we generated 500 random problem instances for each market pool size and each case (I and II), for a total of 3000 problem instances.

Table 4–1 compares the solution quality of problem [R-LM(I_1, I_2)] to problem [LM] for Case II, in which expected demand and demand variance increase linearly with the level of marketing effort. Only for an advertising budget of less than 50% of total expected demand is there a significant integrality gap between the relaxed and mixed integer formulations. As the market pool increases to a size of 50 potential markets, the solution provided by [R-LM(I_1, I_2)] is within 0.31% of

Table 4–1: Results for SNP with limited resources – Case II.

Scenario/Measurement	Advertising Budget: % of Expected Demand				
	25%	50%	75%	100%	200%
10 Markets					
[R-LM(I_1, I_2)]–[LM] %Gap	6.02%	1.21%	0.31%	0.02%	0.00%
[LM] Solution Time	0.00 sec.	0.00 sec.	0.00 sec.	0.00 sec.	0.00 sec.
Avg # Nodes Used	10	11	7	2	1
20 Markets					
[R-LM(I_1, I_2)]–[LM] %Gap	1.11%	0.59%	0.13%	0.00%	0.00%
[LM] Solution Time	0.01 sec.	0.02 sec.	0.01 sec.	0.00 sec.	0.00 sec.
Avg # Nodes Used	17	26	17	1	1
50 Markets					
[R-LM(I_1, I_2)]–[LM] %Gap	0.31%	0.20%	0.06%	0.00%	0.00%
[LM] Solution Time	0.67 sec.	1.17 sec.	0.86 sec.	0.03 sec.	0.03 sec.
Avg # Nodes Used	47	76	59	1	1

the exact [LM] solution. It is also worth pointing out that for problems with 50 or fewer markets, solving [LM] to optimality typically requires less than 1.0 CPU second. We also include the number of subproblems solved in the B&B tree to find the optimal [LM] solution.

Table 4–2 compares the solution quality of problem [R-LM(I_1, I_2)] to problem [LM] for Case I, in which expected demand increases linearly with the level of marketing effort and demand variance remains constant at any advertising level. Notice that the integrality gap for instances when the advertising budget is less than 50% of expected demand is significantly higher than the same results for Case II. Recall that, for [R-LM(I_1, I_2)], we invoke a linearity assumption for all demand function's response to advertising. This implies that, for Case I and II, [R-LM(I_1, I_2)] at the root node will have the same value. (Once subproblem [R-LM(I_1, I_2)] is solved at other nodes in the tree, solutions may be different between the two cases.) However, under Case I, the entire uncertainty cost is spent as soon as the market is entered. So, the [LM] solution provided under Case I will always be less than or equal to the [LM] solution under Case II for the same problem data, and this results in a larger gap. This conclusion is also supported by the fact that

Table 4–2: Results for SNP with limited resources – Case I.

	Advertising Budget: % of Expected Demand				
Scenario/Measurement	25%	50%	75%	100%	200%
10 Markets					
[R-LM(I_1, I_2)]–[LM] %Gap	17.21%	3.03%	0.48%	0.02%	0.00%
[LM] Solution Time	0.00 sec.	0.00 sec.	0.00 sec.	0.00 sec.	0.00 sec.
Avg # Nodes Used	13	12	8	1	1
20 Markets					
[R-LM(I_1, I_2)]–[LM] %Gap	2.42%	1.15%	0.18%	0.00%	0.00%
[LM] Solution Time	0.00 sec.	0.00 sec.	0.00 sec.	0.00 sec.	0.00 sec.
Avg # Nodes Used	28	33	15	1	1
50 Markets					
[R-LM(I_1, I_2)]–[LM] %Gap	0.64%	0.30%	0.08%	0.00%	0.00%
[LM] Solution Time	1.70 sec.	1.81 sec.	0.92 sec.	0.03 sec.	0.03 sec.
Avg # Nodes Used	83	102	60	1	1

solving [LM] requires additional nodes in the B&B tree. Yet, on average, the time required to obtain the [LM] solution is still less than 2 CPU seconds.

As problem instances increase beyond 50 markets, the solution time for the mixed integer formulation may become substantial. Since we observe that [R-LM(I_1, I_2)] at the root node provides a very tight bound on [LM] for these large market pool problems, we could simply solve [R-LM(I_1, I_2)] and round to obtain an integer solution. This heuristic approach is likely to yield high quality solutions.

4.6 Other Considerations

4.6.1 Extension to the Infinite Horizon Planning Problem

While many of the applications in the newsvendor literature involve problems with a single planning period and selling cycle, there are many situations when this cycle will repeat itself in future selling seasons. Therefore, it would be desirable to examine how to select markets and set the order quantity for each selling season. First, we assume that each market has periodic independent, stationary, and identically distributed (iid) demand over an infinite horizon. Any supplier shortages incurred within a period are now back-ordered at a cost of b per unit. Given a multi-period problem, there is no longer a need to salvage excess product. Instead, we maintain the product in inventory at a cost of h per unit in each period. We also assume that the product does not change and can be sold in future periods. The firm must also have an unconstrained supply and either have negligible order costs per period or have scheduled deliveries in every period.

In the infinite horizon problem, we want to maximize the long-run expected profit per period. Assuming full back-ordering of demand, we can extend the SNP model to an infinite horizon in a similar manner shown in Nahmias [56]. To determine the long-run expected profit, we start by assuming an N-period problem. In order to compute the N-period profit, we must know the quantities sold in each period. Let D_t^i represent the realized demand from market i in period t. Then, let D_t^y represent the total quantity demanded across all markets for period t (i.e., $D_t^y = D_t^1 + D_t^2 + \cdots + D_t^I$). In the first period, the firm would order Q. In fact, since each market has iid demand per period, the firm would maintain an order of Q for every period. Since any demand not met is back-ordered, this implies that the firm will always place an order equivalent to last period's total demand. This is shown below:

$$\begin{aligned}
\text{Units sold in period 1} &= \min(Q, D_1^y) \\
\text{Units sold in period 2} &= \max(D_1^y - Q, 0) + \min(Q, D_2^y) \\
\text{Units sold in period 3} &= \max(D_2^y - Q, 0) + \min(Q, D_3^y) \\
&\;\;. \\
&\;\;. \\
&\;\;. \\
\text{Units sold in period } N &= \max(D_{N-1}^y - Q, 0) + \min(Q, D_N^y).
\end{aligned}$$

We can then state the N-period expected profit as

$$\begin{aligned} G(Q,y) &= \sum_{i\in I}((r_i-c)E(D_1^i+D_2^i+\cdots+D_{N-1}^i)-S_i)y_i+\overline{r}E[\min(Q,\sum_{i\in I}D_N^i y_i)]-cQ \\ &\quad -N\left[h\int_0^Q(Q-x)f_y(x)\mathrm{d}x+b\int_Q^\infty(x-Q)f_y(x)\mathrm{d}x\right] \\ &= \sum_{i\in I}((r_i-c)(N-1)\mu_i-S_i)y_i+\overline{r}E[\min(Q,\sum_{i\in I}D_N^i y_i)]-cQ \\ &\quad -N\left[h\int_0^Q(Q-x)f_y(x)\mathrm{d}x+b\int_Q^\infty(x-Q)f_y(x)\mathrm{d}x\right], \end{aligned}$$

where $\overline{r}$ represents the per unit average revenue of the quantity sold in period N. This would be a relatively difficult value to calculate, but it will not be necessary. Dividing by N and letting $N\to\infty$, we can represent the long-run expected profit per period as

$$\begin{aligned} \overline{G}(Q,y) &= \sum_{i\in I}(r_i-c)\mu_i y_i-h\int_0^Q(Q-x)f_y(x)\mathrm{d}x-b\int_Q^\infty(x-Q)f_y(x)\mathrm{d}x \\ &= \sum_{i\in I}(r_i-c)\mu_i y_i-hQ+h\mu_y-(b+h)\int_Q^\infty(x-Q)f_y(x)\mathrm{d}x. \end{aligned}$$

For a given vector y, this expected profit equation is convex in Q, and the first-order condition implies an optimal order quantity, Q_y^*, satisfying

$$F_y(Q_y^*)=\frac{b}{b+h},$$

where we now have $\rho=\frac{b}{b+h}$. For a given vector y, and the resulting loss function of $\Lambda_y(Q)=\int_Q^\infty(x-Q)f_y(x)\mathrm{d}x$, we rewrite the firm's long-run expected profit as:

$$\bar{G}(Q_y^*,y)=\sum\nolimits_{i=1}^n\left[(r_i-c)\mu_i-S_i\right]y_i-hQ_y^*+h\mu_y-(b+h)\Lambda_y(Q_y^*).$$

Assuming normally distributed demand, we use the same optimal order quantity equation (4.2) and identity $\Lambda_y(Q_y^*)=\sigma_y L(z)$ as before. Letting $z(\rho)=z(\frac{b}{b+h})$ and $\bar{r}_i=(r_i-c)\mu_i$, we can write the firm's long-run expected profit as:

$$\bar{G}(Q_y^*,y)=\sum\nolimits_{i=1}^n\bar{r}_i y_i-\{hz(\rho)+(b+h)L(z(\rho))\}\sqrt{\sum\nolimits_{i=1}^n\sigma_i^2 y_i}.$$

Replacing $K(c, v, e)$ from the original model with $K(b, h) = \{hz(\rho) + (b + h)L(z(\rho))\}$, we formulate the selective newsvendor problem over an infinite horizon (SNP-I):

[SNP-I]

$$\begin{array}{lll} \text{maximize} & \sum_{i=1}^{n} \bar{r}_i y_i - K(b,h)\sqrt{\sum_{i=1}^{n} \sigma_i^2 y_i} & \\ \text{subject to:} & y_i \in \{0,1\} & i = 1, \ldots, n. \end{array}$$

Notice that this formulation has the same structure as the single period version of the problem, with the following minor exceptions. First, the market entry cost is not present in the formulation. This is due to the fact that a cost incurred only in the first period will not be significant in the long run. Second, the $K(c, v, e)$ constant has been replaced with $K(b, h)$, which means that the production cost no longer influences uncertainty-related costs. As with formulation [SNP], we can apply the DERU Ratio property to [SNP-I] and determine the optimal selection of markets, where the ratio for each market i is defined by $\bar{r}_i/\sigma_i^2$.

