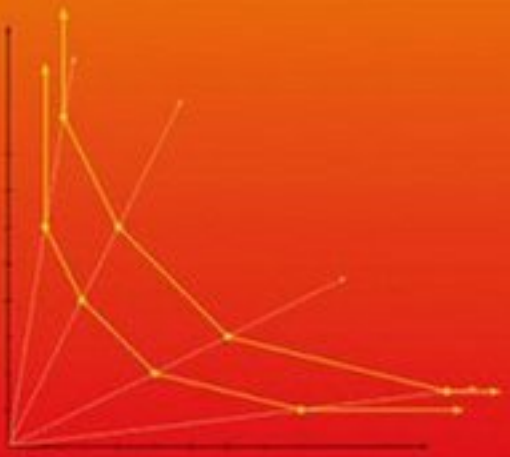
STEVEN T. HACKMAN

Production Economics

Integrating the Microeconomic and Engineering Perspectives



Springer

Production Economics

Steven T. Hackman

Production Economics

Integrating the Microeconomic and Engineering Perspectives

Springer

Steven T. Hackman
Georgia Institute of Technology
Atlanta, GA 30332-0205
USA
shackman@isye.gatech.edu

ISBN 978-3-540-75750-4 e-ISBN 978-3-540-75751-1

DOI 10.1007/978-3-540-75751-1

Library of Congress Control Number: 2007938621

Typesetting and Production: LE-TEX Jelonek, Schmidt & Vckler GbR, Leipzig, Germany
Cover design: WMX Design GmbH, Heidelberg, Germany

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

To my family

Preface

A production economist focuses on assessment. Talk to production economists about a particular firm, and they are likely to ask questions such as: How efficient is the firm in utilizing its input to produce its output? Is the firm using the right mix of inputs or producing the right mix of outputs given prevailing prices? How will the firm respond to a price hike in a critical input? How efficient is the firm in scaling its operations? Has the firm improved its productive capability over time? How does the firm compare to its competitors? A production economist will use an aggregate description of technology to answer these questions.

A production engineer focuses on optimizing resources. Talk to production engineers about a particular firm, and they are likely to ask a completely different set of questions: How can the firm change its operations to be more productive? Should the firm outsource production of a subassembly or should it be made in-house? How can the firm reduce its production lead times? Should resource capacity be expanded and, if so, which resources should be acquired? Can the firm's products be redesigned to improve productive efficiency? Which plants (or operations within a plant) should produce which products at what times? A production engineer will use a detailed description of technology to answer these questions.

Historically, production economists and engineers do not interact with each other. This is counterproductive because each group could benefit from the other group's perspective. Production engineers should be interested in the production economists' questions and the tools they use to answer them. To more accurately answer their questions, production economists should be interested in adopting the micro-level descriptions of technology used by production engineers. The topics in this book bring together under one roof the economics perspective, which focuses on assessment using an aggregate description of technology, and the engineering perspective, which focuses on optimizing resources using a detailed description of technology. *This book offers a unified, integrated point of view that bridges the gap between these two historically distinct perspectives.*

Organization Of This Book

The topics in the book are organized into four main parts:

- Part I: Microeconomic Foundations
- Part II: Efficiency Measurement
- Part III: Productivity and Performance Measurement
- Part IV: Engineering Models of Technology

Due to the integrative and technical nature of the topics presented herein, I have included an extensive Mathematical Appendix (Part V) that covers (and in some cases expands) the core material needed to understand the topics in this book. A detailed overview of the book's topics is provided in Chapter 1.

I emphasize computational and graphical approaches to develop and illustrate the concepts. Computational methods and characterizations facilitate both practical applications and presentation of results. To make the material accessible, I provide many examples, figures, and detailed derivations. I have included many exercises. Some exercises are relatively straightforward; they ask the reader to confirm his or her knowledge of the core concepts presented. Other exercises are quite challenging; they ask the reader to extend the core material in nontrivial ways. To assist the reader, detailed solutions to the exercises are provided. The usual caveat applies: it is best to try the problems first!

Intended Audience

For many years I have taught a graduate course in productivity measurement and analysis at Georgia Tech's School of Industrial and Systems Engineering, a course that has drawn students with diverse backgrounds or majors in engineering, operations research, operations management, economics, and mathematics. The material in this book grew out of the lectures I developed to accommodate their many interests.

Engineering and operations research students study a variety of detailed system representations in their course work and practice. They have modest exposure to topics in Part I, no exposure to the topics in Part II and III, and some exposure to the topics in Part IV (if they took a production operations course). My experience teaching these students suggests they enjoy learning the types of questions economists pose, and the aggregate models the economists use to obtain answers. The models in Part IV go well beyond the textbook presentations they learn in other classes.

Economics and operations management students are familiar with some of the topics in Part I, and have limited exposure to the topics in Parts II–IV. My experience teaching these students suggests that they enjoy learning about the nonparametric models of technology discussed in Part I (not emphasized in traditional microeconomics courses), learning about the practical applications

of economic ideas and computational methods presented in Parts II and III, and being exposed to the engineer's perspective on modeling emphasized in Part IV.

Mathematics students have virtually no exposure to the main topics in this book. My experience teaching these students suggests that they enjoy learning of applications of real and convex analysis, and the style in which the topics are presented.

This book can also be beneficial to professionals on two fronts. Productivity handbooks, each with a different focus, describe practical issues and provide basic formulae that emphasize partial productivity. The contents of Parts I–III, with their emphasis on formal description, graphical representation, and computation, provide context for the formulae and models. This should assist a practitioner to decide which are best suited for the application at hand. With today's information systems and the availability of shop-floor data, it is now possible to develop and implement detailed, dynamic multistage models of technology described in Part IV. These models have been motivated by and applied to real production systems.

Acknowledgments

I would like to thank the many students at Georgia Tech who were willing to learn material outside the traditional industrial engineering curriculum. This book has benefited enormously from their suggestions, comments, and the many typos they found. In particular, I single out Dan Adelman (University of Chicago), Dan Faissol, Dagoberto Garza-Nunez (Tencolgico de Monterrey, Mexico), Yahel Giat (Jerusalem College of Technology, Israel), Marco Gutierrez, Ray Hagtvedt, Nancy Lillo, Michael Pennock, Ray Popovic, Eva Regnier (Naval Post Graduate School), German Riano-Mendoza (Universidad de los Andes, Colombia), Doug Thomas (Penn St. University), Dimitra Vlatsa, Charles Wardell, and Charles Whatley.

During my graduate years at University of California—Berkeley, Kingtim Mak (University of Illinois—Chicago) introduced me to the production economics field and to Ron Shephard, my first thesis advisor. After Shephard's passing, Rob Leachman took over as my advisor. I am indebted to all three individuals, each of whom profoundly influenced my thinking on this subject.

Several of my colleagues at Georgia Tech deserve special mention. Loren Platzman and Dick Serfozo provided extensive editorial suggestions and advice throughout the preparation of this book. In addition to Loren and Dick, conversations with John Bartholdi, Paul Griffin, Craig Tovey, and Alex Shapiro have always proved fruitful and enlightening.

During most of my academic career, I have had the privilege of collaborating with a wise and generous colleague, Ury Passy from the Technion-Israel Institute of Technology. I have made many visits to the Faculty of Industrial Engineering and Management at the Technion. It is a pleasure to acknowledge

their friendship, extraordinary kindness, and support. In particular, I single out Aharon Ben-Tal, Zvi First, Boaz Golany, Michael Mizrach, Dan Peled (Haifa University), the late Meir Rosenblatt, Ishay Weisman, Gideon Weiss (Haifa University), and their spouses.

It is also a pleasure to acknowledge colleagues from whom I have received encouragement, feedback, and a bit of wisdom. From the School of Industrial and Systems Engineering at Georgia Tech: Faiz Al-Khayyal, Earl Barnes, Bill Cook, Jim Dai, Bob Foley, Dave Goldsman, John Jarvis, Leon McGinnis, Gary Parker, Joel Sokol, Mike Thomas, Jerry Thuesen, and John Vande Vate; from other institutions: Lou Billera (Cornell University), Ed Frazelle, Betsy Greenberg (University of Texas), Susan Griffin, Walter Hlawitschka (Fairfield University), Phil Jones (University of Iowa), Mike Magazine (University of Cincinnati), Kathy Platzman, Eli Prisman (York University), Dan Steeples, and Ajay Subramanian (Georgia State University).

Steven T. Hackman
August, 2007

Contents

Part II Efficiency Measurement

Part IV Engineering Models of Technology

Part V Mathematical Appendix

List of Figures

List of Tables

Part I

Microeconomic Foundations

1

Overview

1.1 Introduction

For the purposes herein, a *technology* refers to the process by which a production system transforms its inputs "x" into its outputs "y". In this book, we develop a variety of *mathematical* models of technology that can be used for measuring efficiency, productivity, and performance and for optimizing resources. A mathematical model of technology is a *set*

$$\mathcal{T} := \{(x, y)\}$$

of input-output pairs. Informally, each (x, y) in $\mathcal{T}$ represents an input x from which it is possible to produce output y.

There are two components of a mathematical model of technology. The first component specifies the acceptable *domain* of x and y. For example, are the different inputs and outputs represented as numbers? If so, are these numbers restricted in any way (e.g., non-negative, bounded)? Are the different inputs and outputs represented as functions of time? If so, are they real-valued? Vector-valued? Non-negative? Continuous? Differentiable? Continuously differentiable? The domain is chosen to suit the needs of the analysis. For example, conventional efficiency, productivity, and performance measurement represent inputs and outputs as non-negative vectors, whereas engineering models of technology used for optimization represent inputs and outputs as real-valued functions of time. The choice of domain largely dictates the mathematical and computational tools used to undertake the analysis.

The second component of a mathematical model is a precise statement of the *rules* or *axioms* $\mathcal{A}$ by which membership in $\mathcal{T}$ is specified. The axioms $\mathcal{A}$ are applied to a data set

$$\mathcal{D} := \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$$

collected from N representative production systems of different firms (or different units of the same firm) to generate an *approximation* $\mathcal{T}^a$ to the true technology set $\mathcal{T}$, symbolically represented as

$$\mathcal{D} \xrightarrow{\mathcal{A}} \mathcal{T}^a.$$

How "close" $\mathcal{T}^a$ is to $\mathcal{T}$ depends on the quantity and quality of the data set $\mathcal{D}$ and the choice of axioms $\mathcal{A}$. (We do not discuss in this book how data are defined and collected, *per se*, but the process of data definition and collection can be time-consuming and iterative.)

An example of a basic, obvious axiom is the rule that all observed input-output pairs belong to $\mathcal{T}$. Given this axiom, the data set $\mathcal{D}$ is a subset of $\mathcal{T}^a$. It is possible to stop here; if so, then $\mathcal{T}^a = \mathcal{D}$. Arguably, this is the most conservative approximation to the true technology set. It is certainly simple enough to define, but it has serious deficiencies. For example, most firms will typically be rated efficient if this approximation is used. This is because efficiency measurement compares (in some way) a firm's observed input-output pair to those in $\mathcal{T}^a$ to obtain its efficiency rating, and $\mathcal{T}^a$ here is quite "small." Managers of the firms will be pleased at first, but once they realize that almost all firms are rated efficient, they will desire a more discriminating efficiency measurement system. This can be achieved by expanding the list of axioms in $\mathcal{A}$. When this expanded list of axioms is then applied to the data set $\mathcal{D}$, it will generate a larger technology set $\mathcal{T}^a$. Here are some examples.

- *Disposability axioms.* For most technologies, it is reasonable to assume that it is possible to dispose of unneeded input or output. There are various forms. The most common one, *strong* disposability, requires that an input-output pair (x, y) belongs to $\mathcal{T}$ if there is some observed input-output pair (x_i, y_i) such that x is larger than x_i and y is smaller than y_i.
- *Returns-to-scale axioms.* For some technologies, it is reasonable to assume that when input is doubled, tripled, etc., output will be doubled, tripled, etc., too.
- *Convexity axioms.* It is often assumed that weighted averages (i.e., convex combinations) of observed input-output pairs in the technology also belong to the technology set. A "time-divisibility" argument is sometimes used to justify this axiom; however, this axiom is generally assumed for its analytical properties.

As we shall see, an impressively large technology set $\mathcal{T}^a$ can be generated by applying just a few, relatively innocuous axioms to a data set $\mathcal{D}$.

1.2 Microeconomic Foundations

Microeconomic foundations of production economics used in this book are the subject of Part I.

Neoclassical model of technology

Traditional microeconomic analysis represents a technology set via a *steady-state production function* $\Phi(\cdot)$, the subject of Chapter 2. A production function

$\Phi(x_1, ..., x_n)$ represents the *maximum* scalar output $y \geq 0$ that can be achieved using input vector $x = (x_1, ..., x_n) \geq 0$. Typical input categories include capital (K), labor (L), materials (M), and energy (E). Output y depends on the industry and represents an aggregation of the many outputs a firm produces. Standard examples include: tons of steel, kilowatt-hours of electricity, bushels of corn, barrels of oil, number of printed circuit assemblies, etc. Given that a production function represents the maximum output, if one assumes the axiom that output can be freely disposed, then the technology set associated with $\Phi(\cdot)$ is

$$\mathcal{T} := \{(x, y) : \Phi(x) \geq y\}. \tag{1.1}$$

Since $\mathcal{T}$ is completely characterized by a production function, all other axioms are formulated in terms of $\Phi(\cdot)$. For example, it is typically assumed that $\Phi(\cdot)$ is a non-decreasing function, which implies that the technology exhibits strong disposability of input. Convexity axioms require that $\Phi(\cdot)$ is either concave or quasiconcave.

Sensitivity analysis of an industry production function enables a production economist to answer questions such as:

- What are substitution possibilities among the input factors? For example, if the output level is to remain unchanged, by what percent must the capital input be increased to compensate for a 5% decrease in the labor input?
- How is output affected by a change in an input factor? For example, what will be the percentage increase in output if the labor input is increased by 5%?
- How is output affected by input scale? For example, what will be the percentage increase in output if all inputs are increased by 5%?
- How will an industry adjust its consumption of two particular input factors if their relative prices change and output is to remain the same? For example, suppose the net effect of a proposed government policy will raise the labor wage and/or lower the cost of capital investment. In response, firms can be expected to adjust their labor-capital ratio downwards (i.e., substitute capital input for labor input), the degree of which will affect the employment in this economic sector.

A production function also enables a production economist to compute measures of *technical efficiency*. Suppose a firm uses x to produce y. If $y = \Phi(x)$, then the firm is producing at the maximum level possible given its level of resources. However, if $y < \Phi(x)$, then the firm is *output inefficient*. A natural measure of output efficiency is

$$\theta^*_{output} := \frac{y}{\Phi(x)} \,. \tag{1.2}$$

Since θ^* is a *scalar* (number) between 0 and 1, one may interpret $100\theta^*_{output}$ as a grade. For example, a grade of 50% means the firm should be able to

double its output, whereas a grade of 80% means the firm should be able to increase its output by 25% (for the given input vector).

Even if the firm is output efficient, i.e., $y = \Phi(x)$, it could nonetheless be *input inefficient*, that is, it could be that the firm could reduce its inputs and still achieve the same level of output. Typically, input inefficiency goes hand-in-hand with output inefficiency. For an observed input-output pair (x, y), one possible measure of input efficiency is

$$\theta^*_{input} := \min\{\theta : \Phi(\theta x) \geq y\}. \tag{1.3}$$

Since x was observed to achieve y, which by definition must be less than or equal to $\Phi(x)$, the value of θ^*_{input} must be less than or equal to one. This input efficiency measure is by no means the only way to measure input efficiency, though it is the most popular. Consider this measure of input efficiency:

$$\gamma^*_{input} := \min\left\{\frac{\sum_{i=1}^{n}\gamma_i}{n} : \Phi(\gamma_1 x_1, ..., \gamma_n x_n) \geq y\right\}. \tag{1.4}$$

Note that θ^*_{input} requires an *equiproportional* reduction in all input factors whereas γ^*_{input} does not. Clearly, $\gamma^*_{input} \leq \theta^*_{input}$, and so θ^*_{input} is the more "conservative" of the two. (Which measure would you prefer if you were on the receiving end of an evaluation?)

An axiomatic model of technology

Output of a neoclassical production function represents an aggregation or index of a firm's, industry's, or nation's output (over some period of time). Firms produce a variety of output, and it may be difficult to find an appropriate aggregation of output to serve as an index. For this reason, the neoclassical model of technology must be extended to accommodate vector-valued output.

We adopt an axiomatic approach in the spirit of the pioneering efforts of G. Debreu and R.W. Shephard. That is, we formally define a collection of *possible* axioms that a *well-behaved* technology set could plausibly satisfy. These include different types of the aforementioned disposability, returns-to-scale, and convexity axioms, as well as the topological properties of closure and compactness. As previously mentioned, the modeler chooses which axioms to include in the set of axioms $\mathcal{A}$. When $\mathcal{A}$ is applied to a data set $\mathcal{D}$, it will *by definition* determine a well-behaved approximation $\mathcal{T}^a$ to $\mathcal{T}$. This formal approach to describing technology and how to approximate a technology set via extrapolation of a given data set are the subject of Chapter 3.

Nonparametric models of technology

Application of these axioms generates what are termed *nonparametric* models of technology. In the past two decades, nonparametric models of technology

have been extensively developed and applied. The main motivation for the development of these models, apart from their application to vector-valued output, is that parametric functional forms used to estimate the production function for the neoclassical model of technology often exhibit properties that can be refuted by the data.

The most popular nonparametric models of technology include the fixed-coefficients model due to W.W. Leontief, the constant and variable returns-to-scale *Data Envelopment Analysis (DEA)* models pioneered by A. Charnes, W.W. Cooper and their colleagues, and a model due to G. Hanoch and M. Rothschild, all developed in Chapter 4. A two-input, single-output model is often useful as a prototype to communicate and educate managers and engineers about productivity and efficiency analysis. For this special case, it is possible to construct each of these nonparametric models of technology directly from the data.

A neoclassical production function associated with a single-output nonparametric model is *not* differentiable. (Each isoquant is piecewise linear.) In lieu of calculus, *linear programming* is the main computational tool used to perform sensitivity, efficiency, and productivity analysis.

Dual characterizations of technology

Under appropriate conditions, it is possible to reconstruct the technology set by observing the economic behavior of producers. Economists refer to such reconstructions as *dual characterizations* of technology.

The cost function (Chapter 5), indirect production function (Chapter 6), and distance function (Chapter 7) provide three powerful dual characterizations. Each has enormous value in its own right. A cost function represents the *minimum* cost required to achieve a pre-specified output level given input prices. It can be used to assess effects of a change in output or input prices on cost. An indirect production function represents the *maximum* amount of output a producer can achieve given a budget constraint. In the single output case, it is possible to *graphically* compute the cost and indirect production function.

The distance function underpins much of nonparametric efficiency analysis, and is the basis for a recent nonparametric approach to measuring productivity. Essentially, for a given input-output pair (x, y), an input distance function measures the extent to which the input vector x must be scaled so that it most efficiently obtains the output vector y. (An analogous concept exists for an output distance function.) Distance and cost functions are linked via two, symmetric identities; knowledge of either function uniquely determines the other, and hence the technology set itself.

Nonconvex models of technology

The properties of convex sets (e.g. separation theorems) are central to economic analysis, and are used repeatedly to establish key results in the tradi-

tional microeconomic theory of production. While extremely useful, the property of convexity (e.g., convex sets and quasiconcave functions) is accepted mainly for its analytical expedience.

Nonconvex sets appear when analyzing even simple extensions to popular models. We describe three such examples that arise in resource allocation, producer budgeting, and Data Envelopment Analysis. A generalization of convexity, called *projective-convexity*, can be used to represent these models (and more). This is the subject of Chapter 8.

Projective-convexity possesses a general type of separation property, which can be used to establish a dual characterization of technology. For example, the *multi-dimensional* indirect production function represents the maximum output a producer can achieve given *separate* budget constraints for each input category. Production and multi-dimensional indirect production functions are linked via two, symmetric identities; knowledge of either function uniquely determines the other, and hence the technology set itself.

1.3 Efficiency Measurement

With microeconomic foundations firmly in place, we are in position to formally define different types of efficiency and to show how to compute them, the subject of Part II.

Efficiency analysis

At its core, a measure of efficiency compares an observed input-output pair (x, y) to its projection $(\hat{x}, \hat{y})$ onto the boundary of the technology set $\mathcal{T}$. As previously discussed, there are several ways to approximate a technology $\mathcal{T}$ from a data set $\mathcal{D}$. There are also several ways to project a point onto a boundary of a set. Here are some examples:

- The *radial* measure of output efficiency is

$$\theta^*_{output} := \min\{\theta : \ (x, y/\theta) \in \mathcal{T}\}. \tag{1.5}$$

  Here, (x, y) is projected to $(x, y/\theta^*_{output})$.
- The *radial* measure of input efficiency is

$$\theta^*_{input} := \min\{\theta : \ (\theta x, y) \in \mathcal{T}\}. \tag{1.6}$$

  Here, (x, y) is projected to $(\theta^*_{input} x, y)$.
- The *Russell* measure of input efficiency is

$$\gamma^*_{input} := \min\left\{\frac{\sum_{i=1}^n \gamma_i}{n} : \ ((\gamma_1 x_1, \gamma_2 x_2, \ldots, \gamma_n x_n), \ y) \in \mathcal{T}\right\}. \tag{1.7}$$

  Here, (x, y) is projected to $((\gamma^*_1 x_1, \gamma^*_2 x_2, \ldots, \gamma^*_n x_n), y)$.

- The *hyperbolic* measure of *joint* input-output efficiency is

$$h^* := \min\{h : \ (hx, y/h) \in \mathcal{T}\}. \tag{1.8}$$

  Here, (x, y) is projected to $(h^*x, y/h^*)$.

When $\mathcal{T}$ is characterized by a production function $\Phi(\cdot)$ as in (1.1), then the measures of efficiency (1.5)-(1.7) defined above are identical to their respective counterparts (1.2)-(1.4).

In addition to these measures of efficiency, there are measures of scale and cost efficiency. Each of these efficiency measures, with the exception of the hyperbolic measure, can be computed via linear programming. Formal definitions and computation models of different types of efficiency are the subject of Chapter 9.

Two-dimensional projection

As previously noted, graphical representation is a useful tool for educating managers and engineers about efficiency measurement. It is possible to *graphically* compute a host of efficiency measures in the multi-input, multi-output setting from a *single* graph in the plane. This is the subject of Chapter 10. This graph can be viewed as a two-dimensional projection of an embedded single-input, single-output well-behaved technology. It is possible to compute this projection using a simplex-like pivoting algorithm, which we describe in detail. For the two-input, single-output special case, we will describe a simple procedure for generating this projection directly from the data.

Multi-stage efficiency analysis

The models of technology described so far implicitly assume a single production *stage*, namely, inputs are transformed into final outputs used to satisfy "customer demand." Within the plant or factory, a production system often consists of a *network* of subsystems or stages. Intermediate output of one stage is used as input to follow-on stages. It represents work-in-process. For such a system, conventional efficiency measurement determines *separate* efficiency measures for each stage and the system as a whole. That is, each stage of each firm is compared to its peers using data relevant to that stage, and the data describing the aggregate system are used to compare each firm to its peers. Unfortunately, this approach can lead to two very undesirable results. First, it is possible for each stage to be rated 100% efficient while the system as a whole to be rated inefficient. Second, it is possible for the system as a whole to be rated near efficient while each subsystem is rated highly inefficient.

We will describe an efficiency measurement framework for a production system consisting of subsystems or stages in series that will *never* produce the two undesirable results mentioned above, because it *simultaneously* computes efficiencies of each subsystem and the aggregate system. This is the

subject of Chapter 11. Intermediate outputs, or work-in-process, are explicitly represented. A Pareto efficient frontier characterizes the acceptable set of subsystem efficiencies. Each point in this frontier determines a derived measure of aggregate efficiency. A "consistent pricing principle" characterizes the proposed models.

Application of efficiency analysis to the warehouse and distribution industry

Mathematical models for efficiency measurement have real application. Over a four-year period, through the auspices of the Material Handling Research Center at the Georgia Institute of Technology, data were collected from a large sample of warehouses drawn from a wide array of industries and analyzed. Several models of technology were developed and analyzed (see Chapter 12). The main conclusions are that smaller and less automated warehouses tend to perform more efficiently than larger and more automated warehouses. These conclusions have been used for design and productivity improvement, which is of prime importance to equipment vendors, system designers, consultants, facility engineers, and warehouse managers.

1.4 Productivity and Performance Measurement

Efficiency measurement provides a *static* assessment of each firm. Productivity and performance measurement assesses how well a firm improves its efficiency and technology from one period to the next, the subject of Part III.

Let (x^0, y^0) and (x^1, y^1) denote, respectively, a firm's observed input-output pairs in two consecutive periods (e.g., last year, this year). By how much did the firm improve its productive capability? At its core, a productivity assessment uses a ratio O/I, where O represents a measure of the growth in aggregate output produced, and I represents a measure of the growth in aggregate input consumed. For example, if $O = 1.15$, reflecting a 15% increase in aggregate output, and $I = 1.06$, reflecting a 6% increase in aggregate input, then $O/I \sim 1.09$, suggesting a 9% gain in productivity.

Index numbers

A productivity assessment requires a method for aggregating output and input that yield output and input indexes, the ratio of which determines the productivity index. Consequently, productivity measurement is rooted in the construction of *index numbers*, the subject of Chapter 13. We will define and analyze the most well-known and intuitively appealing price and quantity indexes due to Konus, Laspeyres, Paasche, and Tornqvist. Each price (quantity) index can be used to immediately generate an *implicit* quantity (price) index.

Productivity analysis

We discuss two approaches to productivity measurement (see Chapter 14). The first approach, commonly referred to as the *growth accounting approach* or *total factor productivity*, is widely used and was developed in the late 1950's by R. Solow. It uses a simple formula to measure productivity that is based on a firm's own inputs and output. The second approach was developed in the mid-1990's by R. Fare, S. Grosskopf, and their colleagues. This approach maps data sets $\mathcal{D}^0$ and $\mathcal{D}^1$ collected in consecutive periods to nonparametrically generate approximations to $\mathcal{T}^0$ and $\mathcal{T}^1$, respectively. It uses distance functions described in Part I to disentangle the "efficiency effect," namely, how well a firm improved its own efficiency, from the "technical change effect," namely, how well the technology $\mathcal{T}^1$ improved in comparison to $\mathcal{T}^0$. This approach yields the same assessment of productivity as the growth accounting approach if the technology follows the assumptions of the Solow model.

Performance analysis

At first blush, a firm's observed increase in profits from one year to the next may lead management to conclude the firm has improved its productivity. This conclusion can be erroneous for two fundamental reasons. It ignores the "price effect:" an increase in profits may have more to do with increasing output prices and decreasing input prices than it does with any productivity gains. Second, it ignores the "substitution effect:" when input and output prices change, the choices of inputs and outputs typically change, too. A performance assessment seeks to cleanse these effects from the raw numbers to obtain a fair assessment of productivity growth due solely to the firm's improved capability to transform input into output. Performance measurement is the subject of Chapter 15.

We show how to use index numbers to decompose a firm's performance into its "profitability," "productivity," and "price recovery" components. We also show how to distribute a firm's net gain year-to-year into these categories. The approach we describe adapts the performance measurement system due to J.W. Kendrick.

Economic analysis of productivity

Firms are constantly looking for ways to innovate, either in the product market or in the means of production. Firm productivity can translate into an increase in market share, output, employment, and profits. The extent to which this happens will depend on the degree of competition, namely, the number of competitors and the productivity of each competitor. It is entirely possible for a firm to increase its productivity, but still see a marked decline in its competitive position because the competition improved their respective productivities even more.

To illustrate how this economic analysis is undertaken, we examine the effects of productivity by explicitly determining how each firm's price, output, labor employed, revenue, profit, and market share are affected by the market in which firms compete. This is the subject of Chapter 16. In economics parlance, we analyze a specific market with consumers and producers and derive the market's general equilibrium. We analyze what economists call the *competitive*, *monopolistic competitive*, and *oligopoly* market structures. We will use the analysis to quantitatively assess how a "productivity laggard" or a "productivity leader" will fare against its competition.

1.5 Engineering Models of Technology

With today's information systems, shop-floor data are becoming increasingly available. This data are used by production engineers to build a detailed model of technology for the purpose of optimizing resources. In principle, a production economist can also use detailed models to undertake efficiency, productivity, and performance assessment. Developing models that can serve the interests of both the production engineer and economist is the subject of Part IV.

A production economist typically works with aggregate data that record the cumulative amounts of inputs and outputs in some predetermined period of time (e.g., quarterly, yearly). The time window used by a production economist is too long to be of use to the production engineer. To capture the physical phenomena necessary for optimizing resources, a production engineer needs to know the actual *timing* of the inputs and outputs within a production economist's time window. Consequently, inputs and outputs must incorporate a *time dimension*, namely,

$$x = (x_1(\cdot), x_2(\cdot), \ldots, x_n(\cdot)), \quad y = (y_1(\cdot), y_2(\cdot), \ldots, y_m(\cdot)).$$

We will consider two types of functions to describe the flow of inputs and outputs over time. An *event-based flow* $z(\cdot)$ associates a nonnegative real number $z(\tau)$ to an event that occurs at time τ. For example, $z(\tau)$ might be the quantity of parts initiated into a production process at time τ or a value associated with a completed job at time τ. Event-based flows only take on positive values at those times when events occur. For a *rate-based flow* $z(\cdot)$, the nonnegative real number $z(\tau)$ represents the rate (quantity per unit time) of flow at time τ. Rate-based flows sometimes represent a fluid approximation to event-based flows, and also arise quite naturally when modeling physical processes.

At a detailed level, a production process typically begins with raw materials, parts, subassemblies, and transforms them via several intermediate stages to produce final outputs sold to end users. The core building block of an engineering model of technology is a production stage or *activity*. The first step in building an engineering model of technology, therefore, is to characterize each activity's input-output process.

Dynamic production functions

At a detailed level, each activity's input-output process is conveniently encapsulated by a *dynamic production function*

$$x = (x_1(\cdot), x_2(\cdot), \ldots, x_n(\cdot)) \quad \xrightarrow{f} \quad y = (y_1(\cdot), y_2(\cdot), \ldots, y_m(\cdot)).$$

A dynamic production function $f(\cdot)$ is a *functional*, since both its domain and range are vectors of functions, not vectors of numbers. Since a constant function of time can be identified with a scalar, a dynamic production function is not only compatible with but also extends the neoclassical model of (single-stage) technology.

When an activity's input-output process is not instantaneous, a *lead time* exists between an input start and realized output. For example, it may take time for paint to dry, to complete an inspection, or to transport product. When lead times are sufficiently small, they can be ignored. However, when lead times are significant, they must be accounted for to achieve an accurate model of technology.

The simplest models assume lead times are pre-specified constants, independent of the endogenous behavior of the system. The class of *index-based* dynamic production functions can be used to represent constant lead time processes. This is the subject of Chapter 17. All input and output functions of an index-based process can be represented in terms of a single (index) function. There are three practical ways of defining the index to model constant lead times: indexing "starts," "outs," or non-storable services.

Beyond simple, constant lead time models, a wide variety of non-instantaneous activity processes can be characterized using *distribution-based* dynamic production functions. We will describe the following subclasses:

- constant lead time processes with noninteger lead times and unequal length time periods;
- time-dependent, lead time processes; and
- continuous lead time processes.

The first subclass extends the familiar textbook models of dynamic, single-stage, deterministic production systems. The second subclass uses an exogenously specified "time of completion" function. Here, the output realization depends on the time when input enters the system. The third subclass can be used to model a distribution of output due to inherent randomness or system load, for example, the distribution of parts that fail a time-intensive inspection. Distribution-based dynamic production functions are the subject of Chapter 18.

Often the true dynamic production function is difficult to represent or is not tractable for analysis. It is possible to generate *linear* approximations to the ideal dynamic production function that yield representations amenable for computation. These include:

- *Load-dependent lead time processes.* Here, the lead time is a function of the size of the input queue or "system load"—the larger the queue, the longer the lead time. Such processes arise in manufacturing systems. Given a pre-determined (single) input curve and the queue discipline, the (single) output curve is obtained by integrating a differential equation that characterizes this process. The linear approximation is accurate as long as the input curve is close to the chosen input curve used to generate the approximation. It is possible to extend this formulation to allow the coefficients that define the linear approximation to *depend* on the size of the queue. (This extended formulation uses binary variables, and so this approximation is no longer linear.)
- *Two-point boundary approximation.* Here, just two "boundary" input-output points are used to define the dynamic production function on the whole input domain. We will describe an application of this approximation to modeling work flow in a project-oriented production system such as a naval shipyard.
- *Aggregation.* Aggregate dynamic production functions arise from serial or parallel aggregation of individual activities represented by detailed dynamic production functions.
- *Dynamic activity analysis.* This model extends the steady-state activity analysis models described in Chapter 4 to the dynamic setting.

These approximations are the subject of Chapter 19.

Stochastic model of technology

In a stochastic input-output model, input arrives according to some stochastic process and the resulting output can also incorporate random phenomena. We will describe a simple input-output stochastic model in which the output curve is interpreted as the *expected* output. Both single input and batch input models are developed. We will show how to construct a confidence interval, namely, upper and lower output functions that bound the expected output curve. It is possible to approximate a confidence interval using a linear approximation suitable for computation. This is the subject of Chapter 20.

Multi-stage, dynamic models of technology

Using the single-stage dynamic models as building blocks, we provide a framework to model *multi-stage* systems over time. We describe in detail two examples from manufacturing and assembly with rework. We also show how to use multi-stage, dynamic models of technology for efficiency and productivity measurement. This is the subject of Chapter 21. The multi-stage models we develop are most useful for short-term production planning.

The basic, continuous-time model of dynamic production involves a network of interrelated activities, each of whose technologies is characterized by a

dynamic production function. Storable goods used by each activity will either be acquired from outside the system (i.e., exogenously supplied) or obtained via intermediate product transfers from other activities within the system. Material balance constraints are required to ensure that the requisite storable inputs are available at the time they are used in the production process. Nonstorable services mainly involve the use of different types of equipment and labor, as well as services obtained from activities within the system. Service capacity equations are required to ensure that the rates of the aggregate services available are sufficient to meet internal, aggregate demand.

Specific models of technology are developed by substituting examples of the index-based and distribution-based dynamic production functions described in Chapters 17 and 18 into the fundamental equations. These fundamental equations characterize the technology set $\mathcal{T}$. We will show how to translate these continuous-time models into suitable discrete-time, linear approximations amenable for computation.

When non-instantaneous behavior exists, the initial system state contains relevant information about the work-in-process, namely, the status of the intermediate products in the pipeline. To be accurate, a detailed model of technology must project the future state of this work-in-process. We will describe two practical approaches to handle these initial conditions.

We will also show how to extend the basic model in several practical ways. For example, within a stage there are often several alternative processes that can be used to produce the same output. As another example, there can be load-dependent lead times. Here, the queue in front of a stage reflects all product queues. (This extension uses binary variables, but no more than the ones used to approximate single-input, single-output load-dependent lead time processes.)

The ideas presented in Part IV are applied to model the flow of work inside a warehouse for the purpose of optimizing labor resources. This is the subject of Chapter 22. Specifically, an optimization model is developed that determines the various times personnel (pickers and packers) report to work throughout the day, and how to strategically use overtime and part-time staff. By better matching workers to the *timing* of work requirements, significant reductions in both the number of workers and overtime will be achieved. As a by-product, the model suggests *order release guidelines* that will improve labor efficiency and ease demands for space by reducing unnecessary work-in-process.

1.6 Mathematical Appendix

Our inquiry necessitates the use of mathematics, some basic and some not so basic. The reader should consult the first chapter to become acquainted with the basic notation and mathematical constructs used throughout this book.

Real analysis

Convergence, compactness, continuity, and L^p-spaces constitute the core real analysis used in this book. We review the basic definitions and properties.

Convex analysis

Separation theory of convex sets in Euclidean space (finite-dimensional spaces) is essential to modern economic theory, and is used throughout Part I. The core theorems are presented and proved. Key definitions and properties of the class of concave (convex) and quasiconcave (quasiconvex) functions are provided, too.

Optimization

Economists typically assume economic agents (e.g., consumers, producers) make rational decisions about what to consume, how much to save, invest, produce, etc. Such problems are formulated as optimization problems, which involve an objective function that measures the value to the economic agent of making a particular choice, and a collection of constraints that define feasible choices.

We define several classes of optimization problems, provide necessary and sufficient conditions for optimality, and show how to solve convex optimization problems via Lagrangian duality and the dual formulation. We illustrate these concepts with applications to an economic lot size problem and to linear programming.

Sensitivity analysis

Sensitivity analysis plays a major role in microeconomics. We formally state and prove the extremely useful Envelope Theorem. We use it to revisit sensitivity analysis of the cost function.

A technology can be described via *correspondences*, essentially a point-to-*set* mapping. We provide the core theory of correspondences. This is used to prove Berge's Theorem of the Maximum, which is then used to show the continuity of the cost and indirect production functions.

1.7 A Word of Advice

As noted in the preface, I have tried to make the material accessible. Accessibility ultimately depends on a reader's background, motivation, and patience. In several places, the material requires a careful and thoughtful read. This should not come as a surprise, since many of the ideas presented in this book

were either developed by or built on research of the pioneers in production economics (S.N. Afriat, A. Charnes, W.W. Cooper, G. Debreu, W.E. Diewert, J. Kendrick, T.C. Koopmans, W.W. Leontief, and R.W. Shephard) or by current leaders of the field (R. Fare, S. Grosskopf, R.C. Leachman and C.A.K. Lovell). I hope you enjoy the journey.

2

Production Functions

A production function $\Phi(x_1, ..., x_n)$ represents the *maximum* output u that can be achieved using input vector $x = (x_1, ..., x_n)$. We shall use the symbol u instead of y to emphasize that output is a scalar. A production function can be used by a production economist to undertake different types of sensitivity analysis and to compute different measures of technical efficiency. We begin our discussion of production functions by describing popular parametric forms used for its estimation. Next, we formally describe different ways to estimate rates of substitution among input factors and to measure scale properties. We close this chapter with a description of homothetic production functions, a subclass used often in the theory of productivity analysis.

2.1 Parametric Forms

In standard steady-state productivity analysis, data (x_i, u_i), $i = 1, 2, ..., N$, are collected for N firms within an industry to estimate the parameters of an assumed functional form for $\Phi(\cdot)$.

Definition 2.1. *An* **isoquant** *of the production function $\Phi(\cdot)$ is the collection of input vectors*

$$ISOQ_{\Phi}(u) := \{x : \Phi(x) = u\}$$

that achieve the same level of output.

Definition 2.2. *An* **input possibility set** *or* **upper level set** *of the production function $\Phi(\cdot)$ is the collection of input vectors*

$$L_{\Phi}(u) := \{x : \Phi(x) \geq u\}$$

that achieve at least output level u.

Popular functional forms include the following examples.

Example 2.3. The Cobb-Douglas function.

$$\Phi(x) = A\prod_{i=1}^{n} x_i^{\alpha_i},\ \alpha_i > 0,\ i = 1, 2, \ldots n\ . \tag{2.1}$$

In two dimensions we shall work with $\Phi(K, L) = K^a L^b$, where K and L denote, respectively, the capital and labor inputs. An example is depicted in Figure 2.1.

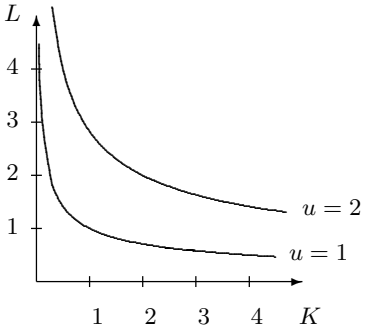


Fig. 2.1. Isoquants for the Cobb-Douglas production function $\Phi(K, L) = K^{1/3}L^{2/3}$.

Example 2.4. The Constant Elasticity-of-Substitution (CES) function.

$$\Phi(x) = A(\sum_{i=1}^{n} \alpha_i x_i^{\rho})^{1/\rho},\ \rho \neq 0,\ \rho \leq 1. \tag{2.2}$$

In two dimensions we shall work with $\Phi(K, L) = (aK^{\rho} + bL^{\rho})^{1/\rho}$. An example is depicted in Figure 2.2.

Example 2.5. Translog function.

$$\ln \Phi(x) = \beta_0 + \sum_i \beta_i \ln x_i + \frac{1}{2}\sum_i \sum_j \beta_{ij} \ln x_i \ln x_j\ , \tag{2.3}$$

where it is further assumed that $\beta_{ij} = \beta_{ji}$. This function generalizes the Cobb-Douglas function.

Example 2.6. Generalized quadratic.

$$\Phi(x) = \beta_0 + \sum_i \beta_i\ g_i(x_i) + \frac{1}{2}\sum_i \sum_j \beta_{ij}\ g_i(x_i) g_j(x_j)\ , \tag{2.4}$$

where it is further assumed that $\beta_{ij} = \beta_{ji}$ and that each $g_i(\cdot)$ is a given twice continuously differentiable function. This function generalizes the translog function.

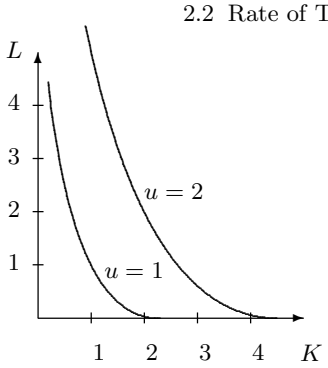


Fig. 2.2. Isoquants for the CES production function $\Phi(K, L) = (2/3\sqrt{K}+1/3\sqrt{L})^2$.

All of the above functional forms are twice continuously differentiable, which facilitates analysis via calculus. For the remainder of this chapter, we assume the production function has the requisite derivatives.

2.2 Rate of Technical Substitution

The *Rate of Technical Substitution (RTS)* measures the degree of substitution between two input factors. We motivate its formal definition below by first examining the two-dimensional setting.

Fix K and L, let $u = \Phi(K, L)$, and let ΔK and ΔL represent small perturbations of K and L, respectively. We wish to find a relationship between ΔK and ΔL to ensure the new input vector $(K + \Delta K, L + \Delta L)$ achieves the same output level u. Since $\Phi(\cdot)$ is differentiable,

$$\Phi(K + \Delta K, L + \Delta L) \approx \Phi(K, L) + \Delta K \frac{\partial \Phi}{\partial K} + \Delta L \frac{\partial \Phi}{\partial L} \tag{2.5}$$

with the error being sufficiently small if both ΔK and ΔL are sufficiently small, too—see (A.6), p. 446. Since the output level is to remain unchanged, it follows that

$$\Delta K \frac{\partial \Phi}{\partial K} + \Delta L \frac{\partial \Phi}{\partial L} \approx 0, \tag{2.6}$$

or

$$\Delta L \approx -\Big[\frac{\partial \Phi / \partial K}{\partial \Phi / \partial L}\Big] \cdot \Delta K. \tag{2.7}$$

The expression in brackets is the rate of technical substitution between capital and labor.

When a differentiable function $L(K)$ exists for which

$$\Phi(K, L(K)) = u \text{ for all } K, \tag{2.8}$$

then differentiating both sides of (2.8) with respect to K yields

$$\frac{\partial \Phi}{\partial K} + \frac{\partial \Phi}{\partial L} L'(K) = 0,$$

which implies

$$L'(K) = -\frac{\partial \Phi / \partial K}{\partial \Phi / \partial L} .$$

This is a formal way of deriving the change of L with respect to the change in K given in (2.7).

Example 2.7. Consider the Cobb-Douglas production function $\Phi(K, L) = K^{1/3} L^{2/3}$ depicted in Figure 2.1, and suppose initially that $(K, L) = (1, 1)$ and $\Phi(1, 1) = 1$. Since $\partial \Phi / \partial K = 1/3(L/K)^{2/3}$ and the $\partial \Phi / \partial L = 2/3(K/L)^{1/3}$, the rate of technical substitution is $-0.5L/K$. Evaluated at the point $(K, L) = (1, 1)$, this ratio equals $-1/2$. Thus, if K increases to 1.01, then L should decrease to approximately 0.995. (The exact value of L is 0.99503719.) If K increases to 1.05, then L should decrease to approximately 0.975. (The exact value of L is 0.975900073 and the percentage error is less than 0.1%.) Keep in mind that the rate of technical substitution is accurate in a small neighborhood of the original point.

As shown in Figure 2.3, the rate of technical substitution is the slope of the line tangent to the graph of $L(\cdot)$. For this example, $L(K) = K^{-1/2}$, $L'(K) = -0.5K^{-3/2}$, and the tangent line $\mathcal{L}$ when $K = 1$ is $L = -0.5K + 1.5$.

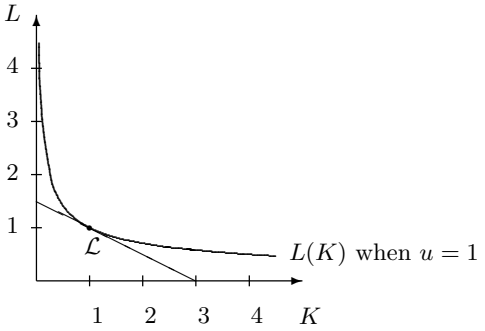


Fig. 2.3. The slope of the line $\mathcal{L}$ tangent to the isoquant equals the rate of technical substitution $L'(K)$.

Remark 2.8. One expects the slope $L'(K)$ to be negative for a typical production technology. As K increases, the *Law of Diminishing Returns* suggests that each additional increment ΔK to K should result in less of a labor reduction ΔL, which implies the (negative) slope $L'(K)$ increases with K. Consequently, $L''(K)$ is positive, and so it is reasonable to assume that $L(\cdot)$ is a decreasing convex function (see Remark D.14, p. 478). For the Cobb-Douglas production function in two dimensions, $L(K)$ is proportional to $K^{-a/b}$, and so $L'(K)$ is

proportional to $-(a/b)K^{-(a/b+1)}$, which is always negative, and $L''(K)$ is proportional to $(a/b)(a/b+1)K^{-(a/b+2)}$, which is always positive. Consequently, $L(\cdot)$ is always a decreasing convex function. For the CES production function in two dimensions, $L(K) = (u^\rho - aK^\rho/b)^{1/\rho}$. When both K and L are positive, it can be readily verified that $L'(K)$ is always negative and $L''(K)$ will be non-negative only when $\rho \leq 1$. Consequently, in order for $L(\cdot)$ to be convex it is necessary to restrict $\rho \leq 1$.

We shall write $\Phi_i(x)$ to denote the $\partial\Phi/\partial x_i$ evaluated at x, which gives the rate at which output changes per additional unit of factor i consumed. It is referred to as the **marginal product of factor i.**

Definition 2.9. *The* **rate of technical substitution between input i and input** j *at a given input vector $x \in \mathbb{R}^n_+$ is minus the ratio of the marginal products, denoted by*

$$RTS_{ij}(x) := -\frac{\Phi_i(x)}{\Phi_j(x)} .$$

Example 2.10. The rate of technical substitution for a generalized Cobb-Douglas production function is

$$-\frac{\alpha_i}{\alpha_j}\frac{x_j}{x_i} . \tag{2.9}$$

Example 2.11. The rate of technical substitution for a generalized CES production function is

$$-\frac{\alpha_i}{\alpha_j}\left(\frac{x_j}{x_i}\right)^{1-\rho} . \tag{2.10}$$

Remark 2.12. The rates of technical substitution for the generalized CES and Cobb-Douglas production functions coincide when $\rho = 0$. While the CES production function is not defined for $\rho = 0$, this observation suggests that the CES production function "converges" to a Cobb-Douglas production function as ρ tends to zero. Indeed, this convergence holds when $\sum_i \alpha_i = 1$. (As we shall learn below, the CES production function always exhibits constant returns-to-scale, whereas the Cobb-Douglas production function exhibits constant returns-to-scale only when $\sum_i \alpha_i = 1$, and so this condition is necessary.)

To see this, fix an x whose coordinates are all positive, and define

$$f_x(\rho) := \left[\sum_i \alpha_i x_i^\rho\right]^{1/\rho} .$$

Since the numerator and denominator of

$$\ln f_x(\rho) = \frac{\ln \sum_i \alpha_i x_i^\rho}{\rho}$$

tend to zero as ρ tends to zero, L'Hopital's rule (see A.12, p. 447) applies. Consequently,

$$\lim_{\rho\to 0} \ln f_x(\rho) = \frac{\lim_{\rho\to 0} \frac{d}{d\rho}[\ln \sum_i \alpha_i x_i^\rho]}{\lim_{\rho\to 0} \frac{d}{d\rho}[\rho]} = \lim_{\rho\to 0} \frac{\sum_i \alpha_i x_i^\rho \ln x_i}{\sum_i \alpha_i x_i^\rho} = \sum_i \alpha_i \ln x_i.$$

Since the exponential function is continuous,

$$\lim_{\rho\to 0} f_x(\rho) = \prod_i x_i^{\alpha_i}.$$

Thus, the CES production function converges (pointwise) to the Cobb-Douglas production function as the parameter ρ approaches zero.

2.3 Elasticity

Consider a production function $\Phi(\cdot)$ of one input factor. Suppose output increases by 100 as a result of an increase of 10 in input. Is this a good return on investment? In absolute terms there is an increase of 10 units of output for each 1 unit increase in input. The return appears excellent. Suppose, however, we learn the initial input level was 10 and the initial output level was 10,000. Measured in *percentage* terms, input increased by 100% while the output increased by only 1%. Using a linear approximation, we estimate an 0.1% increase in output for each 1% increase in input. From this perspective, most likely the return would be considered too low. If, on the other hand, the initial input level was 100 and the initial output level was 100, then the input increased by 10% while the output increased by 100%, yielding a 10% increase in output for each 1% increase in input. In the second case, the return does appear to be excellent. This simple example illustrates that to obtain an accurate assessment of the benefit/cost for the response to a change in a variable, it is often necessary to measure the changes in *percentage* terms instead of absolute terms. The following definition conveys this notion.

Definition 2.13. *Let $f(\cdot)$ be a differentiable real-valued function of one variable. The* **elasticity of $f(\cdot)$ evaluated at** *z is*

$$\epsilon_f(z) := \lim_{\Delta z\to 0} \left[\frac{f(z+\Delta z) - f(z)}{f(z)} \div \frac{(z+\Delta z) - z}{z} \right] = \frac{z f'(z)}{f(z)}.$$

(The definition only applies when z and $f(z)$ are both not zero.)

In Definition 2.13, the numerator measures the percentage change in $f(\cdot)$ and the denominator measures the percentage change in z. Thus, the percentage change in $f(\cdot)$ is estimated to be $\epsilon_f(z)$ when $\Delta z = 0.01z$. That is, to a first-order approximation, $\epsilon_f(z)$ measures the percentage change in $f(\cdot)$ in response to a 1% change in z.

Example 2.14. The elasticity of $f(z) = z^\gamma$, $\gamma \neq 0$, is γ regardless of the value of z. Suppose z changes by 100δ% to $(1+\delta)z$. Now

$$f((1+\delta)z) = (1+\delta)^\gamma z^\gamma \approx (1+\delta\gamma)z^\gamma$$

when δ is sufficiently small since $f(1+\delta) \approx f(1) + f'(1)\delta$. Consequently,

$$\delta\epsilon_f(z) \approx \frac{(1+\delta\gamma)z^\gamma - z^\gamma}{z^\gamma} = \delta\gamma.$$

Remark 2.15. Note that

$$\epsilon_f(z) = \frac{\frac{d}{dz}\ln f(z)}{\frac{d}{dz}\ln z}.$$

Often the "dz" are "canceled" and the elasticity of $f(\cdot)$ evaluated at z is expressed as $d\ln f(z)/d\ln z$. If one interprets $d\ln f(z)$ as $\ln f(z+\Delta z) - \ln f(z)$, then as a first-order (Taylor series) approximation,

$$d\ln f(z) \approx \frac{f'(z)\Delta z}{f(z)} \approx \frac{f(z+\Delta z) - f(z)}{f(z)},$$

which is indeed the percentage change in $f(\cdot)$.

Measuring a variety of elasticities is central to applied production analysis, and we now examine a few of the important ones.