4.6.2 Limited Marketing Effort under a Fixed Contract

Instead, let's assume we are given a pre-defined selection of markets; i.e., the firm is operating under a fixed contract that states it will serve this given set of markets. This assumption, in fact, allows us to formulate a problem for which we will offer a straightforward solution approach. Similar to Section 4.2.1, define a binary vector of market selection variables y, and let I_1 denote the set of markets such that $y_i = 1$. We can write the firm's expected profit equation as

$$G(Q_y^*, y) = \sum_{i \in I_1} (r_i \mu_i(a_i) - t_i a_i - S_i') - (c - v)Q_y^* - v\mu_y - (e - v)\Lambda_y(Q_y^*).$$

Again, assuming normally distributed demand and a marketing effort a_i that only affects the mean value of demand in market i, we can restate the optimal order quantity as

$$Q_y^* = \sum\nolimits_{i\in I_1} \mu_i(a_i) + z(\rho)\sqrt{\sum\nolimits_{i\in I_1} \sigma_i^2}.$$

where $z(\rho) = \Phi^{-1}(\rho)$ is the standard normal variate value associated with the fractile ρ. Through the same simplifications used in Section 4.2.1, the firm's expected profit becomes

$$G(Q_y^*, y) = \sum\nolimits_{i\in I_1} \bar{r}_i(a_i) - K(c,v,e)\sqrt{\sum\nolimits_{i\in I_1} \sigma_i^2}. \tag{4.23}$$

where $\bar{r}_i(a_i) = (r_i - c)\mu_i(a_i) - t_i a_i - S_i'$. Notice that the uncertainty term in the above equation is simply a fixed cost, regardless of the amount of marketing effort expended in any market. Excluding this term from the optimization problem, and assuming a maximum available marketing effort of B, the selective newsvendor with limited resources (SNP-LR) can be formulated as:

[SNP-LR]

$$\begin{aligned} \text{maximize} \quad & \textstyle\sum_{i\in I_1} \bar{r}_i(a_i) & \\ \text{subject to:} \quad & \textstyle\sum_{i\in I_1} t_i a_i \leq B & i \in I_1, \\ & 0 \leq a_i \leq b_i & i \in I_1. \end{aligned}$$

This problem can be classified as a nonlinear knapsack problem, and we will present solution approaches based on the various forms that the expected demand function can take. Using the special cases of linearly increasing expected demand and piecewise linearly increasing expected demand shown in Case I of Section 4.3.2, we now develop solution approaches under a fixed contract.

First, let's consider the case of having piecewise-linear nondecreasing concave expected demand as a function of marketing effort. In Section 4.3.2, we determined that the optimal marketing effort in market i will be equal to a value $\hat{a}_i$ such that $t_i/(r_i - c) \in \partial\mu_i(a_i)$ at $\hat{a}_i$. This is the value where $(r_i - c)s_{ij} \geq t_i \geq (r_i - c)s_{i,j+1}$ for market i. Assume that there are M_i linear demand segments in market i

where $(r_i - c)s_{ij} \geq t_i$ for some j. We would like to choose the most profitable segments from the $\sum_{i \in I_1} M_i$ profitable segments across all markets. We can place these segments in nondecreasing order of $(r_i - c)s_{ij}$ values, and simply assign the marketing effort to the corresponding profitable market segment until the marketing effort limit of B is reached or exceeded. Let k_i represent the number of segments from market i included in the optimal solution, and let a_{k_i} denote the maximum marketing level associated with segment k_i. Finally, let j represent the market for which the limit of B was reached. Then, the optimal marketing effort in market $i \in I_1 \setminus j$ will be $\hat{a}_i = a_{k_i}$, and we can determine the resulting expected demand in market i based on the piecewise-linear function defined in Section 4.3.2. Since segment k_j from market j is the last segment assigned, then the optimal marketing effort in market j is represented by $\hat{a}_j = B - \sum_{i \in I_1 \setminus j} a_{k_i}$. This assignment procedure requires $\mathcal{O}(n \sum_{i \in I_1} M_i \log \sum_{i \in I_1} M_i)$. The reason that this approach will work is that we have committed up-front to entering all of the markets included in [SNP-LR], and the uncertainty cost depicted in equation (4.23) is a constant. Therefore, the only decision is to determine how much profit to obtain from each market that has been entered.

Actually, the linearly increasing case is now just a special case of the piecewise-linear case, where each market only has one linear demand segment (i.e., $M_i = 1$ for all i). Then, the above algorithm will also solve the linearly increasing case optimally as well.

4.7 Conclusions

To fully address demand selection modeling, one must surely consider the effect that demand uncertainty has on our selection decisions. Even when the firm decides a priori to apply certain fixed marketing levels in each market, we are faced with solving an integer problem with a nonlinear profit maximization objective.

As we have shown for the basic SNP, our modeling approach utilizes the problem structure to provide a closed-form solution based on the ranking of each market's attractiveness, which we call the DERU (Decreasing Expected Revenue to Uncertainty) ratio. Once we have an optimal ranking, we offer several what-if scenarios for the firm to consider. Each of these insights can play a role in deciding whether to expand its presence in new markets, increase the marketing effort, or offer some form of price discounting.

We also consider the role that sales and advertising plays in determining the demand distribution observed within each market. When the firm decides a priori to apply certain fixed marketing levels in each market, we are faced with solving an integer problem with a nonlinear profit maximization objective. And we have shown through the DERU ratio property that this problem is surprisingly easy to solve. Beyond this fixed-advertising basic SNP, we evaluated a fairly general set of advertising response functions, each of which has unique properties for the optimal selection of markets and the corresponding advertising levels in these markets.

For the cases in which there was an unlimited marketing budget, we can determine the optimal marketing level to expend in each market, and we show how the problem simplifies to the basic SNP. Combining the effects of a limited marketing budget with a profit objective based on expected revenues with demand uncertainty, we are then faced with a nonlinear knapsack problem with a nonseparable objective function, which is a very difficult problem to solve in general. For the limited resources case, we provide a branch-and-bound procedure to obtain the optimal solution. However, we can actually solve the nonlinear knapsack relaxation subproblems in polynomial time at each node. In fact, for problems with 50 markets, we have also shown that the integrality gap provided by the original problem relaxation is less than 1% for the randomly generated problem instances we tested.

CHAPTER 5
AIRPORT CAPACITY LIMITATIONS – SELECTING FLIGHTS FOR GROUND HOLDING

5.1 Introduction

We now shift our attention from the production and ordering systems described in Chapters 2 - 4 to a demand selection problem arising in the air transportation industry. Throughout this dissertation, we develop solution approaches to decision problems that require selecting the appropriate demand sources to satisfy, based on available resources. In this chapter, we address how to select aircraft for arrival to a single airport experiencing bad weather, through the implementation of a ground holding plan. We also demonstrate the advantages of using a stochastic programming approach to address weather-related uncertainties.

Over the past 20 years, business and leisure air travel have consistently increased in popularity. With more and more passengers wanting to travel, airlines and airports have continued to expand to meet passengers' needs. And now, many airports are at or near capacity with few options for expansion (see U.S. House Subcommittee on Aviation [77]). As more airports approach their capacity, the air travel industry is witnessing higher average delays. While some delays result from an airline's operations (ground servicing, flight crews, late baggage, etc.), a majority of the severe delays are weather related. During bad weather, the Federal Aviation Administration (FAA) imposes restrictions on the number of aircraft an airport can accept in an hour. In technical terms, the airport will be instructed to operate under one of three flight rule policies: VFR (Visual Flight Rules), IFR1 (Instrument Flight Rules 1), or IFR2 (Instrument Flight Rules 2 – more restrictive than IFR1). An airport operates under VFR during good weather or normal conditions. As the weather conditions deteriorate, the FAA may restrict airport capacity by requiring an airport to operate under IFR1 or IFR2. In the most extreme cases, an airport will temporarily close until the poor weather conditions subside.

As a result of these rules, the airport and airlines must decide what to do with all of the aircraft wanting to arrive at an airport experiencing bad weather. The aircraft can be allowed to take off and approach the airport, resulting in some air delays while flight controllers sequence these arriving aircraft. Alternatively, the aircraft can be held at their originating stations, incurring what is called a ground holding delay. Finding the desired balance between ground delays and air delays under severe weather conditions that achieves the lowest cost is the focus of this paper.

The airport acceptance rate (AAR) plays an important role in determining ground holding policies at airports across the nation. Since all airports have finite capacity, there is a continuing effort to maximize capacity utilization, while avoiding unwanted ground and air delays, which impact fuel costs, crew and passenger disruptions, and other intangible costs. We use a stochastic programming approach, which takes into consideration the random nature of the events that put a ground holding plan into place. As these plans cannot predict the future, there may be unnecessary ground holds at originating or upline stations, resulting in unused capacity at the airport in question if the projected capacity reductions (weather-related or not) do not occur.