2.4 Elasticity of Output, Scale and Returns to Scale

Elasticity of output measures the percentage change in output in response to a 1% change in factor input i when all other input factors are held constant. For each scalar z, let

$$\theta_x^i(z) := \Phi(x_1, x_2, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n)$$

denote the derived production function of input i when all other input factors are held constant at x.

Definition 2.16. *The* **elasticity of output with respect to input i evaluated at x** *is*

$$\epsilon_i(x) := \epsilon_{\theta_x^i}(x_i) = \frac{x_i\Phi_i(x)}{\Phi(x)}.$$

Example 2.17. Example 2.14 shows the elasticity of output with respect to input i for the Cobb-Douglas production function is simply α_i.

Example 2.18. The elasticity of output with respect to input i for the CES production function is $\alpha_i x_i^\rho / \sum_i \alpha_i x_i^\rho$. The elasticity of output for the CES and Cobb-Douglas production functions coincide when $\rho = 0$ and $\sum_i \alpha_i = 1$, as they should.

Elasticity of scale measures the percentage change in output in response to a 1% change in *all* input factors. (When all input factors change proportionally the *scale* of operations is said to change.) For each scalar s, let

$$\theta(s) := \Phi(sx)$$

denote the derived production function of the *scale* of operations, as represented by s, when the *input mix*, as represented by x, remains constant.

Definition 2.19. *The* **elasticity of scale evaluated at x** *is*

$$\epsilon(x) := \epsilon_\theta(1) = \frac{d}{ds}\, ln\, \theta(s) \Big|_{s=1}.$$

Perhaps not too surprisingly, elasticity of scale is intimately related to output elasticities.

Proposition 2.20. *The elasticity of scale equals the sum of the output elasticities.*

Proof. Using the chain-rule,

$$\epsilon_\theta(1) = \frac{\sum_i \Phi_i(x) x_i}{\Phi(x)} = \sum_i \epsilon_i(x).$$

Definition 2.21. *If $\epsilon(x) = 1$, then the production function exhibits* **constant returns-to-scale** *at x. If $\epsilon(x) > 1$, then the production function exhibits* **increasing returns-to-scale** *at x. If $\epsilon(x) < 1$, then the production function exhibits* **decreasing returns-to-scale** *at x.*

Example 2.22. Example 2.17 show the elasticity of scale for the Cobb-Douglas production function is $\sum_i \alpha_i$. Thus, the returns-to-scale characterization for a Cobb-Douglas technology is determined by whether the sum of the exponents is greater, equal or less than 1.

Example 2.23. Example 2.18 shows the elasticity of scale for the CES production function is always 1, and hence it always exhibits constant returns-to-scale.

2.5 Elasticity of Substitution

The *elasticity of substitution* provides a unit-free measure of the substitutability between two inputs. Under suitable conditions, it measures the percentage change in the factor ratio x_i/x_j for a 1% change in the factor price ratio p_i/p_j. We shall develop this concept for a two-factor production function, with the extension to general dimensions being straightforward.

Consider the CES production function

$$\Phi(K, L) = [2/3\sqrt{K} + 1/3\sqrt{L}]^2 \tag{2.11}$$

depicted in Figure 2.2. Suppose the prices on capital and labor are 4 and 1, respectively, and the desired output level is 1. Since the output requirement is fixed, a profit-maximizing producer should select the input vector x^* to minimize cost. When we analyze the cost function in Chapter 5, we show that the ratio of the marginal products of capital and labor, $\Phi_K(x^*)/\Phi_L(x^*)$, must equal the ratio of the factor prices, p_K/p_L; otherwise, it will be possible to lower cost. Since the ratio of the marginal products is minus the Rate of Technical Substitution (see Definition 2.9), $-RTS(x^*)$, and the RTS for the CES production function given in (2.10) equals $-(a/b)(L/K)^{1-\rho} = -2\sqrt{L/K}$, the cost-minimum labor-capital ratio equals 4. Substituting $L = 4K$ into (2.11) with $\Phi(K, L) = 1$, the cost minimizing capital input $K^* = 9/16$ and hence the cost minimizing labor input $L^* = 4K^* = 9/4$.

Now suppose as a result of a proposed government tax or stimulus policy, the labor wage is expected to increase and/or the cost of capital investment is expected to decrease thereby lowering the p_K/p_L ratio by 10% to 3.6. The producer is expected to adjust the input mix to accommodate the price changes; in particular, the producer will substitute capital for labor, resulting in lower employment in this economic sector. It is desirable to estimate the change, and the degree of substitution is fundamentally affected by the structure of the technology. Here, the new RTS equals 3.6, and so the new labor-capital ratio will equal $1.8^2 = 3.24$. Thus, in response to a 10% decrease in relative prices the labor-capital ratio declined by 19%. Substituting $L = 3.24K$ into (2.11) the new cost-minimizing input vector is $(K^*, L^*) = (0.6233, 2.0194)$.

Let m denote the labor-capital ratio. Since the RTS for the CES production function is

$$f(m) := m^{1-\rho},$$

whose elasticity is $1-\rho$, the percentage change in the RTS for a 1% change in the labor-capital ratio is approximately 0.5%. Going in the opposite direction, a 1% change in the RTS should result in approximately a 2% change in the labor-capital ratio. In our example, the RTS declined by 10%, and so we would estimate the corresponding decline in the labor-capital ratio to be about 20%. In fact, it was 19%, which is quite close.

Fix the output level u and for each K, let $L(K)$ denote the (assumed) unique value of L for which $\Phi(K, L) = u$. Let

$$m(K) := \frac{L(K)}{K}$$

denote the factor input ratio and let

$$r(K) := -\frac{\Phi_K(K, L(K))}{\Phi_L(K, L(K))}$$

denote the RTS.

Definition 2.24. *For each $x = (K, L)$ the* **elasticity of substitution** $\sigma(x)$ *measures the percentage change in the labor-capital ratio in response to a 1% change in the RTS. Formally,*

$$\sigma(x) := \lim_{\Delta K \to 0} \frac{\frac{m(K+\Delta K)-m(K)}{m(K)}}{\frac{r(K+\Delta K)-r(K)}{r(K)}} .$$

By dividing numerator and denominator by ΔK,

$$\sigma(x) = \frac{\frac{d}{dK} \ln m(K)}{\frac{d}{dK} \ln r(K)} = \frac{d \ln m(K)}{d \ln r(K)} , \tag{2.12}$$

where the interpretation of the second identity is discussed in Remark 2.15.

When the elasticity of substitution is high, the isoquant (in a neighborhood of the point in question) has little curvature. Thus, for a small change in the relative prices of the two inputs, firms will significantly change their ratio mix. Substitutability of input is relatively easy. Conversely, when the elasticity is low, the isoquant has much curvature. Thus, for a small change in the relative prices of the two inputs, firms will not alter their ratio mix by much. Substitutability of input is more difficult.

Remark 2.25. As we have seen with the CES or Cobb-Douglas production functions, sometimes the RTS can be expressed as a function of the ratio L/K, namely, $RTS(x) = f(L/K)$ for some function $f(\cdot)$. In this case, the elasticity of substitution (2.12)

$$\sigma(x) = \frac{\frac{m'(K)}{m(K)}}{\frac{r'(K)}{r(K)}} = \frac{\frac{m'(K)}{m(K)}}{\frac{f'(m(K))m'(K)}{f(m(K))}} = \epsilon_f(m)^{-1}$$

is the reciprocal of the elasticity of $f(\cdot)$ with respect to m. For the CES production function, $f(m) = m^{1-\rho}$ and so $\sigma(x) = 1/(1-\rho)$.

When $\rho = 1$ the CES production function is linear, which means that if both K and L are positive, then the factor price ratio must equal $-a/b$, the constant slope of $L(K)$. If the factor price ratio were to change, the new cost minimal input vector would instantly jump to a boundary point (either K or L would be zero). The elasticity of substitution can be interpreted as infinite and indeed $\sigma(x) = +\infty$.

2.6 Homothetic Production Functions

Homothetic production functions permit a more general returns-to-scale characterization than a constant returns-to-scale technology. This special class of functions plays an important role in the theory of index numbers (see Chapter 13).

Definition 2.26. *A function $F : \mathbb{R}_+ \to \mathbb{R}_+$ is said to be a* **transform** *if (i) it is increasing and continuous, (ii) $F(0) = 0$ and (iii) $F(v) \to \infty$ as $v \to \infty$.*

Remark 2.27. Since $F(\cdot)$ is increasing it has an inverse, which we shall denote by $f(\cdot)$. The inverse function $f(\cdot)$ is also a transform.

Definition 2.28. *A production function $\Phi(\cdot)$ is* **homothetic** *if it can be represented as $F(\phi(\cdot))$, where $\phi(\cdot)$ is a constant returns-to-scale (linearly homogeneous) production function, and $F(\cdot)$ is a transform.*

The function $\phi(\cdot)$ can be interpreted as an index number.

Example 2.29. A constant returns-to-scale production function $\phi(\cdot)$ is trivially homothetic; simply define $F(v) = v$. (Note that $f(v) = v$, too.)

Example 2.30. The general Cobb-Douglas production function $\Phi(x) = A \prod_i x_i^{\alpha_i}$ is homothetic since $\Phi(x) = F(\phi(x))$, where $\phi(x) := A \prod_i x_i^{\alpha_i/S}$, $F(v) := v^S$ and $S := \sum_i \alpha_i$.

Given the unit isoquant of $\phi(\cdot)$,

$$ISOQ_\Phi(1) = \{x : \ \phi(x) = 1\},$$

the returns-to-scale characterization is completely provided by the transform $F(\cdot)$. Each isoquant,

$$ISOQ_\Phi(u) = \{x : \ \Phi(x) = u\},$$

and hence each input possibility set, can be generated from the unit isoquant via radial scaling—as in the special case of constant returns-to-scale technology—except that here, the appropriate scale factor is $f(u)/f(1)$. (Note this ratio equals u in the constant returns-to-scale special case.) This geometrical fact follows from the definitions:

$$\begin{aligned} F(\phi(x)) = 1 &\iff \phi(x) = f(1) \\ &\iff \phi\Big(\frac{f(u)}{f(1)}\, x\Big) = f(u) \\ &\iff F(\phi(\lambda_u x)) = u, \end{aligned}$$

where $\lambda_u := f(u)/f(1)$. Thus, if x lies on the unit isoquant, then $\lambda_u x$ lies on the isoquant corresponding to output u. Consequently, the ratio $f(v)/f(u)$ is the scale factor that will map $ISOQ_\Phi(u)$ to the $ISOQ_\Phi(v)$.

Example 2.31. Figure 2.4 depicts the input possibility sets associated with a homothetic production function when $F(v) = v^2$. Notice how the shape of the isoquants are identical to the unit isoquant, just as in the CRS case; it is only the output label affixed to each input possibility set that is being changed via the transform $F(\cdot)$. To achieve twice as much output one need only scale inputs by a factor of $\sqrt{2}$. The production function exhibits increasing returns-to-scale with this choice of transform $F(\cdot)$.

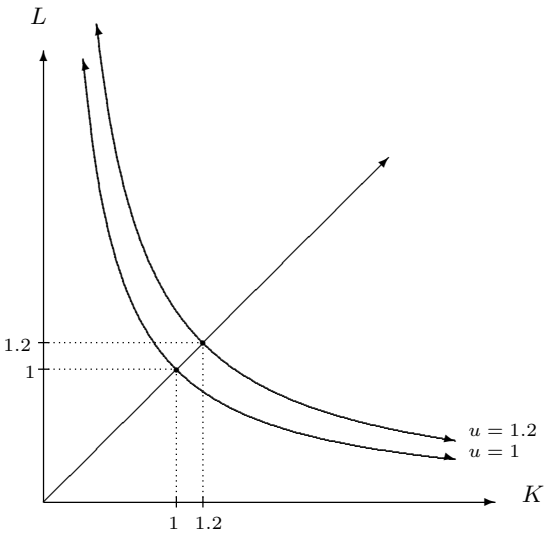


Fig. 2.4. Input possibility sets for a homothetic production function.

As a final remark, we note that the marginal rates of technical substitution among inputs (for a fixed input vector) for homothetic technologies are independent of scale, i.e., they do not change as the input vector is scaled. This is because the marginal rates of technical substitution (in two dimensions) define the slope of the isoquant, and the shapes of the isoquant are identical.

2.7 Exercises

2.1. The *generalized power function* takes the form

$$\Phi(x) = A \prod_{i=1}^{n} x_i^{f_i(x)} e^{g(x)}.$$

Verify that this function generalizes the Cobb-Douglas function.

2.2. A function $\Phi(\cdot)$ is quasiconcave if each of its upper level sets $L_\Phi(u)$ is convex (see Appendix D). Let

$$\phi_\Phi(z_i, z_j) := \Phi(x_1, x_2, \ldots, x_{i-1}, z_i, x_{i+1}, \ldots x_{j-1}, z_j, x_{j+1}, \ldots, x_n) \qquad (2.13)$$

denote the production function of input factors i and j derived from $\Phi(\cdot)$ when all other factor inputs are kept constant at their original levels given by x. Show the geometry depicted in Figure 2.3 holds true for the two-dimensional production function $\phi_\Phi(z_i, z_j)$.

2.3. When the RTS is not a function of the factor ratio, it is possible to use the partial derivatives to determine the elasticity of substitution. Let $\Phi_{ij} := \partial^2\Phi/\partial x_i \partial x_j$ denote the second partial derivatives. Recall that $\Phi_{ij} = \Phi_{ji}$. Show that

$$\sigma(x) = \frac{-\Phi_K \Phi_L (K\Phi_K + L\Phi_L)}{KL(\Phi_{KK}\Phi_L^2 - 2\Phi_{KL}\Phi_K\Phi_L + \Phi_{LL}\Phi_K^2)} \ .$$

2.4. Show that the elasticity of scale is 2 for the production function described in Example 2.31.

2.5. Let $f_k(\cdot)$, $k = 1, 2, \ldots, N$, denote N "micro" production functions, and let $\epsilon_k(x)$ denote the elasticity of scale of f_k at x. Suppose the elasticities of scale of each f_k are identical, i.e., $\epsilon_k(x) := \epsilon(x)$ for each k. Let w_k, $k = 1, 2, \ldots, N$ be positive numbers.

(a) Show that the elasticity of scale of the aggregate production function given by $\Phi(x) = \sum_k w_k f_k(x)$ evaluated at x equals $\epsilon(x)$.
(b) Show that the elasticity of scale of the aggregate production function given by $\Phi(x) = \prod_k f_k(x)^{w_k}$ evaluated at x equals $(\sum_k w_k)\epsilon(x)$.

2.6. Consider a constant returns-to-scale, single-output technology of four input factors, capital, labor, materials and energy. At the current input levels it is known the elasticities of output with respect to labor, materials, and energy respectively equal 0.10, 0.30, and 0.20. What is the elasticity of output with respect to capital?

2.7. Prove that the inverse of a transform is also a transform.

2.8. Show that if $\Phi(\cdot)$ is differentiable and homothetic, then the marginal rates of technical substitution are constant along rays through the origin.

2.8 Bibliographical Notes

This material can be found in graduate-level microeconomic textbooks such as Varian [1992] and Jehle and Reny [2001]. Chambers [1988] provides a thorough treatment of production functions.

2.9 Solutions to Exercises

2.1 Set $f_i(x) = \alpha_i$ for each i and $g(x) \equiv 0$.

2.2 First argue that when $\Phi(\cdot)$ is nondecreasing, differentiable and quasiconcave, so is $\phi_\Phi(\cdot,\cdot)$. Next, use the fact that the gradient vector $\nabla\Phi(x)$ of a differentiable quasiconcave function induces a supporting hyperplane to the upper level set (see Theorem D.12, p. 476) to show that in the two-dimensional (z_i, z_j)-plane, the line that passes through the point $(z_i, z_j) = (x_i, x_j)$ and whose slope is $RTS_{ij}(x)$ *supports* the input possibility set $L_{\phi_\Phi}(\phi_\Phi(z))$ at z.

2.3 First verify the following identities (use the chain rule for the second):

$$\frac{m'(K)}{m(K)} = \frac{L'(K)K - L}{KL}.$$
$$\frac{r'(K)}{r(K)} = \frac{(\Phi_{KK} + \Phi_{KL}L'(K))\Phi_L - \Phi_K(\Phi_{LK} + \Phi_{LL}L'(K))}{\Phi_K\Phi_L}.$$

Next, substitute these identities into the definition for $\sigma(x) = \frac{m'(K)/m(K)}{r'(K)/r(K)}$ and use the fact that $L'(K) = -\Phi_K/\Phi_L$.

2.4 Fix input vector x. Here, $\Phi(sx) = F(\phi(sx)) = F(s\phi(x)) = s^2\phi^2(x)$. Thus, the elasticity of scale equals

$$\epsilon(x) = \frac{d}{ds}\ln\theta(s)\,\Big|_{s=1} = \frac{d}{ds}\{\ln s^2 + \ln\phi^2(x)\}\,\Big|_{s=1} = \frac{2}{s}\,\Big|_{s=1} = 2.$$

In words, when you double the input here the output scales by a factor of four.

2.5 For both parts let $\theta(sx) = \Phi(sx)$. As for part (a),

$$\frac{\theta'(s)}{\theta(s)} = \frac{\sum_k w_k f_k'(sx)}{\sum_k w_k f_k(x)} = \sum_k \Big[\frac{w_k f_k(x)}{\sum_k w_k f_k(x)}\Big]\frac{f_k'(sx)}{f_k(sx)}.$$

Now use the fact that $f_k'(sx)/f_k(sx)$ evaluated at $s = 1$ equals $\epsilon_k(x) = \epsilon(x)$ for each k. As for part (b),

$$\frac{d}{ds}\ln\theta(s) = \frac{d}{ds}\Big[\sum_k w_k \ln f_k(sx)\Big] = \sum_k w_k \frac{d}{ds}\ln f_k(sx).$$

Now use the fact that $d/ds \ln f_k(sx)$ evaluated at $s = 1$ equals $\epsilon_k(x) = \epsilon(x)$ for each k.

2.6 The elasticity of scale for a constant returns-to-scale technology is one. It also equals the sum of the output elasticities (Proposition 2.20). Thus, $1 = 0.10 + 0.30 + 0.20 + \epsilon_K$, which implies the elasticity of output with respect to capital is 0.40.

2.7 Let $F(\cdot)$ be a transform. Since $F(\cdot)$ is increasing its inverse $f^{-1}(\cdot)$ is well-defined and increasing, too. Moreover, it is easily follows from the properties of $F(\cdot)$ that $f(0) = 0$ and that $f(v) \to \infty$ as $v \to \infty$. It remains to show that $f(\cdot)$ is continuous. Fix y_0 and let $x_0 = f(y_0)$. Pick $\epsilon > 0$. We must show there exists a $\delta > 0$ such that $|f(y) - f(y_0)| < \epsilon$ whenever $|y - y_0| < \delta$. Set

$$\delta = \min \left\{ \frac{F(x_0 + \epsilon) - y_0}{2}, \ \frac{y_0 - F(x_0 - \epsilon)}{2} \right\}.$$

The value of δ is positive since $F(\cdot)$ is increasing so that $F(x_0 - \epsilon) < F(x_0) < F(x_0 + \epsilon)$. If $|y - y_0| < \delta$, then by construction

$$F(x_0 - \epsilon) < y_0 - \delta < y < y_0 + \delta < F(x_0 + \epsilon).$$

In particular, $F(x_0 - \epsilon) < y < F(x_0 + \epsilon)$. Since $f(\cdot)$ is increasing, it then follows that $x_0 - \epsilon < f(y) < x_0 + \epsilon$ or that $|f(y) - x_0| = |f(y) - f(y_0)| < \epsilon$, as required.

2.8 We may express $\Phi(x) = F(\phi(x))$ where $\phi(\cdot)$ is linearly homogeneous. Fix x and let $s > 0$. We have that

$$\frac{\partial \Phi(sx)/\partial x_i}{\partial \Phi(sx)/\partial x_j} = \frac{F'(s\phi(x))s\partial\phi(x)/\partial x_i}{F'(s\phi(x))s\partial\phi(x)/\partial x_j} = \frac{\partial\phi(x)/\partial x_i}{\partial\phi(x)/\partial x_j},$$

which is independent of s.

3

Formal Description of Technology

Traditional microeconomic theory of production uses a production function. Output is an aggregation or index of an industry's or nation's output (over some period of time). Firms, however, produce different outputs, and sometimes it is difficult to find an appropriate aggregation of output to serve as an index, especially in the services industries. For this and other reasons, a more general theory of technology is required. There are two components to a theory of production. The *primitive elements* that describe a technology must be defined, and the properties or *axioms* the primitive elements are required to satisfy must be delimited. In this way, one knows precisely what is and what is not an acceptable model of technology. In building a theory one wishes to impose the fewest axioms, and ideally all axioms should be testable, if one wishes to consider it a scientific theory. Often, some axioms are made for analytical convenience and are not testable, *per se*. If so, these axioms should be justified in some way. Finally, the theory should prove fruitful. The theory we develop in this chapter will provide the basis for the nonparametric models of technology developed in Chapter 4 and the efficiency and productivity analysis undertaken in Parts II and III.

3.1 Primitive Elements

In lieu of a production function, a core primitive element of the traditional microeconomic theory of production, a technology will be described via a collection of input-output pairs $\{(x, y)\}$. As in Chapter 2, an input x is an n-dimensional nonnegative vector. Output y is now permitted to be an m-dimensional vector, which we assume is also nonnegative. **We will continue to use the symbol 'u' to denote the output when it is single-dimensional.** When we use the symbol y, keep in mind we do not necessarily mean that $m > 1$, only that m is permitted to be larger than one.

A technology can be defined in one of two equivalent ways. First, we describe a technology via a *single* primitive element $\mathcal{T} \subset \mathbb{R}^n_+ \times \mathbb{R}^m_+$, a subset of

the *joint input-output space*, which abstractly represents those input-output pairs (x, y) that are technologically achievable.

Definition 3.1. *A* **technology** *is a subset* $\mathcal{T}$ *of a* **joint input-output space** $I\!R^n_+ \times I\!R^m_+$. *It will also be referred to as a* **technology set**.

Embedded in a technology set are the input and output possibility sets.

Definition 3.2. *For each* $y \in I\!R^m_+$ *the* **input possibility set** *derived from the technology set* $\mathcal{T}$ *is*

$$L(y) := \{x : \ (x, y) \in \mathcal{T}\}.$$

For each $x \in I\!R^n_+$ *the* **output possibility set** *derived from the technology set* $\mathcal{T}$ *is*

$$P(x) := \{y : \ (x, y) \in \mathcal{T}\}.$$

We suppress the functional dependence of the input and output possibility sets on $\mathcal{T}$.

Remark 3.3. The collection of input possibility sets defines the *input possibility correspondence* (or point to set mapping) $L : \mathbb{R}^m_+ \longrightarrow 2^{\mathbb{R}^n_+}$. Similarly, the collection of output possibility sets defines the *output possibility correspondence* $P : \mathbb{R}^n_+ \longrightarrow 2^{\mathbb{R}^m_+}$. See Appendix G for details about correspondences.

A second approach to defining a technology uses the input or output possibility sets as the primitives.

Definition 3.4. *A* **technology** *is a* **family of input possibility sets**

$$\mathcal{F} = \{L(y) : \ y \in I\!R^m_+\}$$

or a **family of output possibility sets**

$$\mathcal{P} = \{P(x) : \ x \in I\!R^n_+\}.$$

In applications, one first defines the input and output categories and how they will be measured—a challenging and nontrivial task to be sure—and then collects an *observed* input-output data set

$$\mathcal{D} := \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\} \tag{3.1}$$

from representative firms in the industry under consideration. In the single-output setting, parametric analysis can be undertaken to define a production function. Given a production function, such as the ones discussed in Chapter 2, then one obvious way to define a technology set $\mathcal{T}$ is

$$\mathcal{T} := \{(x, y) : \ \Phi(x) \geq u\}. \tag{3.2}$$

The input possibility sets corresponding to $\mathcal{T}$ are

$$L(u) := \{x : \Phi(x) \geq u\},$$

and the output possibility sets corresponding to $\mathcal{T}$ take the special form

$$P(x) := [0, \Phi(x)]$$

of a closed interval. For the nonparametric models of technology developed in Chapter 4, the "smallest" technology set consistent with the chosen set of axioms is represented by a finite set of linear inequalities.

3.2 Input and Output Disposability

We first define concepts pertaining to input and output disposability. We begin with one of the strongest forms of disposability.

Definition 3.5. *A technology set $\mathcal{T}$ exhibits*

- **input free disposability** *if*

$$(x, y) \in \mathcal{T} \text{ and } x' \geq x \Longrightarrow (x', y) \in \mathcal{T}.$$

  In this case the input possibility sets are said to be **nested**, *since $L(y') \subset L(y)$ whenever $y' \geq y$.*
- **output free disposability** *if*

$$(x, y) \in \mathcal{T} \text{ and } y' \leq y \Longrightarrow (x, y') \in \mathcal{T}.$$

  In this case the output possibility sets are said to be **nested**, *since $P(x) \subset P(x')$ whenever $x' \geq x$.*
- **free disposability** *if it exhibits free disposability of both input and output.*

The following weaker notion of disposability is sometimes used.

Definition 3.6. *A technology set $\mathcal{T}$ exhibits*

- **weak input disposability** *if*

$$(x, y) \in \mathcal{T} \Longrightarrow (\lambda x, y) \in \mathcal{T} \text{ for all } \lambda \geq 1.$$

  In this case the input possibility sets are said to be **weakly nested**, *since $L(\mu y) \subset L(y)$ whenever $\mu \geq 1$.*
- **weak output disposability** *if*

$$(x, y) \in \mathcal{T} \Longrightarrow (x, \mu y) \in \mathcal{T} \text{ for all } \mu \leq 1;$$

  In this case the output possibility sets are said to be **weakly nested**, *since $P(x) \subset P(\lambda x)$ whenever $\lambda \leq 1$.*

- **weak disposability** *if it exhibits weak disposability of both input and output.*

Remark 3.7. The preceding notions of output disposability presume that all outputs are "positive." For a "negative" output such as pollution, less of this output is preferred. In this case, the inequality sign associated with this output would simply be reversed in the definitions of disposability. Since this generalization is easily accommodated in the theory and computations to follow, and since it will be notationally convenient to ignore it, we shall do so. In many cases, an alternative approach to representing such output is to take the reciprocal quantity.