Research has been conducted on the ground holding problem for single and multiple airports, and both static and dynamic versions of the problem exist. The static version assumes that the capacity scenarios are defined once at the beginning of the ground holding period under evaluation. Our analysis focuses on the static, stochastic single airport version of the problem. For additional background on the static, single-airport problem, see Ball, et al. [8], Grignon [34], Hoffman and Ball [36], Richetta and Odoni [64], and Rifkin [65]. While our contributions to the static modeling approach offer several new insights, we recognize the potential limitation in correctly representing the real-time changes in airport operations due to weather uncertainties. In contrast, a dynamic version has been studied in which capacity reduction scenarios are updated as the day progresses. In other words, as we move from one period to the next, the capacity estimates for all remaining periods can be updated to reflect more recent weather forecasts. Thus, there could be a total of k unique sets of capacity scenarios for time period k. For research on the dynamic or multiple airport problems, please see Vranas, Bertsimas, and Odoni [80], [81], and Navazio and Romanin-Jacur [57], of which only Vranas et al. [80] considers capacity to be stochastic in nature.

In Section 5.2, we present the Rifkin [65] stochastic formulation of the ground holding problem, along with some solution properties. We adopt this model formulation and develop new findings, which are presented in Sections 5.3 and 5.4. First, Section 5.3 illustrates the benefit of studying a stochastic as opposed to a deterministic ground holding problem. A justification for the use of a stochastic model is presented through a series of computational experiments performed with various input data. Benchmark measures, such as the value of the stochastic solution and expected value of perfect information (see Birge and Louveaux [15]), are used for this justification. In Section 5.4, we consider the effect of introducing risk-averse constraints in the model. In other words, by restricting the size of the worst-case delays, how is the overall expected delay affected? This is not the same as a maximum delay model, which would place a strict upper bound on worst-case delays. Finally, we summarize our findings and discuss directions for future work.

5.2 Static Stochastic Ground Holding Problem

5.2.1 Problem Definition and Formulation

We define a potential ground holding period to consist of a finite number, T, of 15-minute increments or periods. Suppose the arrival schedule contains F flights during the time horizon under evaluation. Further assume that we can group individual flight arrivals into unique time periods, where the sequencing of flights within the time period is not important. We can then denote the number of arrivals initially scheduled to arrive in period t as D_t.

While the airport may have a nominal arrival capacity of X aircraft per period, the estimates based on the poor weather conditions will produce Q possible capacity scenarios within any interval. For each capacity scenario q, there is a probability p_q of that scenario actually occurring. For each time period and scenario, let M_{qt} be the arrival capacity for scenario q during period t. Let c_g denote the unit cost of incurring a ground delay in period t. Assume that ground delays do not increase in cost beyond the first time period for any aircraft. Similarly, define an air delay cost, c_a, as the unit cost of incurring an air delay for one period. We will use these parameters to examine model performance for different ground/air delay ratios in Section 5.3. We next define the following decision variables:

A_t	$=$	Number of aircraft allowed to depart from an upline station and arrive "into the airspace" of the capacitated airport during period t.
W_{qt}	$=$	Number of aircraft experiencing air delay during period t under scenario q.
G_j	$=$	Number of aircraft incurring ground delays in period j. This is the difference between the actual number of arrivals $(\sum_{t=1}^{j} D_t)$ through period j and the total number of expected arrivals $(\sum_{t=1}^{j} A_t)$ through period j.

As stated previously, we focus on the static, stochastic single airport version of the problem. Based on the problem first presented in Richetta and Odoni [64] and later revised in Rifkin [65], the formulation is as follows:

[SSGHP] Static Stochastic Ground Holding Problem (General Case)

$$\begin{aligned}
\text{minimize:} \quad & c_g \sum_{t=1}^{T} G_t + c_a \sum_{q=1}^{Q} \sum_{t=1}^{T} p_q W_{qt} && && (5.1)\\
\text{subject to:} \quad & && && \\
\text{Schedule Arrival Times:} \quad & \sum_{t=1}^{j} A_t \le \sum_{t=1}^{j} D_t && j = 1, \ldots, T, && \\
& \sum_{t=1}^{T+1} A_t = \sum_{t=1}^{T+1} D_t && && (5.2)\\
\text{Arrival Period Capacities:} \quad & A_t + W_{q,t-1} - W_{qt} \le M_{qt} && t = 1, \ldots, T, && \\
& && q = 1, \ldots, Q, && (5.3)\\
\text{Initial Period Air Delays:} \quad & W_{q0} = 0 && q = 1, \ldots, Q, && (5.4)\\
\text{Ground Delays:} \quad & G_j + \sum_{t=1}^{j} A_t = \sum_{t=1}^{j} D_t && j = 1, \ldots, T, && (5.5)\\
\text{Integrality:} \quad & A_t \in Z^+, W_{qt} \in Z^+, G_t \in Z^+ && t = 1, \ldots, T, && \\
& && q = 1, \ldots, Q. && (5.6)
\end{aligned}$$

The objective function minimizes total expected delay cost, accounting for both ground and air delays. Constraint set (5.2) shows that no aircraft will arrive earlier than its planned arrival period. The equality constraint that includes a summation to period $T+ 1$ requires that we account for all aircraft that could not land by period t. Therefore, any aircraft not landing by T will land in period $T + 1$. When examining an entire planning day, this is fairly realistic since even the busiest airports will reduce their operations to 10–20% of capacity late in the evening. When we only want to consider a portion of a planning day, then we need to realize that there may not be enough capacity in period $T + 1$ to land all aircraft. So some additional delay will be present. Constraint set (5.3) requires that, for a given time period, all en-route aircraft, including those on-time and those already experiencing air delay, will either land or experience an air delay until the next time period. Constraint set (5.4) assumes that there are no aircraft currently experiencing delays and waiting to land, prior to the beginning of the ground holding policy. Constraint set (5.5) assigns ground delays to those aircraft not landing in their originally desired time periods. Thus, one can see that aircraft can be assigned ground delay, air delay, or both, when airport capacity is restricted.

5.2.2 Solution Properties

It has been shown (Richetta and Odoni [64], Rifkin [65]) that the constraint matrix of [SSGHP] is totally unimodular, and it follows that the linear programming relaxation of [SSGHP] is guaranteed to yield an integer solution. Thus, constraint set (5.6) can be replaced with the nonnegativity constraint set:

$$\text{Nonnegativity:} \quad A_t, W_{qt}, G_t \geq 0 \quad t = 1, \ldots, T, q = 1, \ldots, Q. \tag{5.7}$$

This property will not hold when we introduce the risk aversion measures in Section 5.4, so we cannot remove the integrality constraints from all of our models in this analysis.

While the original problem itself is not very large, it is always desirable to reduce the problem to a linear program. The number of integer decision variables is $\mathcal{O}(T + T * Q + T) = \mathcal{O}(QT)$, where Q is the number of scenarios and T is the number of 15-minute periods. As will be shown, the experiments presented in this report are based on a 22-scenario, 24-period problem, which implies the evaluation covers a six-hour timeframe. This translates to 576 integer variables, which is very manageable. However, as more accurate data on weather patterns are available, we would most likely have historical data on many more airport capacity scenarios. The resulting model would increase in size very quickly, and the benefit of not requiring the integer restriction now becomes more important.

5.3 Motivation for Stochastic Programming Approach

5.3.1 Arrival Demand and Runway Capacity Data

When a ground holding policy is being considered, the expected arrivals to the airport will be affected. So it is important to know the arrival stream. We consider a six-hour timeframe. Based on actual data from a major airport, an estimated arrival demand in 15-minute increments was obtained. Figure 5–1 presents this arrival demand data. Each chart shows typical arrival patterns for an airport with

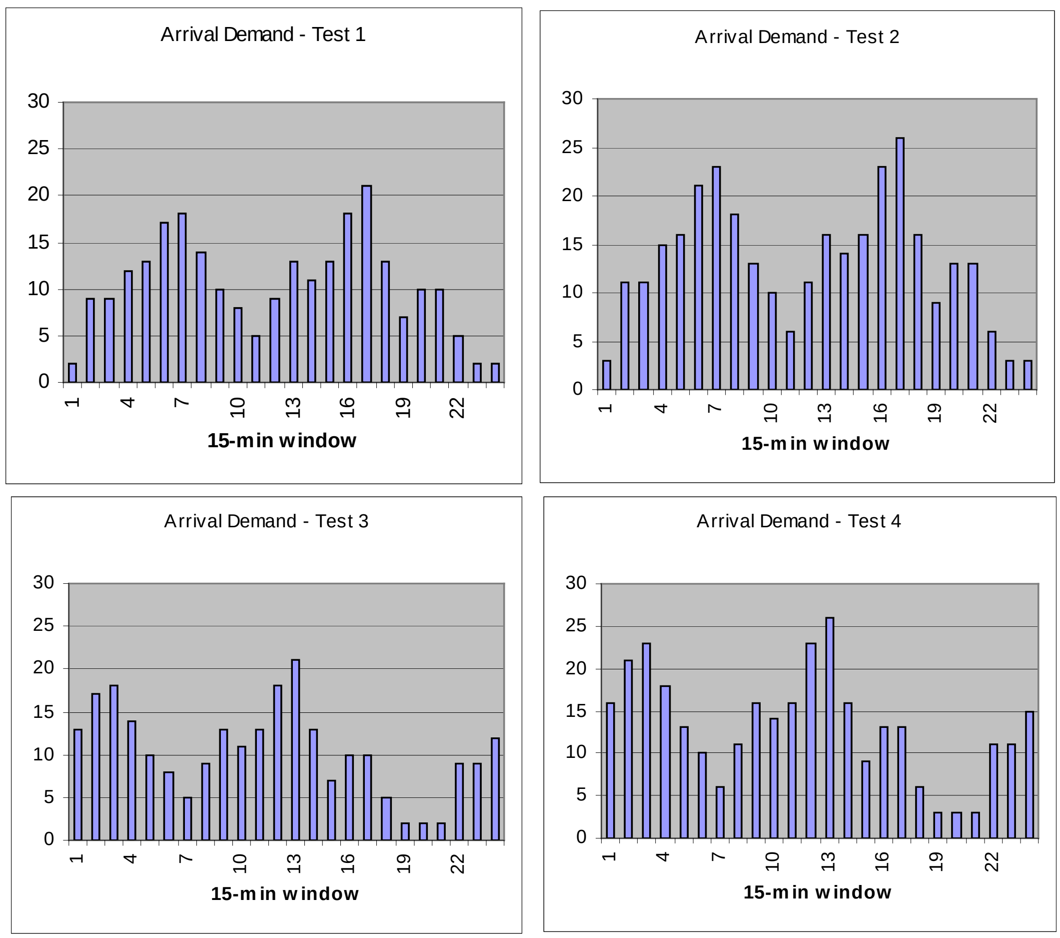


Figure 5–1: Aircraft arrival demand at the capacitated airport.

hub operations in the U.S. This is represented in the cyclical demand for arrivals throughout the period under consideration. Typically, arrivals will dominate the traffic pattern of an airport for approximately one hour, followed by an hour of traffic dominated by departures. The length of these cycles will depend on the airport and airlines operating at the airport. Test 1 and Test 2 use the same underlying arrival pattern, with Test 2 having 25% more arrivals per period. At the beginning of the six-hour period, there is a low arrival demand, indicating that the ground holding policy is being put into place during a departure peak. Test 3 and 4 also use the same distribution as Tests 1 and 2, with a "time shift" to recognize that a ground holding policy is just as likely to begin during a peak arrival flow.