3.3 Efficient Frontiers

Concepts of efficiency and its measurement fundamentally relate to the concepts of an Efficient Frontier. Points in an Efficient Frontier represent a type of "maximal trade-off."

Definition 3.8. *The* **Efficient Frontier of an input possibility set** $L(y)$ *is*

$$\mathcal{EF}(y) := \{x \in L(y) : \text{ if } x' \lneqq x, \text{ then } x' \notin L(y)\}.$$

The **Efficient Frontier of an output possibility set** $P(x)$ *is*

$$\mathcal{EF}(x) := \{y \in P(x) : \text{ if } y' \gneqq y, \text{ then } y' \notin P(x)\}.$$

The **Efficient Frontier** *of a technology set* $\mathcal{T}$ *is*

$$\mathcal{EF} := \{(x, y) \in \mathcal{T} : x \in \mathcal{EF}(y) \text{ and } y \in \mathcal{EF}(x)\}.$$

3.4 Axioms for a Well-Behaved Technology

We are now in position to define a well-behaved technology.

Definition 3.9. *A technology set* $\mathcal{T}$ *is* **well-behaved** *if it satisfies the following axioms:*

A1) *Each input vector can achieve zero output, and a nonzero output vector must require at least some nonzero input.*
A2) *Each output possibility set is compact.*
A3) *It exhibits free disposability.*
A4) *Each input possibility set is closed and convex.*
A5) *The Efficient Frontier of each input possibility set is bounded.*

In mathematical notation, Axiom **A1** states that $(x, 0) \in \mathcal{T}$ for all $x \in \mathbb{R}^n_+$ and $(0, y) \notin \mathcal{T}$ for all $y \in \mathbb{R}^m_+$, $y \neq 0$. This is surely a minimal requirement of a well-behaved technology.

As for Axiom **A2**, a set $S \subset \mathbb{R}^k$ is compact if and only S is closed and bounded. $S \subset \mathbb{R}^k$ is bounded if there exists a uniform bound on the size of any vector in the subset. (Size is measured using a suitable norm, e.g., the Euclidean norm.) A single input vector can be used to obtain possibly different output vectors; however, a single input vector should not be able to produce an arbitrarily large amount of output. Hence the boundedness postulate. $S \subset \mathbb{R}^k$ is closed if whenever $\{x_n\} \subset S$ such that $x_n \longrightarrow x$, then $x \in S$. (Here, convergence is with respect to a suitable norm, e.g., the Euclidean norm.) Closure is a reasonable, albeit technical, property imposed on the technology.

On the input side, axiom **A3** is an innocuous condition. One may view an input vector as an allocation of resources. On the output side, this condition is a little more tenuous, especially for certain types of physical processes.

As for Axiom **A4**, the closure property, while reasonable, is assumed for analytical expedience. The same can be said for the property of convexity. It does have the following "time-divisibility" justification. Suppose input vectors x_1 and x_2 each achieve output level $u > 0$. Pick a $\lambda \in [0, 1]$, and imagine operating $100\lambda\%$ of the time using x_1 and $100(1 - \lambda)\%$ of the time using x_2. At an aggregate level of detail, it is not unreasonable to assume that the weighted average input vector $\lambda x_1 + (1 - \lambda)x_2$ can also achieve output level u. This convexity property will be used repeatedly. Keep in mind that an input possibility set can be empty.

Axiom **A5** guarantees the cost function is properly defined for all nonnegative prices. Consider the Cobb-Douglas production function of capital and labor inputs. If, for some reason, the price of labor were *zero*, then the cost of achieving a given positive output level u could be made arbitrarily close to zero but would never actually be equal to zero. This is because for any arbitrarily small capital input there is a sufficiently large labor input to achieve output level u. The substitution possibilities along the isoquant are *infinite*. In this example, the points along the isoquant define the Efficient Frontier of the input possibility set. Boundedness of the Efficient Frontier is one way to rule out this infinitely substitutable behavior.

Remark 3.10. In general, nonparametric models used to define a technology will *never* generate a technology set that is closed. This is why this stronger axiom is not required.

3.5 Single-Output Technologies

As a test of the usefulness of the theory of technology developed so far, it should be possible to "extract" a traditional production function from a technology set associated with a single-output technology. Indeed, this is so.

Definition 3.11. *Let $\mathcal{T} \subset \mathbb{R}^n_+ \times \mathbb{R}_+$ be a well-behaved, single-output technology set. The* **production function $\Phi^{\mathcal{T}}(\cdot)$ derived from $\mathcal{T}$** *is*

$$\Phi^{\mathcal{T}}(x) := \max \{u : \ (x, u) \in \mathcal{T}\}.$$

It is understood that the technology from which a production function is derived *must* be a well-behaved, single-output technology. We shall refer to $\Phi^{\mathcal{T}}(\cdot)$ as a **derived production function**.

Remark 3.12. Given a technology $\mathcal{T}$, and hence the families of input and output possibility sets, $\mathcal{F}$ and $\mathcal{P}$, respectively, a derived production function can also be defined as

$$\Phi^{\mathcal{T}}(x) := \max \{u : \ u \in P(x)\} = \max\{u : \ x \in L(u)\}.$$

It is acceptable to replace the supremum with a maximum in the definition of the derived production function, since the continuous function $f(u) = u$ will achieve its maximum on the compact set $P(x)$.

The proof of the following proposition follows immediately from the definition of a well-behaved technology.

Proposition 3.13. *A derived production function satisfies the following properties.*

a) $\Phi(x) \geq 0$ *for every* $x \in \mathbb{R}^n_+$ *and* $\Phi(0) = 0$.
b) $\Phi(x)$ *is finite for every input vector* x.
c) $\Phi(\cdot)$ *is nondecreasing.*
d) $\Phi(\cdot)$ *is upper semicontinuous and quasiconcave.*

Suppose one starts with a technology set, extracts the derived production function, and then defines the technology generated from this production function, as in (3.2). It would be most undesirable if the technology set so generated was *not* identical to the original technology set. Indeed, it is identical.

Proposition 3.14. *Let $\mathcal{T}$ be a well-behaved technology set, let $\Phi^{\mathcal{T}}(\cdot)$ denote the derived production function, and let $T(\Phi^{\mathcal{T}})$ be the technology generated from $\Phi^{\mathcal{T}}(\cdot)$ given by*

$$T(\Phi^{\mathcal{T}}) := \{(x, u) : \ \Phi^{\mathcal{T}}(x) \geq u\}.$$

Then $T(\Phi^{\mathcal{T}}) = \mathcal{T}$.

Proof. Pick an $(x, u) \in \mathcal{T}$. Since $\Phi^{\mathcal{T}}(x) \geq u$ by definition of $\Phi^{\mathcal{T}}(\cdot)$, it immediately follows that $\mathcal{T} \subset T(\Phi^{\mathcal{T}})$. As for the reverse inclusion, $T(\Phi^{\mathcal{T}}) \subset \mathcal{T}$, pick an $(x, u) \in T(\Phi^{\mathcal{T}})$. Since $\max\{v : \ (x, v) \in \mathcal{T}\} \geq u$ and $\mathcal{T}$ exhibits output free disposability, it follows that $(x, u) \in \mathcal{T}$, as required. □

3.6 Extrapolation of Technology

Additional properties can be assumed, beyond the minimal ones, to define a well-behaved technology.

3.6.1 Convexity

There are several ways to impose convexity on a technology. We begin with the most obvious way, which will be central to the applications undertaken.

Definition 3.15. *A technology is* **convex** *if the technology set $\mathcal{T}$ is a convex subset of $\mathbb{R}^n_+ \times \mathbb{R}^m_+$.*

A convex technology has the property that

$$(x_1, y_1), (x_2, y_2) \in \mathcal{T},\ \lambda \in [0,1] \implies (\lambda x_1 + (1-\lambda)x_2,\ \lambda y_1 + (1-\lambda)y_2)) \in \mathcal{T}.$$

Operationally, it is easy to "implement" convexity. One simply takes the convex hull of the data set $\mathcal{D}$ in (3.1). Recall the convex hull of a set is the smallest convex set that contains the set. It is represented by a set of linear inequalities, which greatly facilitates computations.

The following proposition established a fundamental relationship between a convex single-output technology and the derived production function.

Proposition 3.16. *The derived production function of a convex technology is concave.*

Proof. Pick $(x_1, y_1), (x_2, y_2) \in \mathcal{T}$ and $\lambda \in [0,1]$. Since $\mathcal{T}$ is convex,

$$(\lambda x_1 + (1-\lambda)x_2,\ \lambda y_1 + (1-\lambda)y_2)) \in \mathcal{T},$$

which immediately implies from the definition of $\Phi^{\mathcal{T}}(\cdot)$ that

$$\Phi^{\mathcal{T}}(\lambda x_1 + (1-\lambda)x_2) \geq \lambda\Phi(x_1) + (1-\lambda)\Phi(x_2).$$

Thus, the derived production function is concave. □

The following definition describes a weaker application of convexity.

Definition 3.17. *A technology is* **bi-convex** *if each input and output possibility set is convex.*

Remark 3.18. In the language of sets, an input possibility set defines a **y-section** of the technology set, and the output possibility set defines an **x-section**. Subsets of an $X \times Y$ space are bi-convex if each section is convex.

There is a property called *projective-convexity* that lies "in between" convexity and bi-convexity. It assumes that if $(x_i, y_i) \in \mathcal{T}$, $i = 1, 2$, then for each convex combination $\lambda x_1 + (1-\lambda)x_2$, $\lambda \in [0,1]$ of x_1 and x_2 (or y_1 and y_2) there is *some* convex combination $\mu y_1 + (1-\mu)y_2$, $\mu \in [0,1]$, of y_1 and y_2 (or x_1 and x_2) for which $(\lambda x_1 + (1-\lambda)x_2, \mu y_1 + (1-\mu)y_2) \in \mathcal{T}$, too. It generalizes convexity, since for a convex set one may take $\mu = \lambda$. Every projectively-convex set is bi-convex; merely set $x_1 = x_2$ or $y_1 = y_2$. Properties of projective-convexity are explored in Chapter 8.

3.6.2 Disposability

Extrapolation of technology almost always assumes some form of disposability. Somewhat similar in spirit to extrapolation defined by the convex hull, the following set-theoretic operations define different types of "disposable hulls."

Definition 3.19. *The* **input free disposable hull** *of a set* $L \subset I\!R^n_+$ *is*

$$\mathcal{IFDH}(L) := \{x' \in I\!R^n_+ : \ x' \geq x \ \ \textit{for some} \ \ x \in L\}.$$

The **output free disposable hull** *of a set* $P \subset I\!R^m_+$ *is*

$$\mathcal{OFDH}(P) := \{y' \in I\!R^m_+ : \ y' \leq y \ \ \textit{for some} \ \ y \in P\}.$$

The **free disposable hull** *of a set* $\mathcal{T} \subset I\!R^n_+ \times I\!R^m_+$ *is*

$$\mathcal{FDH}(\mathcal{T}) := \{(x', y') \in I\!R^n_+ \times I\!R^m_+ : \ x' \geq x, \ y' \leq y \ \ \textit{for some} \ \ (x, y) \in \mathcal{T}\}.$$

Remark 3.20. The set-theoretic addition of two subsets S and T of a linear space is $S + T = \{s + t : \ s \in S, \ t \in T\}$. In set-theoretic notation

$$\begin{aligned} \mathcal{IFDH}(L) &= L + \mathbb{R}^n_+ \\ \mathcal{OFDH}(P) &= \left(P \ + \ \mathbb{R}^m_-\right) \cap \mathbb{R}^m_+ \\ \mathcal{FDH}(\mathcal{T}) &= \left(\mathcal{T} \ + \ \mathbb{R}^n_+ \times \mathbb{R}^m_-\right) \cap \mathbb{R}^n_+ \times \mathbb{R}^m_+. \end{aligned}$$

The following proposition is easily established from the definitions.

Proposition 3.21. *The disposable hull operations have these properties:*

a) If S is a set that (i) possesses a particular disposable property and (ii) contains a respective set (L, P or $\mathcal{T}$), then S contains the disposable hull of that set. Thus, each disposable hull is the smallest set that satisfies both (i) and (ii).[1]

b) If $S \subset T$, then each disposable hull of S is contained within the corresponding disposable hull of T.

c) The disposable hull of the union of two sets is the union of the disposable hulls of the two sets.

d) If a set S possesses a particular disposability property, then the disposable hull of S adds nothing to S; that is, S equals the disposable hull of itself.

Sequential applications of the set-theoretic operations of convexifying and freely disposing are interchangeable and thus preserve both properties.

Proposition 3.22. *Let $L \subset I\!R^n_+$, $P \subset I\!R^m_+$ and $\mathcal{T} \subset I\!R^n_+ \times I\!R^m_+$. Then:*

a) $\mathcal{FDH}(Conv(\mathcal{T})) = Conv(\mathcal{FDH}(\mathcal{T}))$.

[1] The respective disposable hull operations are closed under arbitrary intersections. Hence, each disposable hull is the smallest set with the particular disposable property that contains the given set.

b) $\mathcal{IFDH}(Conv(L)) = Conv(\mathcal{IFDH}(L))$.
c) $\mathcal{OFDH}(Conv(P)) = Conv(\mathcal{OFDH}(P))$.

Proof. We shall only prove (a), as the proofs of (b) and (c) are similar.

By definition, $(x', y') \in \mathcal{FDH}(Conv(\mathcal{T}))$ if and only if there exists $(x, y) \in Conv(\mathcal{T})$ such that $x' \geq x$ and $y' \leq y$. By definition of convex hull, $(x, y) \in Conv(\mathcal{T})$ if and only if there exists $(x_k, y_k) \in \mathcal{T}$ and $\lambda_k \in [0, 1]$, $k = 1, 2, \ldots, K$, such that $x = x(\lambda) := \sum_k \lambda_k x_k$, $y = y(\lambda) := \sum_k \lambda_k y_k$ and $\sum_k \lambda_k = 1$. Thus, $(x', y') \in \mathcal{FDH}(Conv(\mathcal{T}))$ if and only if there exists $\{(x_1, y_1), (x_2, y_2), \ldots, (x_K, y_k)\} \subset \mathcal{T}$ and $\lambda_k \geq 0$ for each $1 \leq k \leq K$ such that $x' \geq x(\lambda)$ and $y' \leq y(\lambda)$.

By definition, $(x, y) \in Conv(\mathcal{FDH}(\mathcal{T}))$ if and only if there exists $(x'_j, y'_j) \in \mathcal{FDH}(\mathcal{T})$ and $\mu_j \in [0, 1]$, $j = 1, 2, \ldots, J$, such that $x = \sum_j \mu_j x'_j$, $y = \sum_j \mu_j y'_j$ and $\sum_j \mu_j = 1$. By definition of the free disposable hull, $(x'_j, y'_j) \in \mathcal{FDH}(\mathcal{T})$ if and only if there exists $(x_j, y_j) \in \mathcal{T}$ such that $x'_j \geq x_j$ and $y_j \leq y'_j$. Let $x(\mu) := \sum_j \mu_j x_j$ and $y(\mu) := \sum_j \mu_j y_j$. Thus, $(x, y) \in Conv(\mathcal{FDH}(\mathcal{T}))$ if and only if there exists $\{(x_1, y_1), (x_2, y_2), \ldots, (x_J, y_J)\} \subset \mathcal{T}$ and $\mu_j \geq 0$ for each $1 \leq j \leq J$ such that $x \geq x(\mu)$ and $y \leq y(\mu)$.

Since the conditions that characterize each set are identical, these two sets coincide. □

Proposition 3.22(a) is the basis for the following definition.

Definition 3.23. *The* **convex, free disposable hull** *of a technology set* $\mathcal{T}$ *is*

$$\mathcal{CFDH}(\mathcal{T}) := \mathcal{FDH}(Conv(\mathcal{T})) = Conv(\mathcal{FDH}(\mathcal{T})).$$

Proposition 3.24. *If* $\mathcal{T}'$ *is a freely disposable, convex technology that contains the technology set* $\mathcal{T}$*, then*

$$\mathcal{CFDH}(\mathcal{T}) \subset \mathcal{T}'.$$

Thus, $\mathcal{CFDH}(\mathcal{T})$ *is the smallest convex, freely disposable technology set that contains* $\mathcal{T}$*.*

Proof. Since $\mathcal{T}'$ is convex and contains $\mathcal{T}$ it follows that $Conv(\mathcal{T}) \subset \mathcal{T}'$. Consequently,

$$\mathcal{CFDH}(\mathcal{T}) = \mathcal{FDH}(Conv(\mathcal{T})) \subset \mathcal{FDH}(\mathcal{T}') = \mathcal{T}',$$

as claimed. □

3.6.3 Constant Returns-to-Scale

Definition 3.25. *A technology exhibits* **constant returns-to-scale** *if*

$$(x, y) \in \mathcal{T} \implies \sigma(x, y) = (\sigma x, \sigma y) \in \mathcal{T} \quad \text{for all} \quad \sigma \in I\!R_+.$$

It will be referred to as a CRS **technology**.

Definition 3.26. *The* **constant returns-to-scale hull** *of a technology set* $\mathcal{T}$ *is*

$$\mathcal{CRS}(\mathcal{T}) := \{(x', y') : \ (x', y') = \sigma(x, y) \ \text{ for some } \ (x, y) \in \mathcal{T} \ \text{ and } \sigma \in I\!R_+\}.$$

The following proposition easily follows from the definitions.

Proposition 3.27. *The constant returns-to-scale hull has these properties:*

a) $\mathcal{CRS}(\mathcal{T})$ *is the smallest cone containing* $\mathcal{T}$.
b) If $\mathcal{T} \subset \mathcal{T}'$, *then* $\mathcal{CRS}(\mathcal{T}) \subset \mathcal{CRS}(\mathcal{T}')$.
c) $\mathcal{CRS}(\mathcal{T} \bigcup \mathcal{T}') = \mathcal{CRS}(\mathcal{T}) \bigcup \mathcal{CRS}(\mathcal{T}')$.
d) If a technology set $\mathcal{T}$ *exhibits constant returns-to-scale, then* $\mathcal{T} = \mathcal{CRS}(\mathcal{T})$.

Proposition 3.28. *The derived production function of a CRS technology is linearly homogeneous.*

Proof. Pick $x \in I\!R^n_+$ and $\sigma > 0$. Since $\mathcal{T}$ exhibits constant returns-to-scale,

$$(\sigma x, u) \in \mathcal{T} \iff (x, \ u/\sigma) \in \mathcal{T}.$$

Consequently,

$$\begin{aligned} \Phi^{\mathcal{T}}(\sigma x) &= \max\{u : \ (\sigma x, u) \in \mathcal{T}\} \\ &= \sigma \max\{(u/\sigma) : \ (x, \ u/\sigma) \in \mathcal{T}\} \\ &= \sigma \max\{v : \ (x, v) \in \mathcal{T}\} \\ &= \sigma \Phi^{\mathcal{T}}(x), \end{aligned}$$

which establishes the desired result. □

Definition 3.29. *The* **convex, constant returns-to-scale hull** *of a technology* $\mathcal{T}$ *is*

$$\mathcal{CCRS}(\mathcal{T}) := \mathcal{CRS}(Conv(\mathcal{T})) = Conv(\mathcal{CRS}(\mathcal{T})).$$

The proof of the next proposition follows similar arguments used to establish the previous propositions and is therefore left to the reader.

Proposition 3.30. *Let* $\mathcal{T}'$ *be a convex, constant returns-to-scale technology that contains the technology* $\mathcal{T}$. *Then*

$$\mathcal{CCRS}(\mathcal{T}) \subset \mathcal{T}'.$$

Thus, $\mathcal{CCRS}(\mathcal{T})$ *is the smallest convex, constant returns-to-scale technology that contains the technology* $\mathcal{T}$.

We close this subsection with a fundamental result due to R.W. Shephard. (A homework exercise asks you to prove this result.)

Proposition 3.31. *A constant returns-to-scale, quasiconcave production function is necessarily concave.*

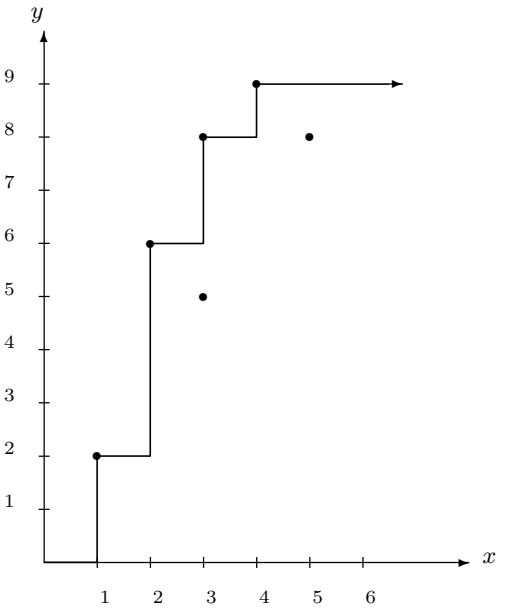


Fig. 3.1. Free disposable hull of the data set.

3.6.4 Example

We illustrate the different hull operations applied to a simple data set $\mathcal{D}$ associated with a single-input, single-output technology provided in Table 3.1.

Table 3.1. Sample input, output data.

Firm	x	u
1	1.0	2.0
2	4.0	9.0
3	2.0	6.0
4	3.0	8.0

The first extrapolation is to form the free disposable hull of the data set, namely, $\mathcal{T}_1 := \mathcal{FDH}(\mathcal{D})$. It is depicted in Figure 3.1. A few observations are in order.

- The technology set $\mathcal{T}_1$ is *not* convex.
- The derived production function is *not* continuous. There are jumps in output precisely at the observed input levels. For example, $\Phi^{\mathcal{T}_1}(x) = 2$ for every $x \in [1, 2)$ but $\Phi^{\mathcal{T}_1}(2) = 6$.

- The derived production function is upper semicontinuous and quasiconcave.[2] An input possibility set associated with an observed output level is a closed interval. As examples, $L(3) = [2, \infty)$ and $L(8) = [3, \infty)$.
- The derived production function is bounded. The highest output level is 9 *regardless* of the input level.

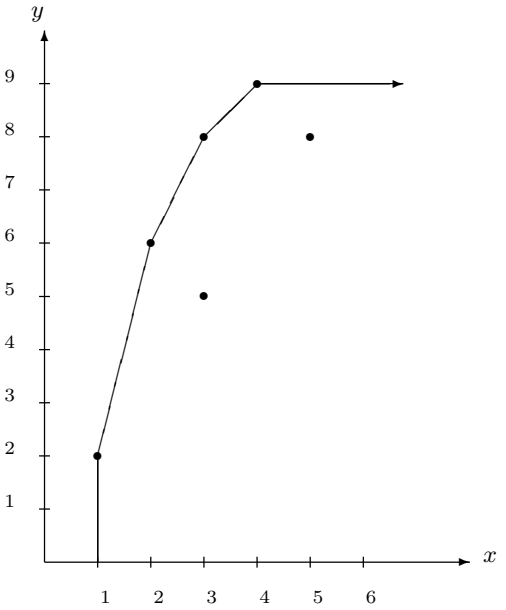


Fig. 3.2. Free disposable, convex hull of the data set.

A less conservative extrapolation is to take the convex hull of the free disposable hull of the data set, namely, $\mathcal{T}_2 := Conv(\mathcal{T}_1)$. (The technology set $\mathcal{T}_2$ is also the free disposable hull of the convex hull of the data set.) It is depicted in Figure 3.2. In this case, the derived production function is concave, piecewise linear, continuous and bounded. Moreover, the derived production function associated with $\mathcal{T}_2$ is (trivially) larger than the one associated with $\mathcal{T}_1$.

The final extrapolation we consider here is the constant returns-to-scale hull of $\mathcal{T}_2$, namely, $\mathcal{T}_3 := \mathcal{CCRS}(\mathcal{T}_2)$. (That is, the technology $\mathcal{T}_3$ is the constant returns-to-scale, free disposable, convex hull of the data set.) It is depicted in Figure 3.3. In this case, the derived production function is (obviously) linearly homogeneous, concave and continuous. Moreover, the derived produc-

[2] Quasiconcavity is only guaranteed in the single-input case. For a typical data set, it will *never* hold when there are multiple inputs.

tion function associated with $\mathcal{T}_3$ is (trivially) larger than the one associated with $\mathcal{T}_2$.

Hopefully, this simple example illustrates how far one can extrapolate a data set to form technology sets with different properties. The "gaps" between $\mathcal{T}_1$, $\mathcal{T}_2$ and $\mathcal{T}_3$ can be quite "large" for a typical data set. Consequently, the efficiency and productivity analysis we undertake in Parts II and III can lead to *significantly* different conclusions depending on the assumptions the modeler is willing to make about the technology.

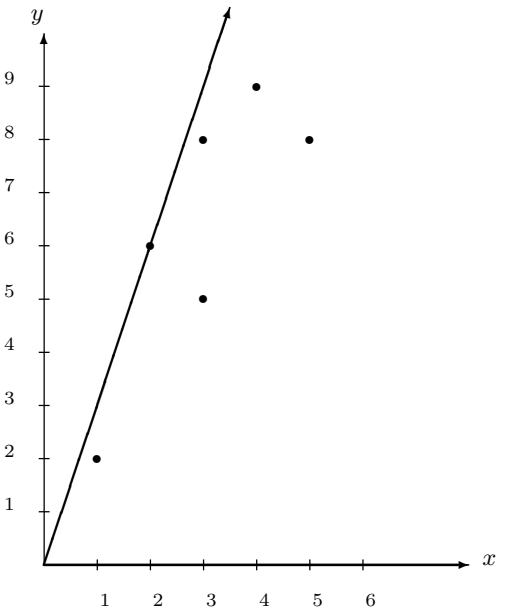


Fig. 3.3. Constant returns-to-scale, free disposable, convex hull of the data set.

3.7 Exercises

3.1. Suppose each $f_k(\cdot)$, $k = 1, 2, \ldots, N$, is quasiconcave. Show that $\Phi(x) = \min_k f_k(x)$ is also quasiconcave.