Even though the FAA may impose only one of three polices (VFR, IFR1, IFR2), the actual weather and flight sequencing will further affect an airport's capacity. So, although there may only be three official AARs under a given runway configuration, many more capacity cases will be seen. Consider first the possibility that no inclement weather materializes. We denote this case as Capacity Scenario 0, or CP0. We then include reduced capacity scenarios in sets of three. For each capacity-reduced set, the first scenario represents a particular weather-induced capacity restriction. The second and third scenarios in the set reduce the capacity in each period by an additional 15% and 30%, respectively. Under these scenarios, there will exist no periods in which the nominal or VFR capacity is realized.

Figure 5–2 presents several "bad weather" scenarios and their effect on realized runway capacity. We only show the first scenario from each set of three created.

In some extreme cases, the bad weather may appear twice within one six-hour period, and this is considered in CP16 – CP21. The probabilities associated with these severely affected capacity scenarios are relatively small. We created seven capacity-reduced sets, for a total of 22 capacity scenarios (including the full capacity scenario, CP0) in our stochastic problem. A second set of "bad weather" scenarios (not presented) was also used in the computational experiments.

With all of the arrival capacity scenarios, we have assigned reasonable probabilities. This, of course, is where much of the difficulty of using stochastic programming is seen. Air traffic controllers, FAA, and airline personnel do not have

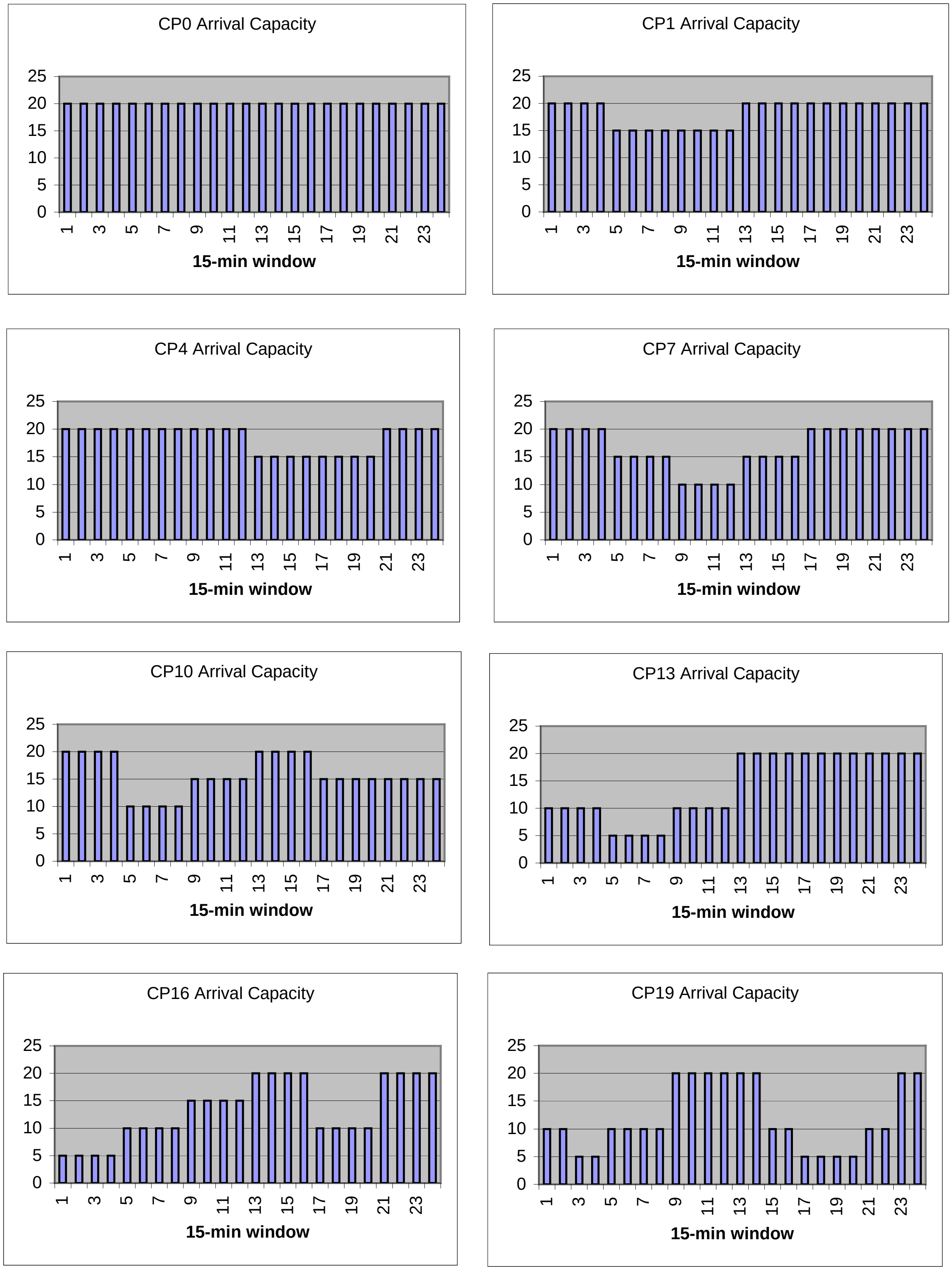


Figure 5–2: Weather-induced arrival capacity scenarios.
(NOTE: Only the first scenario in each set of three is shown.)

historical data to provide them with such capacity scenarios and probabilities. Until this information becomes more readily available, we must make some assumptions about how to obtain such data. Since there would be some debate as to the appropriate probabilities to assign to each scenario, we test three sets of probabilities.

Recall that [SSGHP] uses c_g (or c_a) to denote the unit cost of incurring a ground (or air) delay for one period, t. We evaluate three reasonable estimates for the relative cost of incurring delays on the ground or in the air. Since most of the cost is related to fuel, air delays will usually be much higher. But there may be other negative impacts of holding an aircraft before it takes off. Keeping $c_g = 2$, we create three test cases for $c_a = 3$, 5, and 10. These test cases are also used based on prior experiments conducted and discussed in Ball, et al. [8], Richetta and Odoni [64], and Rifkin [65].

In all, the experiments include four arrival demand profiles, two sets of capacity scenarios, three sets of capacity probabilities, and three ground/air delay ratios, for a total of 72 test problems.

5.3.2 Expected Value of Perfect Information (EVPI) and Value of Stochastic Solution (VSS)

Two key measures to gauge the value of stochastic programming are the expected value of perfect information (EVPI) and the value of the stochastic solution (VSS). EVPI measures the maximum amount a decision maker would be ready to pay in return for complete and accurate information about the future. Using perfect information would enable the decision maker to devise a superior ground holding policy based on knowing what weather conditions to expect. It is not hard to see that obtaining perfect information is not likely. But we can quantify its value to see the importance of having accurate weather forecasts. VSS measures the cost of ignoring uncertainty in making a planning decision. First, the deterministic problem (i.e., the problem that replaces all random variables by their expected values), is solved. Plugging this solution back into the original probabilistic model, we now find the "expected value" solution cost.

This value is compared to the value obtained by solving the stochastic program, resulting in the VSS. Applying this to the GHP, we obtain a solution to the deterministic problem, which provides a set of recommended ground holds per period based on the expected reduction of arrival capacity per period. We then re-solve the stochastic problem using these recommended ground holds as fixed amounts. The difference between the original stochastic solution and the solution using this pre-defined ground holding policy is the VSS. Both the EVPI and VSS measures are typically presented in terms of either unit cost or percent. (We have chosen to show EVPI and VSS as percentage values.) For a more thorough explanation, refer to Birge and Louveaux [15].

We first introduce the four problems that were solved in calculating EVPI and VSS. The "Deterministic Solution" uses an expected arrival capacity per period, $\bar{M}_t$, based on the probability of each weather scenario occurring. Denoting $\bar{M}_t = \sum_{q=1}^{Q} p_q M_{qt}$, we can rewrite [SSGHP] without any scenarios and, thus, without any uncertainty. We present the following formulation:

[DGHP] Deterministic Ground Holding Problem

$$\text{minimize:} \quad c_g \sum_{t=1}^{T} G_t + c_a \sum_{t=1}^{T} W_t \tag{5.8}$$

subject to:

Constraints (5.2),(5.5),

$$\text{Arrival Period Capacities:} \quad A_t + W_{t-1} - W_t \leq \bar{M}_t \quad t = 1, \ldots, T, \tag{5.9}$$

$$\text{Initial Period Air Delays:} \quad W_0 = 0, \tag{5.10}$$

$$\text{Nonnegativity:} \quad A_t, W_t, G_t \geq 0 \quad t = 1, \ldots, T, \tag{5.11}$$

The "Perfect Information Solution" assumes that we know, in advance, the arrival capacity per period. Since we have Q possible capacity scenarios, we solve Q individual problems, setting $\bar{M}_t = M_{qt}$ for each scenario q. Using [DGHP], we determine a minimum cost solution, S_q, for each scenario. Then, we calculate the "Perfect Information Solution" (PIS) by taking the weighted average of the solution values, or PIS $= \sum_{q=1}^{Q} p_q S_q$.

The "Stochastic Solution," our recommended approach, represents the results of solving [SSGHP]. Finally, to calculate the "Expected Value Solution," we will use [SSGHP]. However, we first set the ground delay variables, G_t, and the actual departure variables, A_t, to the values obtained with the "Deterministic Solution." When we solve this version of [SSGHP], we are actually supplying a fixed ground holding plan and observing the additional air delays that result from not taking the randomness of each weather scenario, q, into account explicitly.

Runs were performed across all of the combinations of demand profiles, capacity profiles, probability sets, and ground/air delay ratios. In order to arrive at some summary statistics, the two arrival capacities and three sets of probabilities were grouped together. Thus, each summary test case is an average of six runs. Denote each run's ground/air delay ratio as G2A#, where G2 represents a unit cost of 2 for incurring ground delay and A# represents a unit cost of # for incurring air delay. Each summary test case is then unique based on its arrival demand profile and its ground/air delay ratio (G2A#). Table 5–1 summarizes the results over these groups of test cases.

Both the deterministic and perfect information solutions do not change within a particular arrival demand profile. This indicates that all delays are being taken as ground holds. Since the ground delay cost is less than the air delay cost in each test case, the model will always assign ground delays first, regardless of the magnitude of the capacity reduction. Since deterministic information is not usually

Table 5–1: EVPI and VSS statistics (Minimize total expected delay cost model).

Summary Test Case	Deterministic Solution	Perfect Information Solution	Stochastic Solution	Expected Value Solution	EVPI	VSS
ArrDem1						
G2A3	199	562	821	862	46%	4.9%
G2A5	199	562	1246	1304	122%	4.6%
G2A10	199	562	1874	2409	235%	28.3%
ArrDem2						
G2A3	1034	1459	2033	2154	39%	6.0%
G2A5	1034	1459	2728	2901	87%	6.4%
G2A10	1034	1459	3484	4767	140%	36.8%
ArrDem3						
G2A3	130	602	897	930	49%	3.6%
G2A5	130	602	1387	1464	131%	5.4%
G2A10	130	602	2094	2798	250%	33.3%
ArrDem4						
G2A3	1035	1497	2145	2340	43%	9.1%
G2A5	1035	1497	3007	3210	101%	6.8%
G2A10	1035	1497	3906	5384	162%	37.8%

Note: Delay costs represent unit costs, not monetary amounts.

available, introducing uncertainty through stochastic programming results in solutions with much higher total delay costs. Arrival demand profiles 2 and 4 both increase the amount of traffic arriving to the capacitated airport. This is clearly shown through the large increase in delays, even in the deterministic case.

For the G2A3 cases, the value of the stochastic solution (VSS) is at least 3.6%, which can be quite important given the magnitude in the cost per delay unit. And, as we move to the G2A10 cases, VSS is greater than 28%. For example, in the G2A10 case under Arrival Demand Profile 4, the expected value solution gives a value of 5384, and the expected savings by using a stochastic solution would be 1478. This indicates that, if air delays are expected to be more than five times as costly as ground delays, then evaluating the ground holding policy using stochastic programming is essential. Similarly, with EVPI values ranging from 40% to 250%, it is quite evident that obtaining higher quality weather forecasts would be very beneficial. The EVPI measure can be used as a justification for investing in improved weather forecasting techniques.

5.4 Risk Aversion Measures

5.4.1 Conditional Value at Risk (CVaR) Model

The solution to the SSGHP model is determined by minimizing total expected delay cost over the entire set of scenarios presented. However, there still may be instances where, in certain weather scenarios, the delay incurred as a result of a particular ground holding strategy is much longer than the delay incurred under any other scenario. In this situation, we may want to find solutions that attempt to minimize the spread of delays across all scenarios, or to minimize the extent to which extremely poor outcomes exist. This can be done through the addition of risk aversion measures. Such measures allow us to place a relative importance on other factors of the problem besides total delay. The Value-at-Risk (VaR) measure has been extensively studied in the financial literature. More recently, researchers have discovered that another measure, Conditional Value-at-Risk (CVaR), proves very useful in identifying the most critical or extreme delays from the distribution of potential outcomes, and in reducing the impact that these outcomes have on the overall objective function. For a more detailed description of CVaR and some applications, see Rockafellar and Uryasev [67],[68].

CVaR can be introduced in more than one form for the GHP, depending on the concerns of the airlines, the air traffic controllers, and the FAA. We can define a new objective that focuses on the risk measure, or we can add the risk measure in the form of risk aversion constraints (see Section 5.4.3 for alternate CVaR models). In this section, we present a new formulation that attempts to minimize the expected value of a percentile of the worst-case delays; i.e., we place the CVaR measure in the objective function.

In order to set up the CVaR model, additional variables and parameters are required. Let α represent the significance level for the total delay cost distribution across all scenarios, and let ζ be a decision variable denoting the Value-at-Risk for the model based on the α-percentile of delay costs. In other words, in α% of the scenarios, the outcome will not exceed ζ. Then, CVaR is a weighted measure of ζ and the delays exceeding ζ, which are known to be the worst-case delays. Next, we introduce τ_q to represent the "tail" delay for scenario q. We define "tail" delay as the amount by which total delay cost in a scenario exceeds ζ, which can be represented mathematically as $\tau_q = \text{MAX}\left\{\bar{D}_q - \zeta, 0\right\}$, where $\bar{D}_q = c_g \sum_{t=1}^{T} G_t + c_a \sum_{t=1}^{T} W_{qt}$. The risk aversion problem is now formulated.

[GHP–CVaR] Ground Holding Problem (Conditional Value-at-Risk)

$$
\begin{array}{rlll}
\text{minimize:} & \zeta + (1-\alpha)^{-1} \sum_{q=1}^{Q} p_q \tau_q & & (5.12) \\
\text{subject to:} & & & \\
\text{Schedule Arrival Times:} & \sum_{t=1}^{j} A_t \leq \sum_{t=1}^{j} D_t & j = 1, \ldots, T, & \\
 & \sum_{t=1}^{T+1} A_t = \sum_{t=1}^{T+1} D_t, & & \\
\text{Arrival Period Capacities:} & A_t + W_{q,t-1} - W_{qt} \leq M_{qt} & t = 1, \ldots, T, q = 1, \ldots, Q, & \\
\text{Initial Period Air Delays:} & W_{q0} = 0 & q = 1, \ldots, Q, & \\
\text{Ground Delays:} & G_j + \sum_{t=1}^{j} A_t = \sum_{t=1}^{j} D_t & j = 1, \ldots, T, & \\
\text{Worst-Case Tail Delays:} & \tau_q \geq c_g \sum_{t=1}^{T} G_t + c_a \sum_{t=1}^{T} W_{qt} - \zeta & q = 1, \ldots, Q, & (5.13) \\
\text{Nonnegativity:} & \zeta, \tau_q \geq 0 & q = 1, \ldots, Q, & (5.14) \\
\text{Integrality:} & A_t \in Z^+, W_{qt} \in Z^+, G_t \in Z^+ & t = 1, \ldots, T, q = 1, \ldots, Q. &
\end{array}
$$

This model will actually have an objective function value equal to α-CVaR. In order to compare this solution to the solution provided by [SSGHP], we must still calculate total expected delay cost. Total delay cost, as well as maximum scenario delay cost, can be determined after [GHP–CVaR] is solved.

Operating under this holding policy, we can address the risk involved with incurring extremely long ground and air delays. This may sacrifice good performance under another capacity scenario since the bad weather is not realized under all scenarios. In other words, a capacity scenario that would result in little or no delay may now experience a greater delay based on the holding policy's attempt to reduce the delay under a more severely constrained capacity case. So these differences would need to be dealt with on a case-by-case basis, and we present some alternative models to accommodate the goals of different decision makers in Section 5.4.3.

5.4.2 Minimize Total Delay Cost Model vs. Minimize Conditional Value-at-Risk Model

The CVaR model requires the additional input of a significance level, and $\alpha = 0.9$ is chosen for the analysis. Table 2 presents a comparison of the delay statistics for the Minimize Total Delay Cost Model (SSGHP) and the Minimize Conditional VaR model (GHP–CVaR).

Results of the model comparisons show that in the CVaR model, total expected delay will be increased in order to reduce the worst-case delays across all test cases, which supports our explanation describing the use of risk aversion. By examining the results more closely, we note some interesting findings.

Observe the difference in values for α-VaR and α-CVaR when no risk is modeled. This illustrates the importance of considering the average of worst-case delay costs when you choose to model risk. VaR tends to overlook the differences in delays beyond the critical value, and it may not be able to reduce the worst-case delay costs as effectively. When minimizing α-CVaR in the second model, notice that the α-VaR and α-CVaR values are much closer.

Table 5–2: Overall model comparisons.

	Minimize Total Expected Delay			*Minimize Conditional VaR*		
Summary Test Case	Total Delay Cost	α-VaR	α-CVaR	Total Delay Cost	α-VaR	α-CVaR
ArrDem1						
G2A3	821	1661	2628	1767	1875	2087
G2A5	1246	2337	3943	2059	2222	2471
G2A10	1874	3075	4856	2757	2855	2934
ArrDem2						
G2A3	2033	3667	4777	3426	3505	3717
G2A5	2728	4450	6299	3743	3852	4101
G2A10	3484	4809	6755	4411	4485	4564
ArrDem3						
G2A3	897	2109	3078	1981	2139	2378
G2A5	1387	3087	4703	2371	2531	2749
G2A10	2094	3209	5624	2960	3166	3225
ArrDem4						
G2A3	2145	4315	5520	3878	3953	4192
G2A5	3007	5544	7552	4263	4345	4563
G2A10	3906	5092	7624	4869	4980	5039

Note: Delay costs represent unit costs, not monetary amounts.