3.2. For a single-output, well-behaved technology show that $P(x)$ is upper hemicontinuous.

3.3. Give a simple example of a single-input, single-output technology that is not lower hemicontinuous.

3.4. Give an example of a two-input, single-output technology set that is not closed.

3.5. Let $\mathcal{D}$ be a finite data set and let $\mathcal{T} = \mathcal{FDH}(\mathcal{D})$. For a single-output technology, show that the derived production function is *always* upper semi-continuous and bounded.

3.6. Give an example of a two-input, single-output data set $\mathcal{D}$ such that the derived production function associated with the technology $\mathcal{T} = \mathcal{FDH}(\mathcal{D})$ is *not* quasiconcave.

3.7. Prove Proposition 3.21.

3.8. Prove Proposition 3.27.

3.9. Prove Proposition 3.30.

3.10. Prove Proposition 3.31.

3.8 Bibliographical Notes

The first rigorous modern axiomatic treatment of steady-state production is due to Shephard [1953, 1970] and Debreu [1959]. Tulkens [1993] discusses some of the methodological issues pertaining to the free disposable hull. Hackman and Russell [1995] show how to represent and analyze closeness of two technologies in a formal way (via the topology of closed convergence).

Consult Bunge [1959, 1973] for a philosopher's perspective on the axiomatic and scientific method.

3.9 Solutions to Exercises

3.1 First, the algebraic argument. Pick x, y and $\lambda \in [0,1]$. We have

$$\begin{aligned}\Phi(\lambda x + (1-\lambda y)) &= \min_k f_k(\lambda x + (1-\lambda y)) \\ &\geq \min_k \Big[\min(f_k(x), f_k(y))\Big] \quad \text{since each } f_k \text{ is quasiconcave} \\ &= \min\Big[\min_k f_k(x),\ \min_k f_k(y)\Big] \\ &= \min(\Phi(x), \Phi(y)),\end{aligned}$$

which establishes the quasiconcavity of $\Phi(\cdot)$. Now, the set-theoretic argument. Fix u. As a direct consequence of its definition, it follows that $z \in L^{\geq}_{\phi}(u)$ if and only if $z \in L^{\geq}_{f_k}(u)$ for each k. Thus, $L^{\geq}_{\phi}(u) = \bigcap_k L^{\geq}_{f_k}(u)$. Since each upper level set of $f_k(\cdot)$ is convex and the intersection of convex sets is convex, each upper level set of $\Phi(\cdot)$ is convex. This establishes quasiconcavity.

3.2 Since the correspondence $P(\cdot)$ is obviously compact-valued, we shall use Theorem G.9, p. 498, to establish the desired result. To this end let $x_n \to x$ and pick $y_n \in P(x_n)$. Let $e := (1, 1, \ldots, 1) \in \mathbb{R}^n$ be the vector of ones and pick $\delta > 0$. Eventually $x_n \leq x + \delta e$, which implies that $\Phi(x_n) \leq \Phi(x + \delta e)$. Since $y_n \in P(x_n)$, we have that eventually $y_n \leq \Phi(x_n) \leq \Phi(x + \delta e)$. Thus it is possible to extract a subsequence of the y_n that converge to a point y. (We shall not change notation for the subsequence.) It remains to show that $y \in P(x)$. Pick $\epsilon > 0$. Eventually $y_n \geq y - \epsilon$. Since $\Phi(x_n) \geq y_n$, it follows that eventually $\Phi(x_n) \geq y$ or, equivalently, $x_n \in L^{\geq}_{\Phi}(y)$. As each upper level set of $\Phi(\cdot)$ is closed, we may conclude that $x \in L^{\geq}_{\Phi}(y)$ or, equivalently, that $y \in P(x)$, as desired.

3.3 Let $\Phi(x) = \sqrt{x}$ on $[0,1)$ and let $\Phi(x) = 1 + \sqrt{x}$ on $[1,\infty)$. Set $y = 2$. Obviously $y \in P(1)$. However, as $x_n \to 1$ from below it is not possible to find $y_n \in P(x_n)$ such that $y_n \to y$. Consequently, this technology does not exhibit lower hemicontinuity.

3.4 Consider a technology that achieves exactly output rate one as long as both inputs exceed one. The input possibility set $L(1)$ is the input free disposable hull of the point $(1,1)$ with the point $(1,1)$ removed. This technology is clearly not closed.

3.5 Boundedness follows from the fact that

$$\Phi^{\mathcal{T}}(x) = \max\{y_i :\ x_i \leq x\} \leq \max_i y_i < \infty.$$

As for upper semicontinuity,

$$L(u) = \bigcup_{\{i: y_i \geq u\}} \mathcal{IFDH}(x_i),$$

which is a *finite* union of closed sets and hence closed.

3.6 Let $x_1 = (1, 2)$ and $x_2 = (2, 1)$. Consider a technology that achieves exactly output rate one such that $L(1) = \mathcal{IFDH}(x_1) \bigcup \mathcal{IFDH}(x_2)$. This upper level set satisfies all of the usual properties, except that it is not convex.

3.7 We shall only provide the arguments for the input free disposable hull, as the arguments for the other cases are analogous. We begin with part (a). Say set S possesses input free disposability and set $L \subset S$. It remains to show that $\mathcal{IFDH}(L) \subset S$. Pick $x \in L$ and $x' \geq x$. By definition $x' \in \mathcal{IFDH}(L)$. Since $x \in S$, too, and S possesses input free disposability, it follows that $x' \in S$, which establishes the result. A similar argument establishes part (b). As for part (c), we wish to show that

$$\mathcal{IFDH}(S \cup T) = \mathcal{IFDH}(S) \cup \mathcal{IFDH}(T).$$

Pick $x \in S \cup T$ and $x' \geq x$. By definition $x' \in \mathcal{IFDH}(S \cup T)$. Since x' also belongs to both S and T separately, it follows that $x' \in \mathcal{IFDH}(S)$, $x' \in \mathcal{IFDH}(T)$, too. Thus,

$$\mathcal{IFDH}(S \cup T) \subset \mathcal{IFDH}(S) \cup \mathcal{IFDH}(T).$$

The reverse inclusion follows directly from part (b), and the fact that both S and T are obviously subsets of $S \cup T$.

3.8 Part (a) follows from the fact that if C is any cone containing $\mathcal{T}$ it must also contain $\mathcal{CRS}(\mathcal{T})$ by virtue of it being a cone. As for part (b), pick $(x', y') \in \mathcal{CRS}(\mathcal{T})$. By definition $(x', y') = \sigma(x, y)$ for some $(x, y) \in \mathcal{T}$ and $\sigma \in \mathbb{R}_+$. Since $\mathcal{T} \subset \mathcal{T}'$, $(x, y) \in \mathcal{T}'$, too, from which it immediately follows that $(x', y') \in \mathcal{CRS}(\mathcal{T}')$, as required. As for part (c), pick $(x', y') \in \mathcal{CRS}(\mathcal{T} \cup \mathcal{T}')$. By definition $(x', y') = \sigma(x, y)$ for some $(x, y) \in \mathcal{T} \cup \mathcal{T}'$ and $\sigma \in \mathbb{R}_+$. Without loss of generality, we may assume that $(x, y) \in \mathcal{T}$. It then follows that $(x', y') \in \mathcal{CRS}(\mathcal{T}) \subset \mathcal{CRS}(\mathcal{T}) \cup \mathcal{CRS}(\mathcal{T}')$. Thus,

$$\mathcal{CRS}(\mathcal{T} \cup \mathcal{T}') \subset \mathcal{CRS}(\mathcal{T}) \cup \mathcal{CRS}(\mathcal{T}').$$

The reverse inclusion follows from part (b) and the fact that both $\mathcal{T}$ and $\mathcal{T}'$ obviously belong to their union. As for part (d), by definition $\mathcal{T} \subset \mathcal{CRS}(\mathcal{T})$. Pick $(x', y') \in \mathcal{CRS}(\mathcal{T})$. By definition $(x', y') = \sigma(x, y)$ for some $(x, y) \in \mathcal{T}$ and $\sigma \in \mathbb{R}_+$. Since by assumption $\mathcal{T}$ exhibits constant returns-to-scale, it follows that $(x', y') \in \mathcal{T}$, too, and so $\mathcal{T} = \mathcal{CRS}(\mathcal{T})$.

3.9 Pick $(x', y') \in \mathcal{CCRS}(\mathcal{T})$. By definition $(x', y') = \sigma(x, y)$ for some $(x, y) \in Conv(\mathcal{T})$ and $\sigma \in \mathbb{R}_+$. Since $\mathcal{T}'$ is a convex set, it contains $Conv(\mathcal{T})$, and so $(x, y) \in \mathcal{T}'$, too. Since $\mathcal{T}'$ also exhibits constant returns-to-scale, it follows that $(x', y') \in \mathcal{T}'$, too, which establishes the desired result.

3.10 First, we show that $\Phi(\cdot)$ is *super-additive*, namely, $\Phi(x+y) \geq \Phi(x)+\Phi(y)$. If either $\Phi(x)$ or $\Phi(y)$ equals zero, then super-additivity follows from the monotonicity of $\Phi(\cdot)$. Hereafter, we assume that both $\Phi(x)$ and $\Phi(y)$ are positive.

Since $\Phi(\cdot)$ exhibits constant returns-to-scale, i.e., it is linearly homogeneous, then both $x/\Phi(x)$ and $y/\Phi(y)$ belong to the input possibility set $L^{\geq}_{\Phi}(1)$. Since this input possibility set is convex, it follows that

$$\Big[\frac{\Phi(x)}{\Phi(x)+\Phi(y)}\frac{x}{\Phi(x)} \;+\; \frac{\Phi(y)}{\Phi(x)+\Phi(y)}\frac{y}{\Phi(y)}\Big] \in L^{\geq}_{\Phi}(1).$$

This in turn implies that

$$\Phi\Big(\frac{x+y}{\Phi(x)+\Phi(y)}\Big) \geq 1.$$

Super-additivity now follows from the linear homogeneity of $\Phi(\cdot)$. As for concavity, pick x, y and $\lambda \in [0,1]$. We have

$$\Phi(\lambda x + (1-\lambda y)) \geq \Phi(\lambda x) + \Phi((1-\lambda)y) = \lambda\Phi(x) + (1-\lambda)\Phi(y),$$

as required. (The first inequality follows by super-additivity and the second equality follows by linear homogeneity.)

4

Nonparametric Models of Technology

Parametric forms for the production function are sufficiently differentiable, which facilitates analysis of technology via calculus. Sometimes, however, a parametric form exhibits a property that can be refuted by the data. For example, the CES function has constant elasticity of substitution and constant returns-to-scale *regardless* of the estimated parameters a, b and ρ (hence the appellation). As another example, the Cobb-Douglas function exhibits *multiplicative separability* of the input factors. The translog function has the fewest restrictions, which has made it a desirable and often used functional form.

In this chapter, we show how to estimate a production technology *without* assuming a parametric functional form, namely, via a nonparametric approach.

4.1 Simple Leontief or Fixed-Coefficients Technology

We begin by describing the basic building block or core "atom" of activity analysis known as the **simple Leontief** or **fixed-coefficients** technology, so-named for its inventor W.W. Leontief.

A simple Leontief technology is characterized by a **technical coefficient vector** $a = (a_1, a_2, \ldots, a_n)$. Each component of a is positive. If a simple Leontief technology is to produce at least one unit of output, then the input level of each factor i, $i = 1, 2, ..., n$, must be at least a_i. More generally, if the technology is to produce at least $u \geq 0$ units of output, then the input level of factor i must be at least ua_i.

Example 4.1. Consider a simple Leontief technology using two inputs, capital and labor, whose technical coefficient vector $a = (3, 3)$. If the input vector $x = (3, 3)$, then the output produced will be one. If $x = (6, 6)$, then the output produced will be two, since $x = 2a$, and if $x = (1.2, 1.2)$, then the output produced will be 0.4 since $x = 0.4a$. Now suppose $x = (K, 3)$ for $K > 3$. For any choice of $u > 1$, the vector x is *not* greater than or equal to

$u \cdot a$, and so this input vector cannot produce an output level higher than one. Consequently, the output produced will still be one. By the same reasoning, the output level will also be one if $x = (3, L)$ for $L > 3$.

What is the output $\Phi(x)$ for a simple Leontief technology? Since $\Phi(x)$ is achievable $x_i \geq \Phi(x)a_i$ for all i, which implies that

$$\Phi(x) \leq \min_i \{x_i/a_i\}. \tag{4.1}$$

Since the right-hand side of (4.1) is an output level achievable using x, it follows from the definition of a production function that

$$\Phi(x) = \min_i \{x_i/a_i\}. \tag{4.2}$$

Example 4.2. Continuing with Example 4.1, suppose $x = (40, 33)$. Direct substitution into (4.2) yields $\Phi(x) = \min\{40/3,\ 33/3\} = 11$. In words, since $x \geq 11a = (33, 33)$, this simple Leontief technology will produce at least 11 units of output with this input vector, i.e., $\Phi(x) \geq 11$. Since for any value of $u > 11$, the input vector $x \ngeq ua$, this input vector cannot produce any output level higher than 11. Consequently, the technology will produce an output level of exactly 11, i.e., $\Phi(x) = 11$. For a simple Leontief technology, slack in any one input does not increase output. In this example, the extra $40 - 33 = 7$ units of capital cannot be used to increase output beyond the level of 11.

It is immediate from (4.2) that $\Phi(\cdot)$ exhibits constant returns-to-scale, which implies that

$$L_\Phi(u) = uL_\Phi(1). \tag{4.3}$$

Geometrically, the input possibility set $L_\Phi(u)$ is the radial expansion of the unit input possibility set $L_\Phi(1)$. That is, the isoquant of $L_\Phi(u)$, $ISOQ_\Phi(u) = \{x : \Phi(x) = u\}$, is obtained by multiplying each point in the isoquant $ISOQ_\Phi(1)$ of $L_\Phi(1)$ by u. The input possibility sets for $u = 1$ and $u = 2$ for the simple Leontief technology of Example 4.1 are depicted in Figure 4.1.

Since a simple Leontief technology exhibits constant returns-to-scale, its elasticity of scale is one. As shown in Figure 4.1, each isoquant is "L-shaped," which implies the elasticities of output are each zero. For a simple Leontief technology, the sum of the output elasticities does not equal the elasticity of scale. (Keep in mind the production function is not differentiable.) Technically, both the rate of technical substitution and elasticity of substitution are not defined. However, it seems intuitive that the elasticity of substitution should be zero. Indeed, the vector $x = ua$ achieves output rate u at minimum cost *regardless* of the factor prices, and so a 1% change in the factor prices will have no effect on the factor ratios. The elasticity of substitution for the simple Leontief is defined to be zero. This follows from the fact that the elasticity of substitution for the CES production function is $1/(1-\rho)$, which tends to zero as ρ tends to $-\infty$, and the following remark.

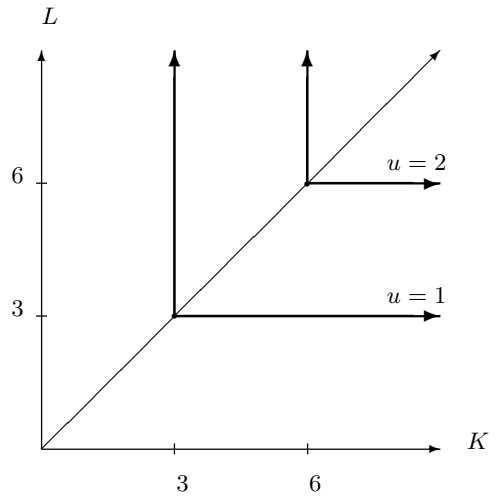


Fig. 4.1. Example of a simple Leontief technology with $a = (3, 3)$.

Remark 4.3. The CES production function $\Phi(x) = [\sum_i \alpha_i x_i^\rho]^{1/\rho}$ converges pointwise to a simple Leontief technology as ρ tends to $-\infty$. To see this, fix x and for each i let $y_i = x_i/a_i$ where $a_i := \alpha_i^{-1/\rho}$. Define

$$\phi_y(\rho) := \left[\sum_i y_i^\rho\right]^{1/\rho},$$

and let $y_{\min} := \min_i y_i$. Since the y_i are positive (i) $\sum_i y_i^\rho \geq y_{\min}^\rho$ and (ii) $y_i^\rho \leq y_{\min}^\rho$ when ρ is negative. Consequently, when ρ is negative,

$$y_{\min} \leq \phi_y(\rho) \leq n^{1/\rho} y_{\min}.$$

Since $n^{1/\rho}$ converges to 1 as ρ tends to $-\infty$, it follows immediately that

$$\lim_{\rho \to -\infty} \phi_y(\rho) = y_{\min} = \min_i \{x_i/a_i\},$$

as claimed.

4.2 General Leontief Technology

Consider a system in which resources can be allocated to any one of N simple Leontief processes $\Phi^k(\cdot)$, $1 \leq k \leq N$. The system output is the sum of the outputs of the simple processes. We shall hereafter refer to this "molecular" system consisting of "N parallel atoms" as a **general Leontief technology.**

4.2.1 Production Function

What is the output $\Phi(x)$ for a general Leontief technology? An *allocation* problem must be solved to determine the maximal output rate, namely, one must disaggregate the aggregate input vector x into allocations x^k to each process k to maximize the sum of the outputs $u^k = \Phi(x^k)$ of the individual processes. That is,

$$\Phi(x) = \max\Big\{ \sum_{k=1}^{N} \Phi^k(x^k) : \ \sum_k x^k \le x, \ x^k \ge 0 \text{ for all } k \Big\}. \tag{4.4}$$

The right-hand side of (4.4) is a *nonlinear* optimization problem. Let $a^k = (a_1^k, a_2^k, \ldots, a_n^k)$ denote the technical coefficient vector for process k. For any desired output rate u^k, it is necessary to allocate at least $u^k a^k$ to process k, and there is no reason to allocate more input to process k. Consequently, we may set $x^k = u^k a^k$, which transforms (4.4) into the *linear program*

$$\Phi(x) = \max\Big\{ \sum_k u^k : \ Au \le x, \ u \ge 0 \Big\}. \tag{4.5}$$

In (4.5), the i^{th} column of the $n \times N$ matrix A is the (transpose of) vector a^i. For the general Leontief technology, one solves linear program (4.5) to determine the production function. The scale of each process or activity k, u^k, is commonly referred to as an **activity intensity**.

4.2.2 Properties

Proposition 4.4. *A general Leontief technology exhibits constant returns-to-scale.*

Proof. Pick $s > 0$. Using (4.5),

$$\begin{aligned} \Phi(sx) &= \max\Big\{ \sum_k u^k : \ Au \le sx \Big\} \\ &= \max\Big\{ s \sum_k (u^k/s) : \ A(u/s) \le x \Big\} \\ &= s \max\Big\{ \sum_k v^k : \ Av \le x \Big\} \\ &= s\Phi(x). \end{aligned} \tag{4.6}$$

□

Since (4.3) holds for a constant returns-to-scale technology, it is only necessary to construct the unit input possibility set $L_\Phi(1)$. To this end, let

$$\mathsf{A} := \{a^1, a^2, \ldots, a^N\}$$

denote the set of the intensity vectors, and consider the convex hull (see Definition C.6, p. 462) of A,

$$Conv(\mathsf{A}) := \Big\{ x = Au : \sum_{k=1}^{N} u^k = 1,\ u^k \geq 0 \text{ for each } k \Big\}. \tag{4.7}$$

Since each point that lies on a line segment joining two intensity vectors a^j and a^k can be represented as $u^j a^j + u^k a^k$ where $u^j + u^k = 1$, $u^j, u^k \geq 0$, each such line segment belongs to the convex hull. In general dimensions, the unit input possibility set is simply the input free disposable hull (see Definition 3.19, p. 42) of the convex hull of the intensity vectors.

Proposition 4.5. $L_{\Phi}(1) = \mathcal{IFDH}(Conv(\mathsf{A}))$.

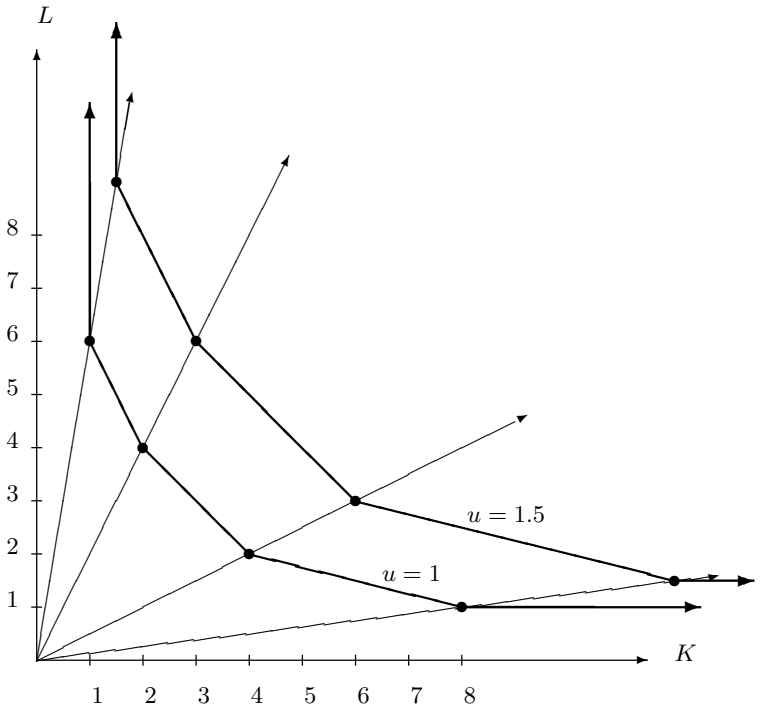


Fig. 4.2. An example of a general Leontief technology.

Example 4.6. Consider a general Leontief technology consisting of four activities whose intensity vectors are $a^1 = (1, 6)$, $a^2 = (2, 4)$, $a^3 = (4, 2)$, and $a^4 = (8, 1)$, respectively. Figure 4.2 depicts the isoquants for output levels 1

and 1.5. The isoquant $ISOQ_{\Phi}(1)$ is constructed by sequentially connecting the intensity vectors, and then adding two rays: one emanating from a^1 heading due north and the other emanating from a^4 heading due east. The input possibility set $L_{\Phi}(1)$ is obtained by adding each vector in $\mathbb{R}^2_+$ to each vector in the isoquant. Algebraically,

$$L_{\Phi}(1) = \Big\{(K,L):\ u^1\begin{pmatrix}1\\6\end{pmatrix} + u^2\begin{pmatrix}2\\4\end{pmatrix} + u^3\begin{pmatrix}4\\2\end{pmatrix} + u^4\begin{pmatrix}8\\1\end{pmatrix} \ \leq\ \begin{pmatrix}K\\L\end{pmatrix} \\ u^1+u^2+u^3+u^4 \ \geq\ 1 \quad \Big\},$$

or in matrix notation,

$$L_{\Phi}(1) = \Big\{(K,L):\ \begin{bmatrix}1 & 2 & 4 & 8\\ 6 & 4 & 2 & 1\\ -1 & -1 & -1 & -1\end{bmatrix}\begin{pmatrix}u^1\\u^2\\u^3\\u^4\end{pmatrix} \ \leq\ \begin{pmatrix}K\\L\\-1\end{pmatrix}\ \Big\}.$$

It is understood the u^i are nonnegative and for convenience we drop these constraints. The output rate $\Phi(x)$ for an input vector x is

$$\Phi(x) \ =\ \max\Big\{u^1+u^2+u^3+u^4:\ \begin{bmatrix}1&2&4&8\\6&4&2&1\end{bmatrix}\begin{pmatrix}u^1\\u^2\\u^3\\u^4\end{pmatrix} \ \leq\ \begin{pmatrix}K\\L\end{pmatrix}\ \Big\}. \tag{4.8}$$

In canonical form, the right-hand side of (4.8) is a linear program

$$\max\{c^T x:\ Ax \leq b,\ x \geq 0\},$$

where $x = (u^1,u^2,u^3,u^4)^T$, $c^T = (1,1,1,1)$, $b = (K,L)^T$ and

$$A = \begin{bmatrix}1&2&4&8\\6&4&2&1\end{bmatrix}.$$

4.2.3 Graphical Construction

When there are two inputs, the output rate can be graphically determined, as follows. Since (4.6) holds we know $x = sx'$ for some $x' \in ISOQ_{\Phi}(1)$; in particular, $s = \Phi(x)$. The four rays passing through the origin and each of the intensity vectors together with the x- and y-axis determine five regions, as depicted in Figure 4.3. (Each region is a convex cone, see Definition C.15, p. 464.) The slopes of the rays identify the boundaries of the labor-capital ratio for each region. Let m denote the labor-capital ratio of x and thus x'. Given m it is easy to identify the region to which x' belongs, say region r. Since x' also belongs to the unit isoquant, we know x' lies at the intersection

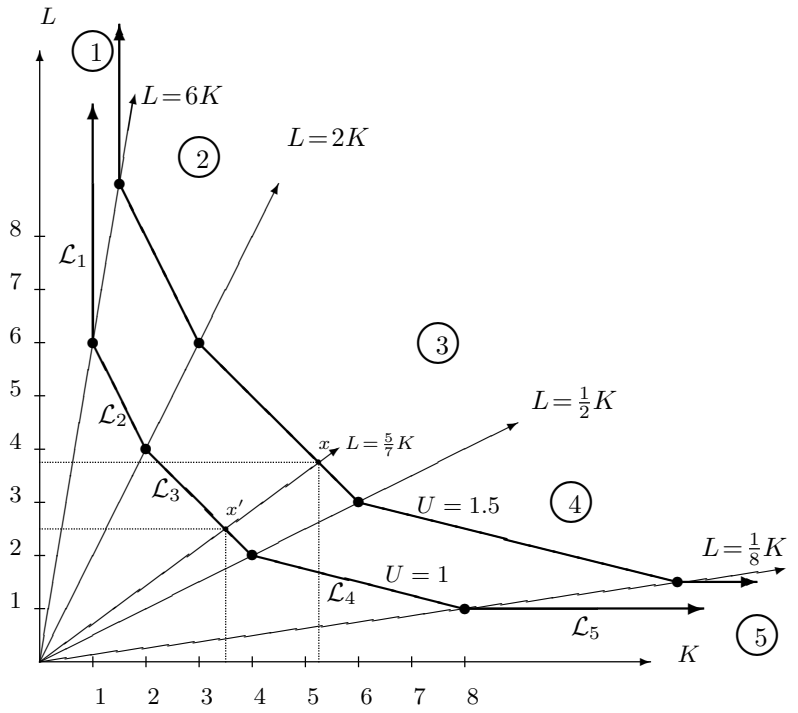


Fig. 4.3. Computing output for a general Leontief technology with two inputs.

of the line $\mathcal{L}_r$ and the line $L = mK$. The two equations uniquely determine the coordinates of x', from which it is easy to determine s. For region 1, the equation of line $\mathcal{L}_1$ is $K = a_1^1$ (the first coordinate of a^1), and for region 5, the equation of line $\mathcal{L}_5$ is $L = a_2^4$ (the second coordinate of a^4).

Example 4.7. Consider the input vector $x = (5.25, 3.75)$. Its labor-capital ratio is $5/7$, and so x' falls within region 3. Since x' also lies on the unit isoquant, it also belongs to the line $\mathcal{L}_3$ determined by points a^2 and a^3, which is $L = -K + 6$. Consequently, x' lies at the intersection of the two lines $L = (5/7)K$ and $L = -K + 6$, which implies $x' = (3.5, 2.5)$ and $s = 1.5$. As Figure 4.3 shows, $x \in ISOQ_{\Phi}(1.5)$ and so $\Phi(x) = 1.5$.

Example 4.8. Consider the input vector $x = (2, 16)$. Its labor-capital ratio is 8 and so x' falls within region 1. Here, x' lies at the intersection of the lines $L = 8K$ and $K = 1$, and so $x' = (1, 8)$ and $s = 2$.

Remark 4.9. The function $L(\cdot)$ that implicitly defines the isoquant of a general Leontief technology is **piecewise linear**. A general Leontief technology belongs to the class of **piecewise linear technologies**.

4.3 Nonparametric Constructions

The central question we now address is how to nonparametrically construct a technology via input possibility sets from observed data (x_i, u_i) on N firms. (The parametric approach constructs the technology via the production function route and some type of estimation technique.) Basically, the goal is to define level sets $L^a(u)$ that approximate the true unknown input possibility sets $L(u)$.

4.3.1 The Hanoch-Rothschild Model of Technology

We begin by assuming the technology is well-behaved. Let

$$\mathcal{X}(u) := \{x_j\}_{j \in \mathcal{I}(u)} \text{ where } \mathcal{I}(u) := \{j : u_j \geq u\}. \tag{4.9}$$

Since the true input possibility set $L(u)$ is convex and exhibits input free disposability,

$$\mathcal{IFDH}(Conv(\mathcal{X}(u))) \subset L(u). \tag{4.10}$$

The Hanoch-Rothschild approach stops here, namely, it approximates $L(u)$ with the smallest convex, input free disposable set that is consistent with the observed data set.

Definition 4.10. *The* **Hanoch-Rothschild model of technology (HR technology)** *is given by the family of input possibility sets*

$$\mathcal{F}^{HR} := \{L^{HR}(u) : u \geq 0\},$$

where for each $u \geq 0$,

$$L^{HR}(u) := \mathcal{IFDH}(Conv(\mathcal{X}(u))).$$

In equation form,

$$L^{HR}(u) := \Big\{z : z \geq \sum_{j \in \mathcal{I}(u)} \lambda_j x_j, \sum_{j \in \mathcal{I}(u)} \lambda_j = 1, \lambda_j \geq 0 \text{ for } j \in \mathcal{I}(u)\Big\}. \tag{4.11}$$

$L^{HR}(u)$ will be empty when $\mathcal{I}(u)$ is empty.

Remark 4.11. Hanoch-Rothschild [1972] developed this simple model as a means to test whether or not a given data set could be consistent with an upper semincontinuous, nondecreasing quasiconcave production function. From an applications perspective, the HR technology is only useful when the technology produces a single output.

The following proposition follows from the definition of the HR technology.

Proposition 4.12. *The HR technology is the smallest well-behaved technology set that contains a given data set.*

By construction, each $L^{HR}(u)$ is the most *conservative* estimate of $L(u)$. The most "liberal" definition of a technology that is consistent with the data is to set each $L(u)$ to be $\mathbb{R}^n_+$.[1] This construction is hardly useful, however!

[1] Less a small neighborhood about zero to be consistent with Axiom **A1**.

Remark 4.13. Each input possibility set of an HR technology can be identified with the input possibility set associated with output rate one of a general Leontief technology whose intensity vectors are given by the vectors in $\mathcal{I}(u)$.

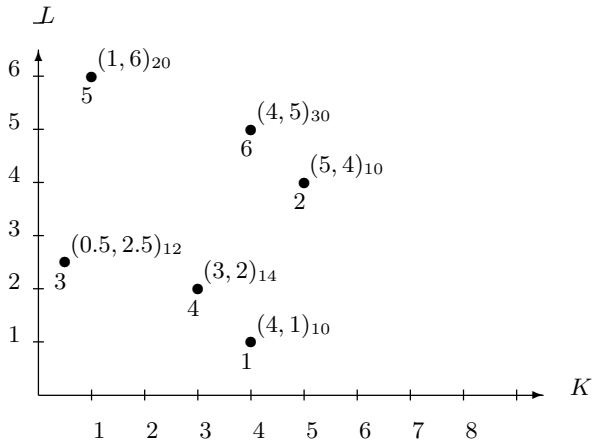


Fig. 4.4. An input-output data set. The number next to each input vector is the output rate.

Example 4.14. For the input-output data shown in Figure 4.4, the input possibility sets for the HR technology for each of the observed output rates are depicted in Figure 4.5. For example,

$$L(14) = \left\{(K,L):\ u^4\begin{pmatrix}3\\2\end{pmatrix} + u^5\begin{pmatrix}1\\6\end{pmatrix} + u^6\begin{pmatrix}4\\5\end{pmatrix} \le \begin{pmatrix}K\\L\end{pmatrix} \right.$$
$$\left. u^4 + u^5 + u^6 = 1 \right\}.$$

Note the data point corresponding to firm 6 does not determine the isoquant of $L(14)$.

4.3.2 Data Envelopment Analysis Models of Technology

Data Envelopment Analysis (DEA), initially developed in 1978 by Charnes, Cooper and Rhodes and extended in many ways over the past 25 years, adopts a less conservative approach to extrapolating the data than the HR-approach. It also can be used in the *multi-output* setting.

To motivate the DEA approach, consider the following hypothetical example. Firm A uses $(K, L) = (4, 4)$ to produce 15. The goal is to measure its input efficiency. To do so requires an estimate the input possibility set $L(15)$. Data on two other firms B and C have been collected: firm B uses $(K, L) = (1, 5)$

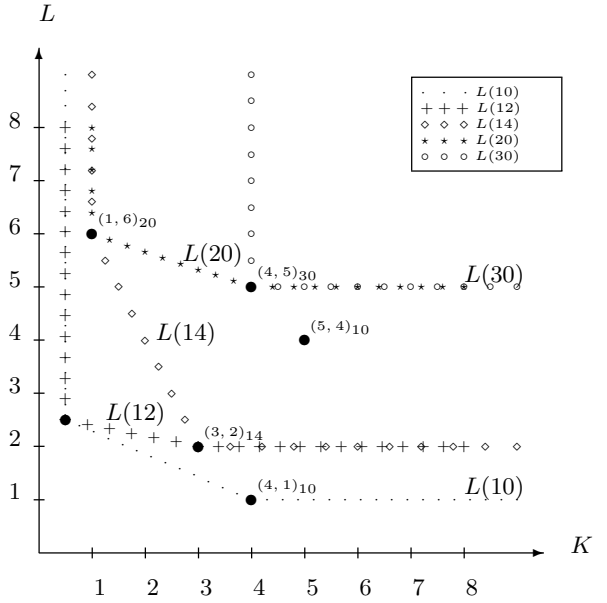


Fig. 4.5. Input possibility sets for the HR technology.

to produce 10 and firm C uses $(K, L) = (5, 1)$ to produce 20. Assuming the HR technology, as depicted in Figure 4.6(a), firm B is irrelevant and firm C cannot reveal firm A to be inefficient since the point (4,4) does not lie "on or above" the point (5,1). Now a 50-50 mixture of the *outputs* of firms B and C results in the number 15. DEA assumes that such output is attainable if a 50-50 mix of the inputs of firms B and C is used. Under this more relaxed assumption, this "composite" firm will produce 15 and use $(K, L) = (3, 3)$. As depicted in Figure 4.6(b), this composite firm would reveal firm A to be 75% efficient.

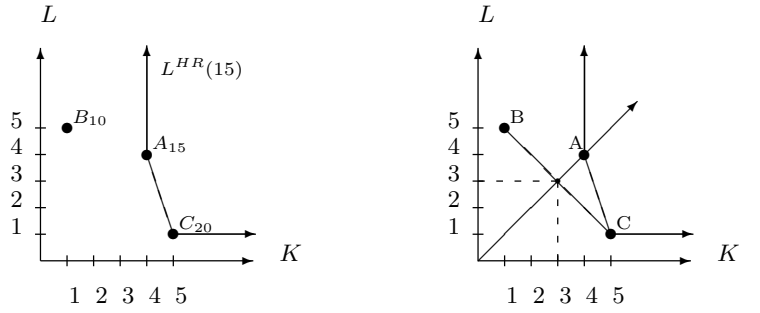
Charnes, Cooper and Rhodes' main contribution is to broaden the application of convexity to *both* inputs and outputs.

Definition 4.15. *The* **Variable Returns-to-Scale DEA model of technology (VRS technology)** *is the smallest closed convex freely disposable set that contains the data set $\mathcal{D}$, namely,*

$$\mathcal{T}^{VRS} := \mathcal{FDH}(Conv(\mathcal{D})).$$

Definition 4.16. *The* **Constant Returns-to-Scale DEA model of technology (CRS technology)** *is the smallest constant returns-to-scale technology that contains the VRS technology, namely,*

$$\mathcal{T}^{CRS} := \mathcal{CRS}(\mathcal{T}^{VRS}).$$



(a) HR technology (b) DEA technology

Fig. 4.6. Comparison between the HR and DEA models of technology. HR technology convexifies input but not output, while DEA convexifies both input and output.

Keep in mind that the definitions of $\mathcal{T}^{CRS}$ and $\mathcal{T}^{VRS}$ do *not* require output to be scalar; that is, in stark contrast to the HR technology, these technologies apply in the multi-output setting.

For each $\lambda \in \mathbb{R}^N_+$, let

$$x(\lambda) := \sum_{i=1}^{N} \lambda_i x_i \quad \text{and} \quad y(\lambda) := \sum_{i=1}^{N} \lambda_i y_i. \tag{4.12}$$

With this notation, that

$$\mathcal{T}^{CRS} = \Big\{(x, y) : \ x \geq x(\lambda), \ y \leq y(\lambda), \ \lambda_i \geq 0\Big\}, \tag{4.13}$$

and

$$\mathcal{T}^{VRS} = \Big\{(x, y) : \ x \geq x(\lambda), \ y \leq y(\lambda), \sum_i \lambda_i = 1, \lambda_i \geq 0\Big\}. \tag{4.14}$$

Remark 4.17. Historically, the first DEA model was the CRS model developed by Charnes, Cooper and Rhodes in 1978. It was extended by Banker, Charnes and Cooper in 1984 to the VRS model. In the parlance of DEA, each firm is called a **Decision-Making Unit (DMU)**.

4.3.3 Graphical Constructions

Given data (x_i, u_i), $i = 1, 2, ..., N$, it is possible to graphically construct the input possibility sets associated with $\mathcal{T}^{CRS}$ and $\mathcal{T}^{VRS}$ in the case of two inputs and scalar output. Such graphical depiction is useful to explain concepts to users, and it is not unusual for a pilot project to involve a single aggregate output with aggregate inputs measuring some form of capital and labor.

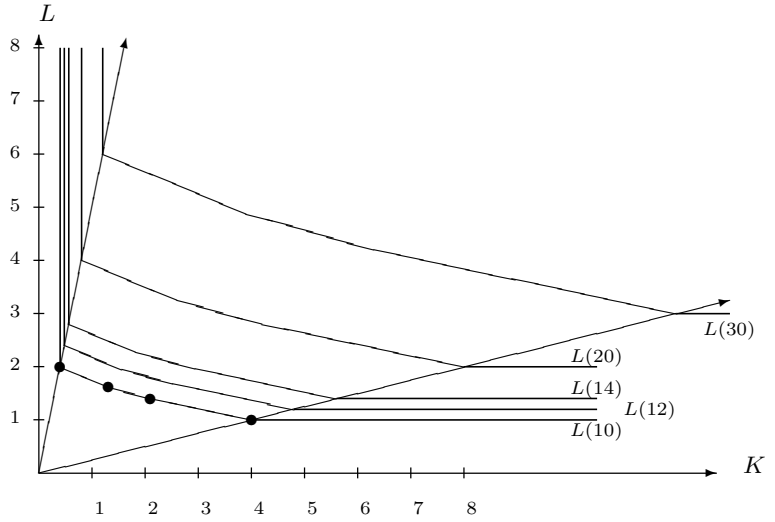


Fig. 4.7. Input possibility sets for the CRS technology.

We first examine the simplest case corresponding to $\mathcal{T}^{CRS}$. Since

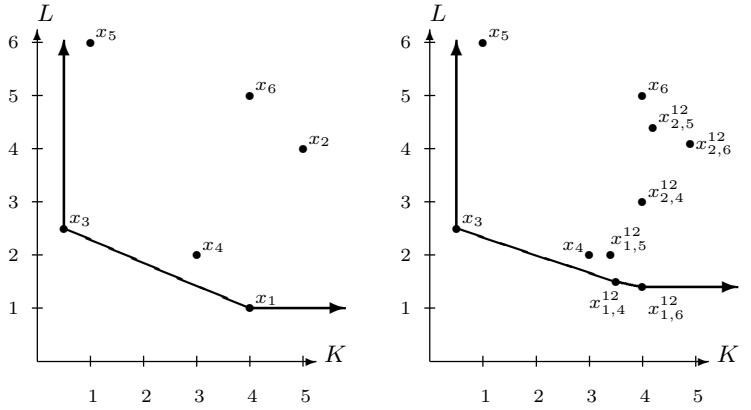
$$L^{CRS}(u) = uL^{CRS}(1),$$

it suffices to show how to construct $L^{CRS}(1)$. Let $\hat{x}_i$ denote the scaled input vector x_i/u_i. Clearly, $\hat{x}_i \in L^{CRS}(1)$ for each i. Since $L^{CRS}(1)$ is convex, all convex combinations of the $\hat{x}_i$ belong to it, too. That is, the convex hull of the $\hat{x}_i$ is a subset of $L^{CRS}(1)$. Since each input possibility set exhibits input free disposability, each input vector whose coordinates are at least as large as some convex combination of the $\hat{x}_i$ must also belong to $L^{CRS}(1)$. This collection of points is precisely the HR-construction applied to the "scaled" data set $(\hat{x}_i, 1)$, $i = 1, 2, ..., N$. Equivalently, the CRS technology can be identified with a general Leontief technology in which the intensity vectors are given by the $\hat{x}_i$.

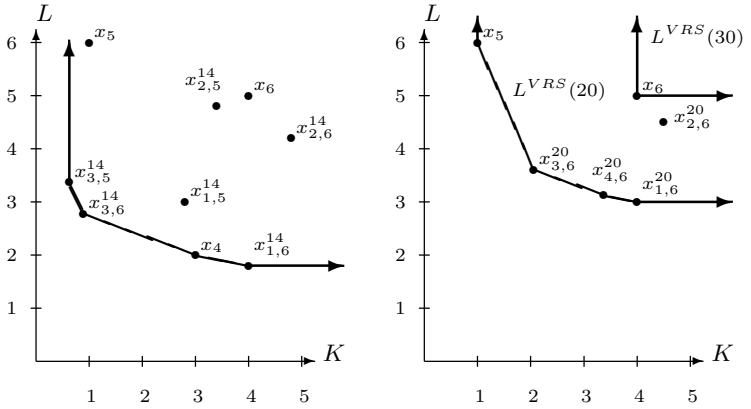
Example 4.18. For the data given in Figure 4.4, Figure 4.7 depicts the input possibility sets for the CRS technology. For example, $L^{CRS}(20)$ is the convex, input free disposable hull of the vectors

$$\hat{x}_1 = (20/10)\begin{pmatrix}4\\1\end{pmatrix},\ \hat{x}_2 = (20/10)\begin{pmatrix}5\\4\end{pmatrix},\ \hat{x}_3 = (20/12)\begin{pmatrix}0.5\\2.5\end{pmatrix},$$
$$\hat{x}_4 = (20/14)\begin{pmatrix}3\\2\end{pmatrix},\ \hat{x}_5 = (20/20)\begin{pmatrix}1\\6\end{pmatrix},\ \hat{x}_6 = (20/30)\begin{pmatrix}4\\5\end{pmatrix}.$$

The construction of the input possibility set $L^{VRS}(u)$ associated with $\mathcal{T}^{VRS}$ is somewhat more involved. We shall define a set $G(u)$ of input vectors or *generators* from which one simply executes the HR-construction; that is, form the convex hull of $G(u)$ and extend "upwards and outwards." Formally,



(a) $L^{VRS}(10)$ (b) $L^{VRS}(12)$



(c) $L^{VRS}(14)$ (d) $L^{VRS}(20)$ and $L^{VRS}(30)$

Fig. 4.8. Input possibility sets for the VRS technology.

$$L^{VRS}(u) = \mathcal{IFDH}(G(u)).$$

First, if $u_i \geq u$, then x_i belongs to $G(u)$. Next, for each j and k for which $u_j > u > u_k$ there exists a *unique* constant $\lambda^u_{j,k} \in (0,1)$ for which

$$u = \lambda^u_{j,k} u_j + (1 - \lambda^u_{j,k}) u_k.$$

Define

$$x^u_{j,k} := \lambda^u_{j,k} x_j + (1 - \lambda^u_{j,k}) x_k. \tag{4.15}$$

Each point $x_{j,k}$ also belongs to $G(u)$.

Example 4.19. For the data given in Figure 4.4, Figure 4.8 depicts the input possibility sets for each of the observed output rates for the VRS technology. For example, the set $G(20)$ contains the original data points x_5 and x_6, as well as the points marked

$$x^{20}_{1,6} := 1/2\ x_1 + 1/2\ x_6, \quad x^{20}_{2,6} := 1/2\ x_2 + 1/2\ x_6,$$
$$x^{20}_{3,6} := 5/9\ x_3 + 4/9\ x_6, \quad x^{20}_{4,6} := 5/8\ x_4 + 3/8\ x_6.$$

Not all points in $G(u)$ will lie on the isoquant of $L^{VRS}(u)$; consider, for example, the point $x^{20}_{2,6}$ in $G(20)$.

4.4 Exercises

4.1. Graphically depict the input possibility set $L(2)$ associated with a simple Leontief technology whose intensity vector is $a = (2, 1)$.

4.2. Consider a general Leontief technology consisting of three activities whose intensity vectors are respectively $a^1 = (1, 5)$, $a^2 = (2, 2)$ and $a^3 = (4, 1)$.

(a) Graphically depict the input possibility set $L(2)$ associated with this technology.
(b) What is $\Phi(12, 28)$?

4.3. Consider the following input-output data obtained from three firms: $x_1 = (2, 4), y_1 = 10$, $x_2 = (3, 16), y_2 = 20$ and $x_3 = (8, 2), y_3 = 20$.

(a) Graphically depict $L^{HR}(20)$.
(b) Graphically depict $L^{CRS}(20)$.
(c) Graphically depict $L^{VRS}(15)$.
(d) Input-output data on a fourth firm has been collected: $x_4 = (9, 7.5), y_4 = 20$. Under the CRS technology:
   i. What is this firm's input efficiency?
   ii. What is this firm's output efficiency?

4.4. Prove Proposition 4.5.

4.5. Prove that $\mathcal{T}^{CRS}$ is the smallest convex cone containing $\mathcal{T}^{VRS}$.

4.6. Let ${\mathcal{T}_F}^{VRS}$ and ${\mathcal{T}_F}^{CRS}$ denote the technologies associated with the standard VRS and CRS assumptions, respectively, except that some of the inputs are *exogenously fixed* or *non-discretionary*—that is, these inputs cannot be scaled downwards. How do the input and output efficiency measures of ${\mathcal{T}_F}^{VRS}$ and ${\mathcal{T}_F}^{CRS}$ compare with those for $\mathcal{T}^{VRS}$ and $\mathcal{T}^{CRS}$?

4.7. This exercise explores the nonparametric estimation of *concave* technologies (see Alon et. al. [2007]). Let $\mathcal{D} := \{(x_1, u_1), (x_2, u_2), \ldots, (x_N, u_N)\}$ denote a data set of N input-output pairs such that $x_i > 0$ (the coordinates of each input vector are strictly positive) and each (scalar) output u_i is strictly positive. Let

$$\mathcal{X} := Conv\{x_1, x_2, \ldots, x_N\} + \mathbb{R}^n_+$$

denote the convex, input free disposable hull of the x_i. Let F denote the set of all functions $f : \mathcal{X} \to \mathbb{R}_+$ such that $f(\cdot)$ is nondecreasing and concave.

(a) For each $x \in \mathcal{X}$, let $\Phi^*(x)$ denote the optimal value of the following linear programming problem:

$$\Phi^*(x) := \max_{\lambda \geq 0} \Big\{ \sum_{i=1}^N \lambda_i u_i \text{ subject to } \sum_{i=1}^N \lambda_i x_i \leq x, \ \sum_{i=1}^N \lambda_i = 1 \Big\}.$$

Prove that $\Phi^*(\cdot) \in \mathcal{F}$.

(b) The data set $\mathcal{D}$ is said to be *concave-representable* if there exists a function $\Phi(\cdot) \in \mathcal{F}$ such that $\Phi(x_i) = u_i$ for each i. Each such $\Phi(\cdot)$ is called a *representation* of $\mathcal{D}$. Prove that if $\mathcal{D}$ is concave-representable, then $\Phi^*(\cdot)$ is its *minimal* representation, namely, $\Phi(\cdot) \geq \Phi^*(\cdot)$ on $\mathcal{X}$ for any other representation $\Phi(\cdot)$ of $\mathcal{D}$.

(c) Prove that the data set $\mathcal{D}$ is concave-representable if and only if $\Phi^*(x_i) = u_i$ for each i.

(d) Provide an example of a single-input, single-output data set $\mathcal{D}$ that is concave-representable. Graphically depict $\Phi^*(\cdot)$ and provide a representation $\Phi(\cdot)$ of $\mathcal{D}$ that does not equal $\Phi^*(\cdot)$.

4.5 Bibliographical Notes

Hanoch-Rothschild's [1972] paper addresses the topic of whether an input-output data set could ever be reconciled with a well-behaved quasiconcave production function. They showed that linear programming models could be used to test this hypothesis, as did Diewert and Parkan [1983]. Activity analysis is due to Koopmans [1951], Debreu [1951] and Leontief [1953].

Data Envelopment Analysis was first introduced in Charnes et. al. [1978] and extended in Banker et. al. [1984]. A family of related models has been developed and applied over the years by many researchers. The monograph by Fried et. al. [1993] and the recent textbooks by Coelli et. al. [2005] and Cooper et. al. [2007] provide detailed descriptions, extensions, and many applications.

4.6 Solutions to Exercises

4.1 It is the input free disposable hull of the point (2,2), which is the set in the (K, L) space given by $\{(K, L) :\ K \geq 2,\ L \geq 2\}$.

4.2 (a) It is the convex, input free disposable hull of the points (1, 5), (2, 2), and (4, 1). This set in the (K, L) space is given by the intersection of the four halfspaces determined by the line parallel to the L axis passing through $(1, 5)$, the line segment joining $(1, 5)$ and $(2, 2)$, the line segment joining $(2, 2)$ and $(4, 1)$, and the line parallel to the K axis passing through the point $(4, 1)$, i.e.,

$$\{(K, L) :\ K \geq 1,\ L \geq -3K + 8,\ L \geq -0.5K + 3,\ L \geq 1\}.$$

(b) The output rate associated with $(12, 28)$ can be obtained by solving the linear program

$$\max\left\{u^1 + u^2 + u^3 :\ \begin{bmatrix}1 & 2 & 4\\ 5 & 2 & 1\end{bmatrix}\begin{bmatrix}u^1\\ u^2\\ u^3\end{bmatrix} \leq \begin{bmatrix}12\\ 28\end{bmatrix}\right\}.$$

As noted in the text, this linear program can be solved geometrically, as follows. The labor-capital ratios corresponding to the intensity vectors a^1, a^2, and a^3 are, respectively, 5, 1 and 0.25. Since the labor-capital ratio of the point $(12, 28)$ is 7/3, the point (12, 28) lies in the cone $\{(K, L) :\ L \leq 5K,\ L \geq K\}$ spanned by the rays emanating from the origin and passing through the points a^1 and a^2, respectively. It follows that the line $L = (7/3)K$ must intersect the line connecting points a^1 and a^2 given by $L = -3K + 8$. The point of intersection is $(1.5, 3.5)$, which, by construction, has output rate equal to one. The output rate associated with the point $(12, 28)$ equals the scalar u for which $(12, 28) = u(1.5, 3.5)$. Here, $u = 8$ and $\Phi(12, 28) = 8$.