Also, the percentage increase in total expected delay cost between the two models is more drastic for smaller air delay costs. But as the air delay cost rises, the total delay cost incurred when minimizing α-CVaR is not severely affected. Consider the following results observed for the Arrival Demand 4 profile. The G2A3 case experiences an increase in average delay cost of 80%, while the G2A10 case experiences only a 25% increase. The magnitude of air delay costs will significantly impact the effectiveness of using risk constraints. Recall from Table 5–1 the example that was previously described. In the G2A10 case the expected value solution gives a value of 5384, and the expected savings without risk constraints would be 1478. Now, by minimizing the 10% worst-case delays (using the GHP–CVaR model), the expected savings reduces to 515. But we also have reduced the worst-case delays from 7624 to 5039.

Since the CVaR analysis up to this point only considers using $\alpha = 0.9$, it is worthwhile to show how CVaR can shape the distribution of outcomes at another significance level. Consider $\alpha = 0.75$, which implies that the 25%-largest delay costs are minimized. Figure 5–3 shows the distribution of delays across all scenarios for a particular test case (ArrDem 1, G2A5, with arrival capacity 1 and probability set 1). We present the delay distributions with and without CVaR, and at both levels of significance for the CVaR model. Notice how the delays by scenario vary greatly when minimizing total expected delay. In fact, there are three scenarios which would result in delay costs exceeding 4500. By minimizing α-CVaR, these long delays would not occur. The trade-off with this modeling approach is that there are no scenarios with only minimal delays. With further adjustments to the value of α, the decision makers have some control over how the resulting delay outcomes would look. It then depends on what the decision makers are willing to accept. This is one of the underlying powers of introducing risk aversion such as minimizing α-CVaR.

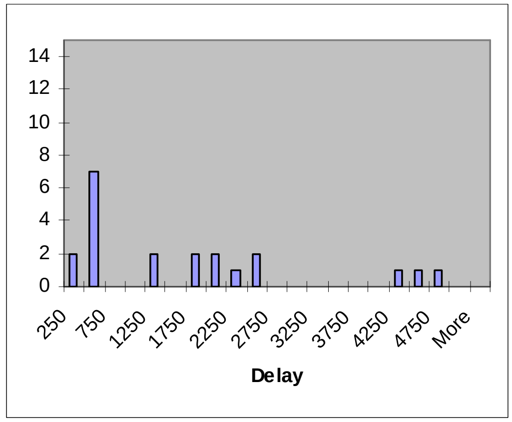
5.4.3 Alternate Risk Aversion Models

Depending on the input from each group involved in constructing a ground holding policy, there will be conflicting desires to reduce total expected delay and to reduce the worst-case outcomes. For these purposes, we can actually choose among several risk aversion models. If your sole desire were to reduce the total expected delay cost, you would not require the use of risk aversion. But if you want to reduce maximum delay costs, you might use the GHP–CVaR model. For other cases, which will account for most collaborative efforts, some combination of these models will be chosen.

For this reasoning, we introduce a more general formulation of the ground holding problem. Now, we consider that the objective may be to minimize total expected delay, to minimize worst-case delay (α-CVaR), or to minimize some combination of these measures. We present the first alternate GHP formulation as:

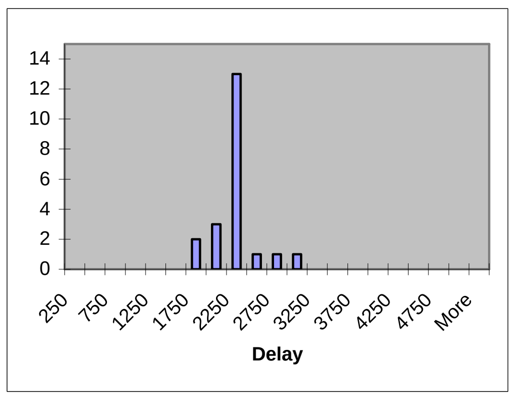
Minimize Total Expected Delay Cost

Total Exp. Delay Cost = 1114



Minimize Conditional VaR (alpha = 0.90)

Total Exp. Delay Cost = 2063



Minimize Conditional VaR (alpha = 0.75)

Total Exp. Delay Cost = 1609

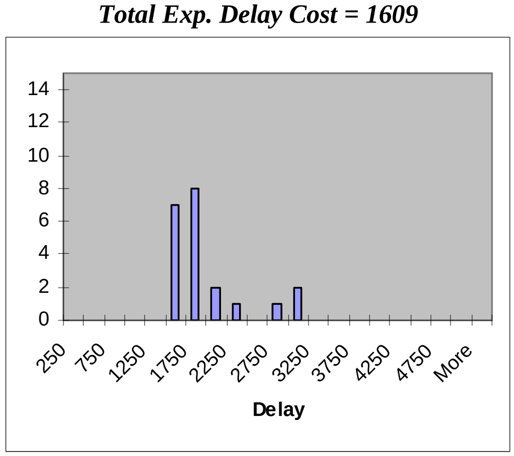


Figure 5–3: Total delay output for arrival demand level 1.

[GHP–CVaR1] Alternate GHP Risk Formulation 1

$$\text{minimize:} \quad w_1\left(c_g\sum_{t=1}^{T}G_t + c_a\sum_{q=1}^{Q}\sum_{t=1}^{T}p_qW_{q,t}\right) + w_2\left(\zeta + (1-\alpha)^{-1}\sum_{q=1}^{Q}p_q\tau_q\right) \tag{5.15}$$

subject to:

$$\text{Constraints } (2-6, 13-14).$$

By including weights w_1 and w_2, the decision maker can have complete control over the importance of each objective measure. Note that for $w_1 = 0$ and $w_2 =$ any constant, we have the special case of [GHP–CVaR]. Likewise, for $w_1 =$ any constant and $w_2 = 0$, we have the special case of [SSGHP].

We introduce a second alternate formulation that imposes a restriction on allowable losses. We use the original objective function from [SSGHP], minimizing the expected total delay cost, while satisfying a constraint requiring the percentile of worst-case delays to be no more than some parameter, v.

[GHP–CVaR2] Alternate GHP Risk Formulation 2

$$\text{minimize:} \quad c_g\sum_{t=1}^{T}G_t + c_a\sum_{q=1}^{Q}\sum_{t=1}^{T}p_qW_{qt}$$

subject to:

$$\text{Constraints } (2-6, 13-14),$$

$$\text{Worst-Case Delay Bound:} \quad \zeta + (1-\alpha)^{-1}\sum_{q=1}^{Q}p_q\tau_q \le v. \tag{5.16}$$

By placing an upper bound on a loss function, as in (5.16), it approaches a maximum loss constraint. But some scenarios can actually exceed this parameter value, as long as the weighted average of losses within the percentile remains below v. In Table 5–3, we provide an illustration of the effect of using each model to set the ground holding policy.

Notice that SSGHP provides the lowest expected total delay cost, based on considering the likelihood of each weather scenario actually occurring. On the other

Table 5–3: Performance comparison of alternate risk models.

Model	Total Expected Delay Cost	α-VaR	α-CVaR	Maximum Delay Cost
SSGHP	3336	4890	7354	8570
GHP–CVaR	4325	4446	4521	5076
GHP–CVaR1 ($w_1 = w_2$)	3746	4136	4703	5696
GHP–CVaR2 ($v = 5000$)	3566	4068	5000	5898
GHP–CVaR2 ($v = 6000$)	3383	4332	6000	6882

Note: Results are based on (ArrDem 2,G2A10,Arrival Capacity 1,Probability Set 1).

hand, GHP–CVaR produces the best value of α-CVaR and the lowest maximum delay cost in any scenario. The tradeoff is that the more likely scenarios will now encounter increased ground holdings. Combining these two objectives with GHP–CVaR1, we gain a substantial amount of the benefit of the previous two models, with total expected delay cost at 3746 and maximum scenario delay cost at 5696. And we can even fine-tune our objective further through the use of the CVaR constraint. As α-CVaR is increased, we approach our original SSGHP model.

In addition to the above risk models, Rifkin [65] briefly presents the Maximum Air Delay Model (MADM). MADM can be thought of as a maximum-loss constraint for any scenario, and if such a number exists, this could be added to any of the above formulations. What MADM fails to address is the continuing effort to minimize total delays. A max-loss constraint can be added to any of the formulations presented in this paper, allowing the user additional insight into a particular airport's ground holding policies. As with the parameter v, setting the maximum loss too tight may prevent the model from finding a feasible solution.

There is no one answer when deciding which problem formulation to use. Each will shape the resulting total delay in different ways, and thus it is dependent on the groups making the decisions in determining the amount of acceptable delay.

5.5 Conclusions

As we have shown in this chapter, the airport ground holding problem is essentially a demand selection problem in which we desire the optimal allocation of flight arrival requests to the airport's runway capacity, the limited resource. By selecting when to hold aircraft at their upline stations, we can reduce the quantity of costly air delays while maintaining some desired level of service.

We have also shown that modeling the ground holding problem as a stochastic problem is most certainly beneficial. Even under cases when delay costs are low and uniform, the value of the stochastic solution is significant. Additionally, introducing risk aversion allows a decision maker to offer several potential outcomes based on various worst-case delay scenarios. The FAA, airport authorities, and airlines all work within some pre-defined performance measures, and providing a model that allows constraints to be adjusted to meet such performance measures is very important.

There are several issues that were not addressed and are areas for future research. By modeling the problem at the individual flight detail, we may be able to gain more accuracy in determining the true capacity and realized arrival flows into an airport. Additional research is still required to determine whether the added benefits of unique flight information merit the undertaking of working with a more complex model. Once the model is at the flight level, arrival sequencing, banking, and other arrival/departure disruptions can be modeled. We would also like to incorporate specific flight durations to more accurately represent the traffic for the airport in question.

As we briefly mentioned, the nature of a static GHP model does not allow for system updates to airport capacity, which in reality, are very likely to occur. We handle this by simply re-solving the static model as time progresses. Alternatively, we would like our first ground holding decision to take into account the possible changes to future weather conditions, and developing a dynamic model would provide this additional level of forecasting detail.

Considering the originating stations of the aircraft could also be worthwhile. A multiple airport model would be able to provide more realistic information on the decisions and actions of each individual departing aircraft en route to the capacitated airport under study. However, these models will grow in size quickly, even under the assumption that airport capacities are deterministic, as in Navazio and Romanin-Jacur [57].

Finally, Collaborative Decision Making (CDM), described in Ball, et al. [7], has been an area of focus recently. It allows airlines to be involved in the decisions on which aircraft will be delayed during a ground holding plan. This is likely to achieve a reduction in overall costs to individual airlines by allowing "more critical" aircraft to take off at their scheduled departure times and not incur ground holding delays. CDM may be more difficult to model, but it is important to include this fundamental interactive approach in order to represent or simulate the actual environment.

CHAPTER 6
CONCLUDING REMARKS

When a decision maker has discretion to accept or deny demand sources, especially when facing limited resources, determining the best set of demands to select based on the resulting revenue and production/delivery costs can be quite challenging. In order to adequately address this challenge, we studied a wide variety of demand selection problems. The production-related problems included uncapacitated and capacitated versions of the order selection problem, demand selection with pricing as a decision variable, and also stochastic demand selection problems that allow the decision maker to influence demand through marketing effort. We then tied these efforts in with the applications area of airport operations. The demand selection problem, especially in the manufacturing setting, has gone relatively unnoticed in the literature until recently. We have provided a thorough discussion of a family of models that exist in this area.

The models we present serve as a starting point for future research on more general models. We first address future research areas specific to the multi-period problems presented in Chapters 2 and 3. Suppose that instead of picking and choosing individual orders by period, the producer must satisfy a given customer's orders in *every* period if the producer satisfies that customer's demand in any single period. In other words, a customer cannot be served only when it is desirable for the producer, since this would result in poor customer service. We might also pose a slightly more general version of this problem, which requires serving a customer in some contiguous set of time periods, if we satisfy any portion of the customer's demand. This would correspond to contexts in which the producer is free to begin serving a market at any time and can later stop serving the market at any time in the planning horizon; however, full service to the market must continue between the start and end period chosen.

Future research might also consider varying degrees of producer flexibility (where certain minimum order fulfillment requirements must be met) or modeling more complex market interactions and price as a decision variable (where market demand is a function of the producer's price and/or the price offered by competitors). We might also consider a situation in which the producer can acquire additional capacity (either in terms of production capacity or marketing budget) at some cost in order to accommodate more orders than current capacity levels allow. These generalizations of the order selection models may further increase net profit from integrated order selection, capacity planning, and production planning decisions.

Some of the major areas for future research in the airport operations ground holding context will focus on modeling the individual demand sources (or flights) at a finer level of detail. As was shown for the demand selection problems in Chapters 2 – 4, the unique characteristics of each demand source can significantly influence the desirability of fulfilling the demand source. The same will hold true for determining which aircraft are most suitable for ground holds. We also would like to introduce capacity updates to the ground holding problem, which implies that our first ground holding decision would take into account the possible changes to future weather conditions. However, as we include these additional details, our demand selection problem becomes a multi-period, dynamic stochastic decision problem, which could prove to be very difficult to solve. And, for this reason, it also remains a very interesting topic for future research.

REFERENCES

[1] P. Afentakis and B. Gavish. Optimal lot-sizing algorithms for complex product structures. *Operations Research*, 34:237–249, 1986.

[2] P. Afentakis, B. Gavish, and U. Karmarkar. Computationally efficient optimal solutions to the lot-sizing problem in multistage assembly systems. *Management Science*, 30:222–239, 1984.

[3] R.K. Ahuja, T.L. Magnanti, and J.B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice-Hall, Englewood Cliffs, New Jersey, 1993.

[4] K.R. Baker. Requirements planning. In S.C. Graves, A.H.G. Rinnooy Kan, and P.H. Zipkin, editors, *Handbooks in Operations Research and Management Science; volume 4: Logistics of Production and Inventory*, chapter 11. North-Holland, Amsterdam, The Netherlands, 1993.

[5] K.R. Baker, P. Dixon, M. Magazine, and E. Silver. An algorithm for the dynamic lot size problem with time-varying production capacity constraints. *Management Science*, 24:1710–1720, 1978.

[6] A. Balakrishnan and J. Geunes. Production planning with flexible product specifications: An application to specialty steel manufacturing. *Operations Research*, 51(1):94–112, 2003.

[7] M. Ball, R. Hoffman, C. Chen, and T. Vossen. Collaborative decision making in air traffic management: Current and future research directions. Technical Report T.R. 2000-3, NEXTOR, University of Maryland, 2000.

[8] M. Ball, R. Hoffman, A. Odoni, and R. Rifkin. The static stochastic ground holding problem with aggregate demands. Technical Report T.R. 99-1, NEXTOR, University of Maryland and Massachusetts Institute of Technology, 1999.

[9] I. Barany, J. Edmonds, and L.A. Wolsey. Packing and covering a tree by subtrees. *Combinatorica*, 6:245–257, 1986.

[10] I. Barany, T.J. Van Roy, and L.A. Wolsey. Strong formulations for multi-item capacitated lot sizing. *Management Science*, 30(10):1255–1261, 1984.

[11] M.S. Bazaraa, H.D. Sherali, and C.M. Shetty. *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons, New York, New York, 2nd edition, 1993.

[12] P.P. Belobaba. Airline yield management: An overview of seat inventory control. *Transportation Science*, 21:63–73, 1987.

[13] S. Bhattacharjee and R. Ramesh. A multi-period profit maximizing model for retail supply chain management: An integration of demand and supply-side mechanisms. *European Journal of Operational Research*, 122:584–601, 2000.

[14] S. Biller, L.M.A. Chan, D. Simchi-Levi, and J. Swann. Dynamic pricing and the direct-to-customer model in the automotive industry. Working paper, School of Industrial and Systems Engineering, Georgia Institute of Technology, 2003.

[15] J. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer Series in Operations Research, New York, New York, 1997.

[16] S. Brealy and S. Myers. *Principles of Corporate Finance*. Irwin McGraw-Hill, New York, New York, 7th edition, 2002.

[17] M. Burman. Dynamic pricing applied to manufacturing. In *Proceedings of INFORMS Conference on OR/MS Practice (Creating Value in the Extended Enterprise)*, Phoenix, AZ, 2003. INFORMS.

[18] C. Cachon. Competitive supply chain inventory management. In S. Tayur, M. Magazine, and R. Ganeshan, editors, *Quantitative Models for Supply Chain Management*. Kluwer Academic Publishers, Boston, Massachusetts, 1999.

[19] S. Carr and I. Duenyas. Optimal admission control and sequencing in a make-to-stock/make-to-order production system. *Operations Research*, 48(5):709–720, 2000.

[20] S. Carr and W. Lovejoy. The inverse newsvendor problem: Choosing an optimal demand portfolio for capacitated resources. *Management Science*, 46(7):912–927, 2000.

[21] L.M.A. Chan, D. Simchi-Levi, and J. Swann. Dynamic pricing strategies for manufacturing with stochastic demand and discretionary sales. Working paper, School of Industrial and Systems Engineering, Georgia Institute of Technology, 2003.

[22] S. Chopra and P. Meindl. *Supply Chain Management: Strategy, Planning , and Operations*. Prentice-Hall, Upper Saddle River, New Jersey, 2nd edition, 2004.

[23] D. Erlenkotter. A dual-based procedure for uncapacitated facility location. *Operations Research*, 26:992–1009, 1978.

[24] A. Federgruen and M. Tzur. A simple forward algorithm to solve general dynamic lot-sizing models with n periods in $O(n \log n)$ or $O(n)$ time. *Management Science*, 36(4):449–466, 1990.

[25] C.H. Fine and R.M. Freund. Optimal investment in product-flexible manufacturing capacity. *Management Science*, 36(4):449–466, 1990.

[26] M.L. Fisher. The Lagrangian relaxation method for solving integer programming problems. *Management Science*, 27(1):1–18, January 1981.

[27] M.L. Fisher, A. Raman, and A. McClelland. Rocket science retailing is almost here–are you ready? *Harvard Business Review*, July-August 2000.

[28] M. Florian and M. Klein. Deterministic production planning with concave costs and capacity constraints. *Management Science*, 18:12–20, 1971.

[29] M. Florian, J.K. Lenstra, and A.H.G. Rinnooy Kan. Deterministic production planning: Algorithms and complexity. *Management Science*, 26(7):669–679, 1980.

[30] G. Gallego and G. van Ryzin. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Science*, 40(8):999–1020, 1994.

[31] G. Gallego and G. van Ryzin. A multi-product dynamic pricing problem and its applications to network yield management. *Operations Research*, 45(1):24–41, 1997.

[32] A.M. Geoffrion. Lagrangian relaxation for integer programming. *Math. Prog. Study*, 2:82–114, 1974.

[33] S.M. Gilbert. Coordination of pricing and multi-period production for constant priced goods. *European Journal of Operational Research*, 114:330–337, 1999.

[34] L. Grignon. Analyses of delay in an air traffic system with weather uncertainty. Master's thesis, University of Washington, December 2002.

[35] A.Y. Ha. Inventory rationing in a make-to-stock production system with several demand classes and lost sales. *Management Science*, 43(8):1093–1103, 1997.