4.3 (a) $L^{HR}(20)$ is the convex, input free disposable hull of the points (3, 16) and (8, 2). This set in the (K, L) space is given by the intersection of the three halfspaces determined by the line parallel to the L axis passing through $(4, 8)$, the line segment joining $(4, 8)$ and $(8, 2)$, and the line parallel to the K axis passing through the point $(8, 2)$, i.e.,

$$\{(K, L) :\ K \geq 3,\ L \geq -(14/5)K + 122/5,\ L \geq 2\}.$$

(b) The point $(4, 8)$, which is twice the first intensity vector, will achieve an output rate of 20. Thus, $L^{CRS}(20)$ is the convex, input free disposable hull of the points (3, 16), (4, 8), and (8, 2). This set in the (K, L) space is given by the intersection of the four halfspaces determined by the line parallel to the L axis passing through $(3, 16)$, the line segment joining $(3, 16)$ and $(4, 8)$, the line segment joining $(4, 8)$ and $(8, 2)$, and the line parallel to the K axis passing through the point $(8, 2)$, i.e.,

$$\{(K, L) :\ K \geq 3,\ L \geq -8K + 40,\ L \geq -(6/4)K + 14,\ L \geq 2\}.$$

(c) The points $(3, 16)$ and $(8, 2)$ obviously belong to $L^{VRS}(15)$. It remains to determine the additional generators that belong to the set $G(15)$ defined in the text. Here there are two: since $15 = 0.5(10) + 0.5(20)$, the midpoint of the line segment joining points (2, 4) and (3, 16), which is (2.5, 20), is one and the midpoint of the line segment joining points (2, 4) and (8, 2), which is (5, 3), is the other. The set $L^{VRS}(15)$ is then the convex, input free disposable hull of the points $\{(3, 16), (8, 2), (2.5, 10), (5, 3)\}$. This set is the intersection of the four halfspaces determined by the line parallel to the L axis passing through $(2.5, 10)$, the line segment joining $(2.5, 10)$ and $(5, 3)$, the line segment joining $(5, 3)$ and $(8, 2)$, and the line parallel to the K axis passing through the point $(8, 2)$, i.e.,

$$\{(K, L) : \ K \geq 2.5, \ L \geq -(7/2.5)K + 17, \ L \geq -(1/3)K + 14/3, \ L \geq 2\}.$$

(d) The point (9, 7.5) lies on the line given by $L = (5/6)K$, which intersects the line $L = -1.5K + 14$ passing through the points (4,8) and (8, 2). The point of intersection is (6, 5), which, by construction, has output rate of 20. Now $(6, 5) = 2/3(9, 7.5)$ and so the CRS input efficiency is 2/3. For the CRS model of technology, the output and input efficiencies are equal, and so the output efficiency is also 2/3.

4.4 Use (4.5) and the input free disposability of $L_{\Phi}(1)$ to argue that $CO(\{a^k\}) + \mathbb{R}^n_+ \subset L_{\Phi}(1)$. To show the reverse inclusion, pick an x in $L_{\Phi}(1)$. Use (4.5) to find a vector u for which $x \geq Au$ and $\tilde{u} := \sum_k u^k \geq 1$. Let λ denote the vector in $\mathbb{R}^n_+$ whose coordinates are given by $\lambda^k := u^k / \sum_k u^k$, $1 \leq k \leq n$. Write x as $A\lambda + (x - A\lambda)$ and argue that the vector $x - A\lambda$ is nonnegative.

4.5 Let C denote an arbitrary convex cone containing $\mathcal{T}^{VRS}$. It is sufficient to show that $\mathcal{T}^{CRS} \subset C$. To this end, pick $(x(\lambda), y(\lambda)) \in \mathcal{T}^{CRS}$ and let $s := \sum_k \lambda_k$. The case $s = 0$ is trivial, so assume $s > 0$. Let $\mu_i := \lambda_i / \sum_k \lambda_k$ for each i. Show that $(x(\lambda), y(\lambda)) = s(x(\mu), y(\mu))$.

4.6 The input and output efficiency measures of ${T_F}^{VRS}$ and ${T_F}^{CRS}$ are *higher* than their counterparts for T^{VRS} and T^{CRS}.

4.7 (a) Suppose $x_1 \leq x_2$ with each $x_i \in \mathcal{X}$. Any feasible choice λ for the linear program defined by $\Phi^*(x_1)$ is also feasible for the linear program defined by $\Phi^*(x_2)$, which immediately implies that $\Phi^*(x_1) \leq \Phi^*(x_2)$. This establishes that $\Phi^*(\cdot)$ is nondecreasing. It remains to prove that $\Phi^*(\cdot)$ is concave. To this end, pick $x_1, x_2 \in \mathcal{X}$ and $\mu \in [0, 1]$. Let λ_i^*, $i = 1, 2$, denote optimal solutions for the linear programs defined by $\Phi^*(x_i)$, respectively. The vector $\mu\lambda_1^* + (1-\mu)\lambda_2^*$ is feasible for the linear program defined by $\Phi^*(\mu x_1 + (1-\mu)x_2)$. This implies that

$$\begin{aligned} \Phi^*(\mu x_1 + (1-\mu)x_2) &\geq (\mu\lambda_1^* + (1-\mu)\lambda_2^*) \cdot y \\ &= \mu(\lambda_1^* \cdot y) + (1-\mu)(\lambda_2^* \cdot y) \\ &= \mu\Phi^*(x_1) + (1-\mu)\Phi^*(x_2), \end{aligned}$$

which established concavity.

(b) Pick a representation $\Phi(\cdot)$ of $\mathcal{D}$ and pick $x \in \mathcal{X}$. For each feasible choice λ for the linear program that defines $\Phi^*(x)$, we have that $x \geq \sum_i \lambda_i x_i$, which implies that $\Phi(x) \geq \Phi(\sum_i \lambda_i x_i)$ since $\Phi(\cdot)$ is nondecreasing. Since $\Phi(\cdot)$ is also concave and $\Phi(x_i) = y_i$, we have that

$$\Phi(x) \geq \Phi(\sum_i \lambda_i x_i) \geq \sum_i \lambda_i \Phi(x_i) = \sum_i \lambda_i y_i.$$

It now follows by the definition of $\Phi^*(x)$ that $\Phi(x) \geq \Phi^*(x)$. Finally, we must show that $\Phi^*(\cdot)$ is indeed a representation, which boils down to showing that $\Phi^*(x_i) = y_i$ since we established in (a) that $\Phi^* \in \mathcal{F}$. Pick an index i, $1 \leq i \leq N$. The vector $\lambda \in \mathbb{R}_+^N$ such that $\lambda_i = 1$ and $\lambda_j = 0$ for $j \neq i$ is feasible for the linear program defining $\Phi^*(x_i)$. Thus, $\Phi^*(x_i) \geq y_i$. On the other hand, we are given that $\Phi(x_i) = y_i$ and we have already established that $\Phi(x) \geq \Phi^*(x)$ for any $x \in \mathcal{X}$. It follows then that $\Phi^*(x_i) = y_i$, as required. (Here is where we usr the assumption that $\mathcal{D}$ is concave-representable so that such a $\Phi(\cdot)$ exists.)

(c) Suppose $\Phi^*(x_i) = y_i$ for each i. By (a) we know that $\Phi^*(\cdot) \in \mathcal{F}$, which means that $\Phi^*(\cdot)$ is itself a representation. The converse was shown in (b).

(d) Let $\mathcal{D} = \{(1,2),(2,3)\}$. Here, $\mathcal{X} = [1,2]$. We have established that $\Phi^*(\cdot)$ is the smallest concave nondecreasing function passing through these two points. Obviously $\Phi^*(x) = x + 1$ on $\mathcal{X}$. The function

$$\Phi(x) = \begin{cases} 2x, & 1 \leq x \leq 1.5, \\ 3, & 1.5 \leq x \leq 2, \end{cases}$$

is nondecreasing and concave and also passes through these two points; however, $\Phi(x) > \Phi^*(x)$ on $(1,2)$.

5

Cost Function

A cost function represents the *minimum* cost required to achieve a predetermined level of output given the prices for the factors of production. It is an essential tool of applied production analysis. A well-behaved technology can be reconstructed by observing the cost minimizing behavior of producers. We shall use this fundamental *duality* between the cost and production functions in Part II.

5.1 Definition

Definition 5.1. *The* **cost function** *is*

$$Q(y,p) := \min\{p \cdot x : \ x \in L(y)\}. \tag{5.1}$$

It is understood that the cost function is defined only for those output vectors that are attainable; that is, the input possibility set must be nonempty.

Technically, the minimum in Definition 5.1 should be replaced with an infimum, unless one proves it is always possible to find an $x \in L(u)$ that achieves the minimal cost. For a well-behaved technology, this will be the case. See the Appendix to this chapter for a proof.

5.2 Properties

5.2.1 Geometry

Since $Q(y,p)$ represents minimum cost, it follows that $p \cdot z \geq Q(y,p)$ for each $z \in L(y)$, or

$$L(y) \subset \{z : p \cdot z \geq Q(y,p)\}. \tag{5.2}$$

This means that the input possibility set $L(y)$ lies within the closed halfspace "lying on or above" the hyperplane

$$H(p, Q(y,p)) := \{z : p \cdot z = Q(y,p)\}.$$

Since there is a cost minimizer x^* for which $p \cdot x^* = Q(y,p)$, the hyperplane $H(p, Q(y,p))$ *supports* $L(y)$.

In the two-dimensional input case with both input prices positive, the set of points lying on the hyperplane $\{z : p \cdot z = Q\}$ is equivalent to defining a line

$$\{(K,L) : L = -(p_K/p_L)K + Q/p_L\} \tag{5.3}$$

with slope $-(p_K/p_L)$ and intercept Q/p_L. This is known as an **isocost line**. The slope of the line does *not* change as the value of Q changes, and so the lines induced by different values of Q are all *parallel*. From a geometrical perspective, to obtain a cost minimizer and minimum cost given prices p and input possibility set $L(y)$, all one has to do is to plot a line with slope $-(p_K/p_L)$ and adjust the intercept so that the line is *tangent* to the isoquant $ISOQ(y)$.

Example 5.2. In Figure 5.1, an isocost line associated with the minimal cost is tangent to the isoquant at the point $(3, 2)$. Two isocost lines associated with Q-values lower than the minimal cost are also displayed.

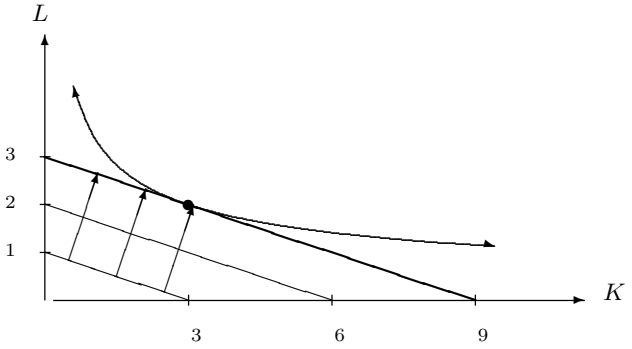


Fig. 5.1. Determining minimal cost.

In the differentiable setting, the rate of technical substitution between K and L at a cost minimizer x^* defines the slope of a line tangent to the isoquant $ISOQ(y)$ at x^*. (See Figure 2.3 on p. 22.) Since there is only one tangent line, it follows that the rate of technical substitution equals minus the ratio of the input prices. Since the rate of technical substitution is minus the ratio of the partial derivatives of the production function, the *price vector is proportional to the gradient vector evaluated at a cost minimizer*. This fact extends to the

general case of $n > 2$ inputs. We shall verify this fact when we solve for $Q(y, p)$.

5.2.2 Homogeneity

A function $f : \mathbb{R}^k \longrightarrow \mathbb{R}$ is linearly homogeneous if $f(s \cdot x) = sf(x)$ for any positive scalar s. It is obvious that $Q(y, p)$ is linearly homogeneous in p. Scaling a price vector merely scales the minimum cost, which is a necessary property of any well-behaved cost function!

5.2.3 Concavity

In addition to being linearly homogeneous, the cost function is also concave.

Proposition 5.3. *$Q(y, p)$ is concave in p for fixed y.*

Proof. Pick $p_i \in \mathbb{R}^n_+$, $i = 1, 2$ and $\lambda \in [0, 1]$. By definition of the cost function,

$$\begin{aligned} Q(y, \lambda p_1 + (1-\lambda)p_2) &= \min\{(\lambda p_1 + (1-\lambda)p_2) \cdot x : \ x \in L(y)\} \\ &= \min\{\lambda(p_1 \cdot x) + (1-\lambda)(p_2 \cdot x) : \ x \in L(y)\} \\ &\geq \min\{\lambda(p_1 \cdot x) : \ x \in L(y)\} + \min\{(1-\lambda)(p_2 \cdot x) : \ x \in L(y)\} \\ &= \lambda Q(y, p_1) + (1-\lambda)Q(y, p_2). \ \square \end{aligned} \tag{5.4}$$

Concavity of the cost function does *not* require the input possibility set to be convex. (The proof above does not use this property.)

Remark 5.4. Concavity of $Q(y, \cdot)$ follows from the general fact that a minimum of concave functions is always concave.

5.3 Example: Cobb-Douglas Technology

Since a Cobb-Douglas function is differentiable, we will determine its cost function by using the first-order optimality conditions. We motivate this derivation by first analyzing a two-input, single-output example.

Example 5.5. Consider the production function $\Phi(K, L) = KL^2$. Suppose the cost per unit of capital is 4 and the cost per unit of labor is 12. The desired output rate is 12. We seek to minimize cost. We have

$$\begin{aligned} Q(u, p) &= \min\{p_k K + p_L L : \ KL^2 \geq u\} \\ &= \min_{L>0}\{\ \theta(L) := p_k(u/L^2) + p_L L\}. \end{aligned} \tag{5.5}$$

At the cost minimizer, $(K(u, p), L(u, p)) = (K^*, L^*)$, the derivative of $\theta(\cdot)$ at L^* must vanish; thus,

$$L^* = L(u,p) = (2p_K/p_L)^{1/3}u^{1/3} = 2. \tag{5.6}$$

Since $\Phi(K^*, L^*) = u$,

$$K^* = K(u,p) = (p_L/2p_K)^{2/3}u^{1/3} = 3. \tag{5.7}$$

Substituting the general solutions for capital and labor, the closed-form expression for minimum cost is

$$\begin{aligned} Q(u,p) &= p_K K(u,p) + p_L L(u,p) \\ &= \left\{[(1/2)^{2/3} + 2^{1/3}]p_K^{1/3}p_L^{2/3}\right\}\ u^{1/3} \\ &= Q(1,p)\ u^{1/3}. \end{aligned} \tag{5.8}$$

The minimum cost is 36.

Remark 5.6. The cost function given in (5.8) is homogeneous of degree one in p, as expected, and is also *multiplicatively separable* in prices and output; that is, it factors into a product of a function solely of prices and a function solely of output. Technologies whose cost functions exhibit such a decomposition are called *homothetic*, and we shall study them later. Note also the unit cost function $Q(1,p)$ has a Cobb-Douglas form.

To solve for $Q(u,p)$, we may also use the method of *Lagrange multipliers*, and transform the constrained problem into an unconstrained one. The *first-order optimality conditions* (see E.7, p. 481) ensure that if x^* is a cost minimizer whose coordinates are all strictly positive and if the constraint is tight at the optimum (which it will be), then there must be a positive scalar λ^* for which the partial derivatives of the *Lagrangian* function (see E.10, p. 483)

$$L((K,L),\lambda^*) = (p_K K + p_L L) - \lambda^*(KL^2 - u) \tag{5.9}$$

with respect to the factor inputs K and L must vanish when evaluated at the optimum (K^*, L^*). Thus, $p_K = \lambda L^2$ and $p_L = \lambda 2KL$, from which it follows that the ratio $p_k/p_L = 0.5L/K$ or $K = 1.5L$. Using the fact that $\Phi(K,L) = u$ we obtain the same solution as before.

The cost function for a general Cobb-Douglas technology admits a very useful characterization in terms of the cost or expenditure shares.

Definition 5.7. *Let $x(y,p)$ denote an optimum choice of inputs to minimize the cost function $Q(y,p)$. The* **cost or expenditure share of factor input i** *(with respect to $x(y,p)$) is the ratio*

$$S_i(y,p) := p_i x_i(y,p)/Q(y,p).$$

When the cost minimal input vector is unique, we write S_i in lieu of $S_i(x(y,p))$. The cost shares always sum to one, i.e.,

$$\sum_i S_i(x(y,p)) = 1.$$

For a general Cobb-Douglas technology,

$$Q(u,p) = \min\{\, p \cdot x = \sum_i p_i x_i :\ A \prod_i x_i^{\alpha_i} \geq u\}.$$

The Lagrangian function associated with the minimum cost problem is

$$L(x,\lambda) = p \cdot x - \lambda(\Phi(x) - u), \tag{5.10}$$

and the first-order optimality conditions are[1]

$$0 = \nabla_x L(x,\lambda) = p - \lambda \nabla \Phi(x). \tag{5.11}$$

The first-order optimality conditions imply that

$$p_i = \lambda A \alpha_i x_i^{\alpha_i - 1} \prod_{j \neq i} x_j^{\alpha_j}. \tag{5.12}$$

Multiplying both sides of (5.12) by x_i,

$$p_i x_i = \lambda u \alpha_i, \tag{5.13}$$

since the constraint must be tight at the optimum. It is immediate from (5.13) that $S_i/S_j = \alpha_i/\alpha_j$. Given that $\sum_i S_i = 1$, we arrive at a simple, yet fundamental result concerning cost minimization of a Cobb-Douglas technology, namely, *the expenditure shares are fixed a priori* as

$$S_i = \frac{\alpha_i}{\sum_k \alpha_k}\ . \tag{5.14}$$

Since p_i is known, knowledge of $Q(u,p)$ immediately renders

$$x_i(u,p) = \frac{Q(u,p) S_i}{p_i}.$$

It remains to find $Q(u,p)$.

To this end, sum both sides of (5.13) over i to obtain

$$Q(u,p) = \lambda u \sum_i \alpha_i. \tag{5.15}$$

Keep in mind that $\lambda = \lambda(u,p)$ is a function of both u and p. Now we must find $\lambda(u,p)$.

From (5.13) and the identity

$$A \prod_i (p_i x_i)^{\alpha_i} = \Big[\prod_i p_i^{\alpha_i} \Big]\ u,$$

[1] We shall drop the * on λ when solving these equations.

we have

$$A\lambda^{\sum_i \alpha_i}\left(\prod_i \alpha_i^{\alpha_i}\right) u^{\sum_i \alpha_i} = \left[\prod_i p_i^{\alpha_i}\right] u.$$

Thus,

$$\lambda(u,p) = \left[\frac{\prod_i (p_i/\alpha_i)^{\alpha_i/\sum_i a_i}}{A^{1/\sum_i a_i}}\right] u^{1/\sum_i \alpha_i - 1}. \tag{5.16}$$

Using (5.15), we finally arrive at

$$Q(u,p) = \left(\sum_i \alpha_i\right)\left[\frac{\prod_i (p_i/\alpha_i)^{\alpha_i/\sum_i \alpha_i}}{A^{1/\sum_i \alpha_i}}\right] u^{1/\sum_i \alpha_i}. \tag{5.17}$$

An observation is in order. An examination of (5.11) shows that the gradient vector of the production function at the cost minimizer must be *proportional* to the price vector. Equivalently, the ratio of the prices equals the ratio of the partial derivatives. For the two-dimensional example, it can be verified that $\lambda^* = 1$, and so $\nabla\Phi(K,L) = (L^2, 2KL)$ at $(K^*, L^*) = (3,2)$ exactly coincides with the price vector.

5.4 Sensitivity Analysis

We continue with the single-output setting and examine two types of sensitivities:

- How does the minimum cost vary with output?
- How does the minimum cost vary with factor prices?

For the developments to follow, we recall the concept of elasticity. Let $f(\cdot)$ be a real-valued differentiable function of scalar x. The elasticity of $f(\cdot)$ with respect to x measures the percentage change in $f(\cdot)$ for a 1% percentage change in x, and is

$$\epsilon_f(x) = \frac{x}{f(x)}\, f'(x). \tag{5.18}$$

When $f(x) = x^r$ for some non-zero r, the elasticity of $f(\cdot)$ with respect to x is easily seen to be r, independent of x.

5.4.1 Sensitivity to Output

The *elasticity of cost with respect to output* for the cost function given in (5.8) is 1/3, which implies that the $\partial Q/\partial u = 1$, since here $u = 12$ and $Q(u,p) = 36$.

Computing the elasticity of output of the cost function boils down to computing the partial derivative $\partial Q/\partial u$. In this special case, it is quite easy. In fact, it turns out that no calculation (other than what was already performed)

is required! The partial derivative here is 1, which just happens to equal the optimal Lagrange multiplier λ^*. This is definitely not a coincidence.

In what follows, we assume there is a *unique* cost minimizer $x(u,p)$, which is differentiable. (There are a number of conditions that will ensure this is true.) Thus,

$$Q(u,p) = p \cdot x(u,p)$$

and

$$\frac{\partial Q(u,p)}{\partial u} = \sum_i p_i \frac{\partial x_i(u,p)}{\partial u}. \tag{5.19}$$

Since

$$\Phi(x(u,p)) = u, \tag{5.20}$$

we may differentiate both sides of the equality with respect to u to obtain that

$$\sum_i \frac{\partial \Phi}{\partial x_i} \frac{\partial x_i}{\partial u} = 1. \tag{5.21}$$

From our previous discussion, we know that

$$p = \lambda^* \nabla \Phi(x^*). \tag{5.22}$$

Substituting (5.22) into (5.19) and using (5.21) yields the conjectured result, namely,

$$\frac{\partial Q}{\partial u} = \lambda^*. \tag{5.23}$$

Remark 5.8. An inspection of (5.17) and (5.16) shows that (5.23) holds for the general Cobb-Douglas setting, as it should.

5.4.2 Sensitivity to Price: Shephard's Lemma

Consider the sensitivity of minimal cost to price. Suppose the price of a unit of capital changes from 4 to $4 + \Delta\pi$. By approximately how much will the minimum cost change when $\Delta\pi$ is small? The answer is

$$\frac{\partial Q(u,p)}{\partial p_K} \cdot \Delta\pi,$$

and so we could directly differentiate the expression in (5.8). Before we undertake that exercise, an examination of (5.8) shows that the elasticity of the cost function with respect to the price of capital, namely

$$\frac{p_K}{Q(u,p)} \frac{\partial Q}{\partial p_K}, \tag{5.24}$$

is 1/3, which immediately gives us the value of the partial derivative to be 3, since $p_K = 4$ and $Q(u,p) = 36$. The number 3 just happens to coincide

with the optimum amount of capital, K^*. Once again, this is definitely not a coincidence.

Since

$$\frac{\partial Q(u,p)}{\partial p_k} = \sum_i p_i \frac{\partial x_i(u,p)}{\partial p_k} + x_k(u,p), \tag{5.25}$$

we may differentiate both sides of (5.20) this time with respect to p_k to obtain that

$$\sum_i \frac{\partial \Phi}{\partial x_i} \frac{\partial x_i}{\partial p_k} = 0. \tag{5.26}$$

Substituting (5.22) into (5.25) and using (5.26) yields the conjectured result, namely

$$\frac{\partial Q}{\partial p_k} = x_k(u,p). \tag{5.27}$$

This famous result is known as **Shephard's Lemma**, and is one of the cornerstone results in applied production analysis.

5.5 Nonparametric Estimation

5.5.1 Leontief Technologies

Consider first a simple Leontief technology. It is clear from Figure 4.1, p. 55, that choosing $x = ua$ will achieve minimum cost for any price vector since there is no benefit to adding slack to the inputs. Thus,

$$Q(u,p) = u(p \cdot a) = u\, Q(1,p). \tag{5.28}$$

Consider next a general Leontief technology. For any choice of activity intensities $u = (u^1, u^2, \ldots, u^N)$ choosing $x = Au$ will achieve minimum cost for any price vector. (Once again, there is no benefit to adding slack to the inputs.) The activity intensities, however, must satisfy the constraint $\sum_k u^k = u$. Consequently, the problem of finding the minimum cost can be reformulated as

$$\min\{p \cdot (Au) : \sum_k u^k = u\}. \tag{5.29}$$

Let $Q^k(u,p)$ denote the cost function for the k^{th} simple Leontief process. Since

$$p^T A = (Q^1(1,p), Q^2(1,p), \ldots, Q^N(1,p)), \tag{5.30}$$

the minimum cost problem is equivalent to

$$\min\left\{\sum_k Q^k(1,p)u^k : \sum_k u^k = u\right\}, \tag{5.31}$$

whose solution is easily seen to be

$$Q(u,p) = u \min_k Q^k(1,p). \tag{5.32}$$

Equation (5.32) should not be surprising. Since the general Leontief technology exhibits constant returns-to-scale, the cost at any positive level u is simply u times the cost of achieving output rate one. Since the input possibility set corresponding to output rate one is the convex, input free disposable hull of the fixed coefficient vectors a^k, the minimum cost to achieve output rate one corresponds to the activity whose cost (at these prices) is minimum. (If there are ties, then any weighted combination of such activities will achieve minimum cost.) From a geometrical perspective, any line (hyperplane) that supports a piecewise linear isoquant must contain one of the vertices that define the isoquant.

5.5.2 *HR* Technology

The input possibility set $L^{HR}(u)$ is the smallest closed, convex, input free disposable set containing those input vectors that achieve at least output rate u. Consequently, the minimum cost of achieving output rate u corresponds to the cost of the activity that achieves output rate u at minimum cost. (Once again, if there are ties, then any weighted combination of such activities will achieve minimum cost.)

5.5.3 *CRS* and *VRS* Technologies

The input possibility sets for each technology are each characterized by a set of linear constraints; accordingly, it is possible to formulate a linear program to obtain the minimum cost for a particular choice of u and p. Linear programming is unnecessary in the scalar output case—the cost function for these two technologies can be *graphically* computed.

To this end, we first define the concept of the output-cost set.

Definition 5.9. *For each* $p \geq 0$ *the* **output-cost set** *is*

$$\mathcal{OC}^{\mathcal{T}}(p) := \{(u, p \cdot x) : \ (x,u) \in \mathcal{T}\}. \tag{5.33}$$

An output-cost set represents the collection of all output-cost pairs that are technologically feasible when cost is measured at prices p. If $\mathcal{T}$ is a well-behaved convex technology, then it is straightforward to show that each output-cost set will be a convex subset of the output-cost (u,c)-space in $\mathbb{R}^2_+$. Moreover, each output-cost set will exhibit free disposability—in this context, this means that if $(u,c) \in \mathcal{OC}^{\mathcal{T}}(p)$ and if $c' \geq c$ and $u' \leq u$, then $(u',c') \in \mathcal{OC}^{\mathcal{T}}(p)$, too.

With this definition in hand, let $\mathcal{T}^{CRS}$, $\mathcal{OC}^{CRS}(p)$ and $\mathcal{T}^{VRS}$, $OC^{VRS}(p)$ denote the technology and output-cost sets corresponding to the CRS and VRS technologies, respectively. The output-cost sets corresponding to the

Table 5.1. Input, output and cost data. Price of each input is 0.10.