[36] R. Hoffman and M. Ball. A comparison of formulations for the single-airport ground holding problem with banking constraints. Technical Report T.R. 98-44, NEXTOR, University of Maryland, 1998.

[37] R. Horst and H. Tuy. *Global Optimization: Deterministic Approaches*. Springer-Verlag, Berlin, Germany, 1990.

[38] W. Huang, H.E. Romeijn, and J. Geunes. Continuous-time single-sourcing problem with capacity constraints. Technical Report 2004-2, Industrial and Systems Engineering Department, University of Florida, Gainesville, Florida, 2004.

[39] A.V. Iyer and M.E. Bergen. Quick response in manfacturer-retailer channels. *Management Science*, 43(4):559–570, 1997.

[40] J. Johansson. Advertising and the s-curve: A new approach. *Journal of Marketing Research*, 16:346–354, 1979.

[41] H. Kunreuther and L. Schrage. Joint pricing and inventory decisions for constant priced items. *Management Science*, 19(7):732–738, 1973.

[42] E.L. Lawler. *Combinatorial Optimization: Networks and Matroids*. Holt, Rinehart, and Winston, New York, New York, 1976.

[43] C.Y. Lee, S. Çetinkaya, and A.P.M. Wagelmans. A dynamic lot sizing model with demand time windows. *Management Science*, 47(10):1384–1395, 2001.

[44] H.L. Lee. Ultimate enterprise value creation usning demand-based management. Report SGSCMF-W1-2001, Stanford Global Supply Chain Management Forum, 2001.

[45] H.L. Lee and S. Nahmias. Single-product, single-location models. In S.C. Graves, A.H.G. Rinnooy Kan, and P.H. Zipkin, editors, *Handbooks in Operations Research and Management Science; volume 4: Logistics of Production and Inventory*, chapter 1. North-Holland, Amsterdam, The Netherlands, 1993.

[46] H.L. Lee, V. Padmanabhan, and S. Whang. Information distortion in a supply chain: The bullwhip effect. *Management Science*, 43(4):546–558, 1997.

[47] J. Leung, T.L. Magnanti, and R. Vachani. Facets and algorithms for capacitated lot sizing. *Mathematical Programming*, 45:331–359, 1989.

[48] G.L. Lilien, P. Kotler, and K.S. Moorthy. *Marketing Models*. Prentice-Hall, Englewood Cliffs, New Jersey, 1992.

[49] M. Loparic. *Stronger Mixed-01 Models for Lot-Sizing Problems*. PhD thesis, Université catholique de Louvain, 2001.

[50] M. Loparic, Y. Pochet, and L.A. Wolsey. The uncapacitated lot sizing problem with sales and safety stocks. *Mathematical Programming*, 89:487–504, 2001.

[51] S. Love. A facilities in series inventory model with nested schedules. *Management Science*, 18:327–338, 1972.

[52] V. Mahajan and E. Muller. Advertising pulsing policies for generating awareness for new products. *Marketing Science*, 5(2):89–106, 1986.

[53] H.M. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

[54] G.N. Monahan, N.C. Petruzzi, and W. Zhao. The dynamic pricing problem from a newsvendor's perspective. *Manufacturing and Service Operations Management*, 6(1):73–91, Winter 2004.

[55] I. Moon and E.A. Silver. The multi-item newsvendor problem with a budget constraint and fixed ordering costs. *Journal of the Operational Research Society*, 51:602–608, 2000.

[56] S. Nahmias. *Production and Operations Analysis.* Irwin McGraw-Hill, Boston, Massachusetts, 4th edition, 2001.

[57] L. Navazio and G. Romanin-Jacur. The multiple connections multi-airport ground holding problem: Models and algorithms. *Transportation Science*, 32(3):268–276, 1998.

[58] G.L. Nemhauser and L.A. Wolsey. *Integer and Combinatorial Optimization.* John Wiley & Sons, New York, New York, 1988.

[59] N.C. Petruzzi and M. Dada. Pricing and the newsvendor problem: A review with extensions. *Operations Research*, 74:183–194, 1999.

[60] N.C. Petruzzi and G.E. Monahan. Managing fashion goods inventories: Dynamic recourse for retailers with outlet stores. *IIE Transactions*, 35:1033–1047, 2003.

[61] Y. Pochet. Valid inequalities and separation for capacitated economic lot sizing. *Operations Research Letters*, 7:109–116, 1988.

[62] E.L. Porteus. Stochastic inventory theory. In *Handbooks in Operations Research and Management Science; Volume 2*, chapter 12, pages 605–652. North-Holland, Amsterdam, The Netherlands, 1990.

[63] H. Rhim, T.H. Ho, and U.S. Karmarkar. Competitive location, production, and market selection. *European Journal of Operational Research*, 149:211–228, 2003.

[64] O. Richetta and A. Odoni. Solving optimally the static ground-holding policy problem in air traffic control. *Transportation Science*, 27(3):228–238, 1993.

[65] R. Rifkin. The static stochastic ground holding problem. Master's thesis, Massachusetts Institute of Technology, 1998.

[66] R.T. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, New Jersey, 1970.

[67] R.T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *The Journal of Risk*, 2(3):21–41, 2000.

[68] R.T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking and Finance*, 26(7):1443–1471, 2002.

[69] H.E. Romeijn, J. Geunes, and K. Taaffe. Solution methods for a class of nonlinear, non-separable knapsack problems. Working paper, Industrial and

Systems Engineering Department, University of Florida, Gainesville, Florida, 2004.

[70] 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, *Handbooks in Operations Research and Management Science; volume 4: Logistics of Production and Inventory*, chapter 8. North-Holland, Amsterdam, The Netherlands, 1993.

[71] Z.J. Shen, C. Coullard, and M. Daskin. A joint location-inventory model. *Transportation Science*, 37(1):40–55, 2003.

[72] E.A. Silver and H.C. Meal. A heuristic for selecting lot size requirements for the case of a deterministic time-varying demand rate and discrete opportunities for replenishment. *Production and Inventory Management Journal*, 14(2):64–74, 1973.

[73] H. Simon. ADPULS: An advertising model with wearout and pulsation. *Journal of Marketing Research*, 19:352–363, 1982.

[74] J. Thomas. Price-production decisions with deterministic demand. *Management Science*, 16(11):747–750, 1970.

[75] J. Thomas. Price and production decisions with random demand. *Operations Research*, 22(3):513–518, 1974.

[76] A. Tsay, S. Nahmias, and N. Agrawal. Modeling supply chain contracts: A review. In S. Tayur, M. Magazine, and R. Ganeshan, editors, *Quantitative Models for Supply Chain Management*. Kluwer Academic Publishers, Boston, Massachusetts, 1999.

[77] U.S. House Subcommittee on Aviation. *Aviation Congestion is Reaching Crisis Proportions, Congressional Oversight Hearing*, April 2001.

[78] D. Vakratsas, F. Feinberg, F. Bass, and G. Kalyanaram. The shape of advertising response functions revisited: A model of dynamic probabilistic thresholds. *Marketing Science*, 23(1):109–119, Winter 2004.

[79] G. van Ryzin and J. McGill. Revenue management wihout forecastings or optimization: An adaptive algorithm for determining airline seat protection levels. *Management Science*, 46(6):760–775, 2000.

[80] P. Vranas, D. Bertsimas, and A. Odoni. Dynamic ground-holding policies for a network of airports. *Transportation Science*, 28(4):275–291, 1994.

[81] P. Vranas, D. Bertsimas, and A. Odoni. The multi-airport ground-holding problem in air traffic control. *Operations Research*, 42(2):249–261, 1994.

[82] A. Wagelmans, S. van Hoesel, and A. Kolen. Economic lot sizing: an $O(n \log n)$ algorithm that runs in linear time in the Wagner-Whitin case. *Operations Research*, 40-S1:S145–S156, 1992.

[83] H. Wagner and T. Whitin. Dynamic version of the economic lot size model. *Management Science*, 5:89–96, 1958.

[84] E.L. Williamson. *Airline Network Seat Inventory Control: Methodologies and Revenue Impacts*. PhD thesis, Massachusetts Institute of Technology, 1992.

[85] C.A. Yano. Setting planned leadtimes in serial production systems with tardiness costs. *Management Science*, 33(1):95–106, 1987.

[86] W.I. Zangwill. A backlogging model and a multi-echelon model of a dynamic economic lot sizing problem. *Management Science*, 15(9):506–527, 1969.

BIOGRAPHICAL SKETCH

I was born on April 27, 1966, in Berwyn, Illinois. After maintaining a steady interest in mathematics throughout middle school and high school, I attended the University of Illinois at Urbana-Champaign, where I received a B.S. and M.S. in industrial engineering in 1988 and 1990. I began my professional career at American Airlines Decision Technologies (AADT) in Fort Worth, Texas. I consulted on a variety of projects concerning air and passenger transportation systems, using simulation and other modeling tools to aid in making project recommendations. In 1994, the technology groups were reorganized under the name Sabre, a company independent from American Airlines. In addition to American, our clients included airport authorities, airport boards, governmental agencies, package delivery companies, hotels, car rental agencies, as well as other airlines. I spent eight years with AADT and Sabre, serving in roles from consultant to senior project manager.

In search of deeper professional reward, I took steps in a new direction in 1998, working as an Instructor in the Department of Industrial Engineering at Northern Illinois University. I worked there for two years and gained invaluable experience, teaching several different undergraduate courses. Realizing I still had one final step to take, I chose to leave NIU in 2000 and pursue a Ph.D. in industrial and systems engineering at the University of Florida. I was awarded a Stephen C. O'Connell Presidential Fellowship for the duration of the doctoral program. During my graduate studies, I acquired a new area of expertise in production and inventory control and supply chain management, and these specializations now complement my existing experience and interest in transportation and logistics.