Firm	x_1	x_2	u	$p \cdot x$	$(p \cdot x)/u$
1	0.5	1.5	1.0	0.2	0.200
2	3.0	1.0	1.6	0.4	0.250
3	1.2	1.8	2.0	0.3	0.150
4	7.0	1.0	3.0	0.8	0.267
5	2.0	3.0	4.0	0.5	0.125
6	8.5	6.5	5.0	1.5	0.300
7	12.0	8.0	6.5	2.0	0.308
8	11.0	11.0	8.0	2.2	0.275

VRS and CRS technology sets are convex, since each of these technologies is convex.

Example 5.10. Consider the data shown in Table 5.1. The price vector $p = (0.10, 0.10)$, and so the cost of each input vector is 10% of the sum of the two inputs. The output-cost set for the VRS technology is depicted in Figure 5.2. The piecewise linear, convex "lower boundary" defines the *Efficient Frontier* in the output-cost space. The output-cost set for the CRS technology is depicted in Figure 5.3. Its Efficient Frontier is defined by the ray that begins at the origin and passes through the point $(4.0, 0.50)$, which has the *lowest cost-to-output ratio.* The definitions of $Q(u, p)$ and the output-cost set $\mathcal{OC}^{\mathcal{T}}(p)$ imply that this Efficient Frontier is the *graph* of the cost function! For example, consider the minimum cost to achieve output rate 3. Since the line $u = 3.0$ intersects the output-cost set at the midpoint of the line segment joining points $(2.0, 0.3)$ and $(4.0, 5)$, the minimum cost to achieve output rate 3.0 for the VRS technology is therefore 0.40. Since the boundary of the output-cost set for the CRS technology is determined by the line $c = 0.125u$, the line $u = 3.0$ intersects it at the point $(3, 0.375)$. Consequently, the minimum cost to achieve output rate 3.0 for the CRS technology is 0.375. As for the HR technology, the minimum cost to achieve output rate 3.0 is 0.5.

5.6 Reconstructing the Technology

For a well-behaved technology, the hyperplane $\{z : p \cdot z = Q(y, p)\}$ supports $L(y)$ at a boundary point. Each point that lies on the boundary of $L(y)$ necessarily has a price vector p that supports it. These two statements together imply that the cost function characterizes the boundary of each input possibility set. Since the boundary of each input possibility set is sufficient information to generate the input possibility set itself, *the cost function contains enough information to reconstruct the underlying technology from which it is derived.*

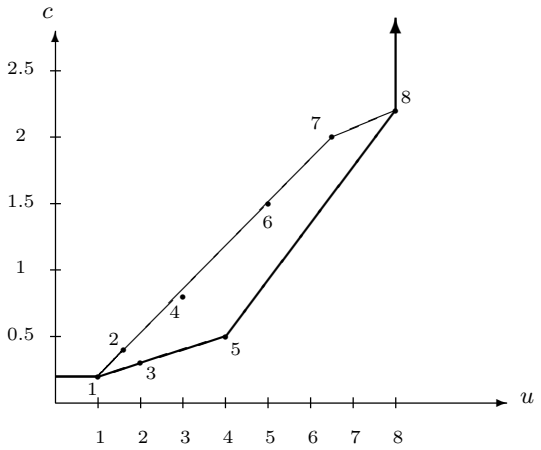


Fig. 5.2. Output-cost set for the VRS technology for the data given in Table 5.1.

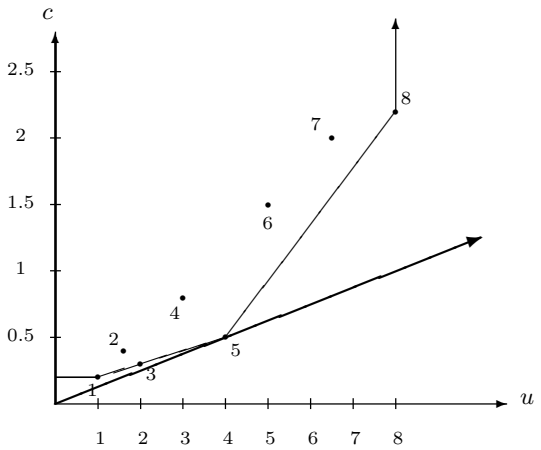


Fig. 5.3. Output-cost set for the CRS technology for the data given in Table 5.1.

Let $Q(y, p)$ denote a cost function derived from some well-behaved technology whose input possibility sets are given by $L(y)$. Define

$$L^*(y) = \bigcap_{p \geq 0} \{z : \ p \cdot z \geq Q(y, p)\}. \tag{5.34}$$

Theorem 5.11. $L^*(y) = L(y)$.

Proof. First, it follows from (5.34) that $L^*(y)$ is a closed, convex, input free disposable set that contains $L(y)$. To show the reverse inclusion, pick an $\bar{x} \notin L(y)$. The strict separation theorem (see Corollary C.11, p. 463) guarantees existence of a price vector $\bar{p} \geq 0$ for which

$$\bar{p} \cdot \bar{x} < \alpha < \bar{p} \cdot z \text{ for all } z \in L(y),$$

which immediately implies that $\bar{x} \notin L^*(y)$. Thus, the claim is established. □

The essence of (5.34) is that the cost function can be used to reconstruct each input possibility set, the family of which defines the technology.

Here are two applications in the scalar output case.

5.6.1 Outer Approximation of Technology

Suppose input, output and price data

$$\{(x_1, u_1, p_1), (x_2, u_2, p_2), \ldots, (x_N, u_N, p_N)\}$$

are given for N firms. Under assumption of cost-minimization on the part of each firm, we have that $p_i \cdot x_i = Q(u_i, p_i)$, from which it follows that

$$L(u_i) \subset \{z : \ p_i \cdot z \geq p_i \cdot x_i\}.$$

Let $J(u) := \{i : u_i \leq u\}$. Define

$$L^O(u) = \bigcap_{i \in J(u)} \{z : \ p_i \cdot z \geq p_i \cdot x_i\}.$$

Clearly, $L^O(u)$ is a closed, convex and input free disposable set. Moreover,

$$L(u) \subset L^O(u)$$

due to the nestedness of the input possibility sets. In fact, $L^O(u)$ is the *largest closed, convex, input free disposable set containing* $L(u)$ *consistent with the data and the assumption of cost minimization.* It is often referred to as an **outer approximation** to the true input possibility set; the symbol 'O' here refers to outer. Recall that the input possibility set of an HR technology is

$$L^{HR}(u) = \mathcal{IFDH}(Conv(\mathcal{X}(u))),$$

where

$$\mathcal{X}(u) = \{x_j\}_{j \in \mathcal{I}(u)} \text{ and } \mathcal{I}(u) = \{i : u_i \geq u\}.$$

$L^{HR}(u)$ is the smallest closed, convex, input free disposable set consistent with the data, and represents an **inner approximation** to the true technology. Thus,

$$L^{HR}(u) \subset L(u) \subset L^O(u).$$

See Figure 5.4.

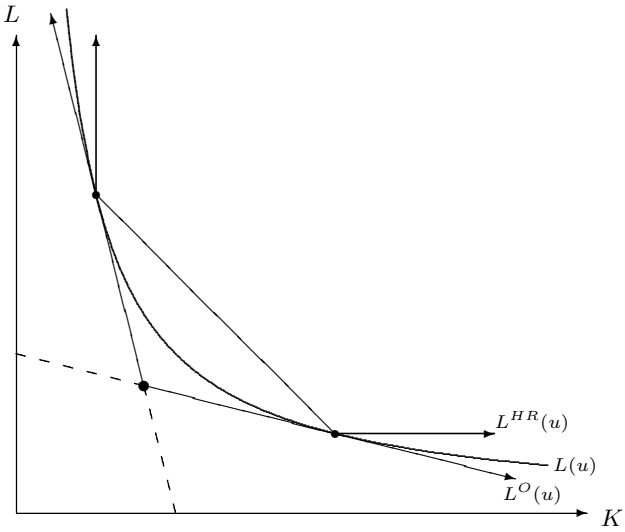


Fig. 5.4. $L(u)$ represents the true input possibility set. Inner approximation given by $L^{HR}(u)$. Outer approximation given by $L^{O}(u)$.

5.6.2 Cost and Production

Consider a given vector $x_0 > 0$ and a target level of output u_0. Can x_0 achieve u_0, i.e., is $x_0 \in L(u_0)$? If the cost function and price vector p are known, we know that $p \cdot x \geq Q(u, p)$ for all $x \in L(u)$. However, it is possible that $x_0 \notin L(u)$ but $p \cdot x_0 \geq Q(u, p)$, so checking the value of cost of x_0 at existing prices, $p \cdot x_0$, will not provide a definitive answer.

It certainly would be reasonable to hope that if $x_0 \notin L(u)$, there is *some* price system, p_0, under which the cost of x_0, $p_0 \cdot x_0$, would be strictly less than the minimum cost of achieving u_0 at p_0, $Q(u_0, p_0)$. (The converse is immediately true by definition of minimal cost.) For a well-behaved technology, the hope is realized. If $x_0 \notin L(u_0)$, then the strict separation theorem (see Corollary C.11, p. 463) guarantees existence of a hyperplane that strictly separates x_0 from $L(u_0)$; thus, a p exists for which

$$p \cdot x_0 < \alpha < p \cdot x \text{ for all } x \in L(u). \tag{5.35}$$

The coordinates of p must all be nonnegative since $L(u)$ exhibits input free disposability, and clearly α is positive. Thus, we have shown the existence of a price vector $p_0 = p/\alpha$ for which

$$Q(u_0, p_0) \geq 1 \text{ and } p_0 \cdot x_0 < 1. \tag{5.36}$$

To make matters concrete, assume $L(u_0)$ is approximated via the HR technology, and let $x_1, x_2, \ldots, x_N$ denote the observed input data (each of which can be used to obtain at least u_0). Let

$$\mathcal{P} := \{p \geq 0 : \ p \cdot x_i \geq 1 \text{ for all } i = 1, 2, \ldots, N\}. \tag{5.37}$$

$\mathcal{P}$ is defined by a finite set of linear inequalities. Clearly, if $Q(u_0, p) \geq 1$, then $p \in \mathcal{P}$. Conversely, suppose $p \in \mathcal{P}$. By definition of the HR technology, each $z \in L(u_0)$ must be at least as large as some convex combination of the x_i's, i.e.,

$$z \geq \sum_i \lambda_i x_i \text{ for some } \lambda \in \mathbb{R}^N_+. \tag{5.38}$$

Since the cost of each x_i at p is at least one, the cost of z,

$$p \cdot z \geq p \cdot \left(\sum_i \lambda_i x_i \right) = \sum_i \lambda_i (p \cdot x_i), \tag{5.39}$$

must be at least one, too. We have just shown that $p \in \mathcal{P}$ if and only if $Q(u_0, p) \geq 1$. Since we have previously argued that $x_0 \notin L(u_0)$ if and only if there exist a p_0 for which (5.36) holds, it follows that $x_0 \notin L(u_0)$ *if and only if the optimal value to the linear program*

$$\min\{p \cdot x_0 : \ p \in \mathcal{P}\} \tag{5.40}$$

is strictly less than one! We will return to this example when we discuss distance functions.

5.7 Homothetic Technologies

The cost function $Q(u, p)$ associated with a homothetic production function (see Definition 2.28, p. 29) has an important structural property. It *factors* as a product $f(u)P(p)$, where $f(\cdot)$ is a transform and $P(\cdot)$ is homogeneous of degree one. For a homothetic technology, if the prices remain constant, then the economies of scale are determined by the transform $f(\cdot)$.

Proposition 5.12. *If the production function $\Phi(\cdot)$ is homothetic, then $Q(u, p) = f(u)P(p)$, where $f(\cdot)$ is a transform and $P(\cdot)$ is homogeneous of degree one.*

Proof. Let $\Phi(x) = F(\phi(x))$ be a homothetic production function. By definition of minimum cost and the fact that $\phi(\cdot)$ is homogeneous of degree one,

$$\begin{aligned} Q(u, p) &= \min\{p \cdot x : \ F(\phi(x)) \geq u\} \\ &= \min\{p \cdot x : \ \phi(x) \geq f(u)\} \\ &= f(u) \left(\min \left\{ p \cdot \left(\frac{x}{f(u)} \right) : \ \phi\left(\frac{x}{f(u)} \right) \geq 1 \right\} \right) \\ &= f(u) \ \min\{p \cdot z : \ \phi(z) \geq 1\} \\ &:= f(u)P(p). \end{aligned}$$

Clearly, $P(\cdot)$ is homogeneous of degree one. □

Remark 5.13. $P(p)$ is the minimal cost of achieving at least one unit of output at prices p.

Remark 5.14. Proposition 5.12 shows that the factorability of the cost function is a *necessary* condition for a homothetic technology. We shall subsequently show this factorabilty condition is *sufficient* to imply the property of homotheticity. Thus, homotheticity is *equivalent* to the factorability of the cost function.

5.8 Appendix

We prove that the cost function is well-defined. Pick a nonzero $p \in \mathbb{R}^n_+$ and a $y \in \mathbb{R}^m_+$ such that $L(y)$ is not empty. We begin by establishing the following Lemma.

Lemma 5.15. $L(y) = \mathcal{IFDH}(cl(\mathcal{EF}(y)))$.[2]

Proof. Pick an $x \in L(y)$ and define

$$F := L(y) \cap \{z \in \mathbb{R}^n_+ : \ z \leq x\}.$$

Let z^* denote a point in F that minimizes the distance to the origin. Such a point exists since $L(y)$ is closed by Axiom **A4** and hence F is compact. The point z^* belongs to the Efficient Frontier $\mathcal{EF}(y)$ of $L(y)$; otherwise, it would be possible to find a point in F closer to the origin. Obviously, $x \geq z^*$ and $x = z^* + (x - z^*)$, which shows that $L(y) \subset \mathcal{IFDH}(cl(\mathcal{EF}(y)))$. The reverse inclusion $\mathcal{IFDH}(cl(\mathcal{EF}(y))) \subset L(y)$ is an immediate consequence of $\mathcal{EF}(y) \subset L(y)$, the closure of $L(y)$, and Axiom **A3**, the input free disposability of $L(y)$. □

Define the cost function as

$$Q(y, p) := \inf\{f(x) : x \in S\},$$

where $f(x) := p \cdot x$ and $S := L(y)$. The linear function $f(\cdot)$ is continuous. If S were closed and bounded (and hence compact), then the infimum exists can be replaced by a minimum. The input possibility set $L(y)$ is closed by Axiom **A4**, but clearly not bounded. The idea, however, is to show that it is possible to restrict the domain of the minimization problem to a compact subset S of $L(y)$. When $p > 0$, this is easy: simply pick any $z \in L(y)$ and take S as $L(y) \cap T$, where $T := \{x \in \mathbb{R}^n_+ : \ f(x) \leq f(z)\}$. (The set T is closed and bounded since $f(\cdot)$ is continuous.) If any of the components of p are zero, however, the set $L(y) \cap T$ will still be unbounded. This is where Axiom **A5**,

[2] Recall that $\mathcal{EF}(y)$ is the Efficient Frontier of the input possibility set and that $\mathcal{IFDH}(L)$ is the input free disposable hull of L. See Chapter 3.

boundedness of the Efficient Frontier, comes to the rescue. Simply take S to be $cl(\mathcal{EF}(y))$, the closure of the efficient subset $\mathcal{EF}(y)$. Obviously, S will be closed and bounded. To complete the proof, one must show that the minimum cost defined over S is in fact the true minimum. But this follows immediately from Lemma 5.15, since for the purpose of minimizing cost there is no need to consider input vectors "larger" than those in $cl(\mathcal{EF}(y))$. □

5.9 Exercises

5.1. Consider the production function $\Phi(K, L) = K^{0.2}L^{0.4}$. The price vector $p = (p_K, p_L) = (1, 2)$ and desired output rate is 8.

(a) Determine the minimal cost $Q(u, p)$ and the cost minimum input vector.
(b) Determine $\partial Q(u, p)/\partial u$ and use it to estimate the increase in cost when the desired output rate is increased to 8.08.
(c) Determine the elasticity of cost with respect to output.
(d) Determine the exact new minimum cost when desired output is increased by 1%. Compute the percentage increase and compare with your answer to part (c).
(e) Determine $\partial Q(u, p)/\partial p_K$ (with the original numbers) and use it to estimate the increase in cost when p_K increases to 1.01. Compute the exact minimum cost when p_K increases to 1.01 and compare your answers.

5.2. For a Cobb-Douglas production function of two input factors, show how to use (5.14) to easily solve for the minimum cost input vector *without* directly using the first-order optimality conditions. Solve problem 5.1(a) using this technique.

5.3. For the production function $\Phi(K, L, M) = K^{0.25}L^{0.15}M^{0.40}$ suppose it is known that $Q(u, p) = 200$. How much of the 200 was spent on the input factor M?

5.4. For cost-minimizing producers show that the elasticity of cost with respect to the price of an input factor equals the cost share of that input factor.

5.5. Use the method of Lagrange multipliers to derive a closed-form expression for the cost function for the general CES production function $\Phi(x) = [\sum_i \alpha_i x_i^\rho]^{1/\rho}$.

5.6. Consider the production function $\Phi(K, L) = [16K^{1/3} + 9L^{1/3}]^3$. The current input vector is $x = (K, L) = (8, 27)$.

(a) What is the output rate?
(b) What is the rate of technical substitution?
(c) Suppose the labor input decreases by 1%. Explain how to use your answer to (b) to estimate the capital input required to maintain the current output level.

(d) What is the factor price ratio p_K/p_L if x is a cost minimizer?
(e) Suppose as a result of a labor bargaining agreement the factor price ratio declines by 2%. Use the elasticity of substitution to estimate the new cost-minimizing input vector.

5.7. Consider the input-output data shown in Table 5.2.

Table 5.2. Input-output data for Exercise 5.7.

Firm	x_1	x_2	u
1	2	4	2
2	4	2	3
3	6	6	5
4	8	6	6
5	6	8	7

(a) Suppose the price vector $p = (2, 1)$ and desired output rate is 5. Determine the minimal cost for the VRS and CRS technologies using the output-cost set.
(b) Answer (a) when the price vector $p = (1, 4)$.

5.8. Determine the minimal cost function for the technology characterized by the production function $\Phi(x_1, x_2, x_3) = x_1/a_1 + \min(x_2/a_2, x_3/a_3)$.

5.9. This exercise generalizes the previous exercise. For this problem the input vector $x = (x^1, x^2, \ldots, x^M)$ is separated into M subsets of inputs. (Each component $x^k \in \mathbb{R}_+^{n^k}$ and $\sum_{k=1}^M n^k = n$.) The technology is characterized by an *additively-separable* form $\Phi(x) = \sum_k \phi^k(x^k)$. Each component function $\phi^k(\cdot)$ is a well-behaved production function in its own right. Moreover, each $\phi^k(\cdot)$ is homothetic, which means its minimal cost function can be represented as

$$Q^k(u^k, p^k) = \min\{p^k \cdot x^k : \ \phi^k(x^k) \geq u^k\} = f^k(u^k)P^k(p^k).$$

(a) Show that

$$Q(u, p) = \min\Big\{\sum_k f^k(u^k)P^k(p^k) : \ \sum_k u^k \geq u\Big\}.$$

(b) Derive an exact expression for $Q(u, p)$ when $f^k(u^k) = u^k$ for all k.
(c) Derive an exact expression for $Q(u, p)$ when $f^k(u^k) = (u^k)^\beta$, $\beta > 1$ for all k.
(d) Show how to compute $Q(u, p)$ when $f^k(u^k) = (u^k)^{\beta_k}$, $\beta_k > 1$ for all k. (It is not possible to obtain a general closed-form solution.)

5.10. Determine the minimal cost function for the technology characterized by the production function $\Phi(x_1, x_2, x_3, x_4) = \max\{b_1x_1 + b_2x_2,\ b_3x_3 + b_4x_4\}$.

5.11. Use the Envelope Theorem (see Appendix F) to prove (5.23) and Shephard's lemma.

5.12. Fix $u > 0$ and a strictly positive price vector $p_0 \in \mathbb{R}^n_{++}$. Let x_0 be a cost minimizing input vector for the $Q(u, p_0)$ problem. Assume that $Q(u, \cdot)$ is differentiable in p. Define $\psi(p) := Q(u, p) - p \cdot x_0$.

(a) Explain why $\psi(p) \leq 0$ for all $p \in \mathbb{R}^n_+$.
(b) Explain why $\psi(p_0) = 0$.
(c) Use parts (a) and (b) to establish Shephard's lemma.

5.13. The purpose of this problem is to provide a graphical proof in two dimensions of the concavity of the cost function with respect to prices. To simplify matters a bit, assume the well-behaved production function $\Phi(K, L)$ is differentiable so that the isoquant is smooth. Pick $p^i \in \mathbb{R}^2_{++}$, $i = 1, 2$, and assume that p^1 and p^2 are not proportional. For each $\lambda \in [0, 1]$ define

$$\begin{aligned}
p^\lambda &:= \lambda p^1 + (1 - \lambda)p^2 \\
Q^\lambda &:= \lambda Q(u, p^1) + (1 - \lambda)Q(u, p^2) \\
L^i &:= \{z \in \mathbb{R}^2_+ :\ p^i \cdot z = Q(u, p^i)\},\ i = 1, 2, \\
H^i &:= \{z \in \mathbb{R}^2_+ :\ p^i \cdot z \geq Q(u, p^i)\},\ i = 1, 2, \\
L^\lambda &:= \{z \in \mathbb{R}^2_+ :\ p^\lambda \cdot z = Q^\lambda\} \\
H^\lambda &:= \{z \in \mathbb{R}^2_+ :\ p^\lambda \cdot z \geq Q^\lambda\}.
\end{aligned}$$

(a) Let z_{12} be the point of intersection of the lines L^1 and L^2. Explain why $z_{12} \notin L(u)$.
(b) Intuitively the slope of the line L^λ should be a convex combination of the slopes of the lines L^1 and L^2 and should pass through the point z_{12}. Verify this algebraically.
(c) Provide a graphical illustration of why $L(u) \subset H^1 \cap H^2 \subset H^\lambda$.
(d) Using the fact that the hyperplane

$$\{z \in \mathbb{R}^n_+ :\ p^\lambda \cdot z = Q(u, p^\lambda)\}$$

supports $L(u)$, show how parts (a)-(c) immediately implies (5.4), the concavity of $Q(u, \cdot)$.
(e) Illustrate this graphical proof of concavity with the following example: $\Phi(K, L) = \sqrt{KL}$, $u = 2$, $p^1 = (4, 1)$, $p^2 = (1, 4)$ and $\lambda = 0.5$.

5.14. Suppose the set $L \subset \mathbb{R}^n_+$ is closed and exhibits input free disposability. The set L, however, may not be convex. For $p \in \mathbb{R}^n_+$ define

$$Q(p) := \inf\{\ p \cdot z :\ z \in L\} \tag{5.41}$$

$$Q^*(p) := \inf\{p \cdot x :\ x \in L^*\}. \tag{5.42}$$

and define

$$L^* := \{z \in \mathbb{R}^n_+ : \ p \cdot z \geq Q(p) \text{ holds for all } p \in \mathbb{R}^n_+\}. \tag{5.43}$$

(a) Prove that $Q^*(p) = Q(p)$.
(b) Give a concise proof in words why L^* is closed, convex and exhibits input free disposability.
(c) Prove that L^* is the smallest closed, convex and input free disposable set that contains L.
(d) Is the assumption of convexity for the input requirement sets $L(u)$ unnecessarily restrictive for the economic analysis of cost (e.g., the sensitivity of cost to output, prices, etc.)? What behavior on the part of producers must be assumed? How much of the underlying technology can be recovered? Briefly discuss.

5.15. Given a well-behaved production function, the cost function

$$Q(u, p) := \min\{p \cdot x : \ \Phi(x) \geq u\}$$

is concave in p for fixed u.

(a) Prove that for fixed p the cost function is *convex* in u when the production function $\Phi(\cdot)$ is concave.
(b) Give a concrete example of a production function for which the cost function is *concave* in u for fixed p.

5.10 Bibliographical Notes

Shephard [1970], Varian [1992], Jehle and Reny [2001], Mas-Colell et. al. [1995] and Chambers [1988] provide extensive developments of the theory. Further reading on duality theory can be found in Fuss and McFadden [1978], Diewert [1982] and Fare [1988].

5.11 Solutions to Exercises

5.1 (a) First-order optimality conditions give:

$$p_K = \lambda 0.2 K^{-0.8} L^{0.4},$$
$$p_L = \lambda 0.4 K^{0.2} L^{-0.6},$$

which implies that $p_K/p_L = (0.2/0.4)(L/K)$ or that $L^* = K^*$. Since the desired output rate is 8, we must have $L^{0.2}L^{0.4} = 8$, which gives $L = 32 = K$. The minimum cost is therefore $1(32) + 2(32) = 96$.

(b) Use the fact that $\partial Q(u,p)/\partial u = \lambda^*$. For example,

$$1 = \lambda^* 0.2(32^{-0.8}32^{0.4}) = 0.05\lambda^*,$$

which gives $\lambda^* = 20$. Since the change in output rate is $+0.08$ the minimum cost is estimated to increase by $(0.08)(20) = 1.6$.

(c) The elasticity of cost with respect to output is $\frac{u\partial Q(u,p)/\partial u}{Q(u,p)} = 8(20)/96 = 1.6\bar{6}$.

(d) The first-order optimality conditions do not change, so once again $L^* = K^*$. Thus, $L^{0.2}L^{0.4} = 8.08$, which gives $L^* = 32.535 = K^*$ and a minimum cost of $1(32.535)+2(32.535) = 97.605$. Note that $100(97.605-96)/96 = 1.67\%$, which is quite close to the answer for part (c), as expected.

(e) Using Shephard's Lemma, we know that $\partial Q(u,p)/\partial p_K = K^* = 32$. Thus, the new cost should increase by approximately $32(0.01) = 0.32$. To compute the actual cost, since p_L remains unchanged, the first-order optimality conditions imply that $L^* = p_K K^* = 1.01K^*$. Thus, $1.01^{0.4}K^{0.6} = 8$, which gives $K^* = 31.7884$ and $L^* = 32.1063$. The minimum cost equals $1.01K^*+2(1.01K^*) = 3.03K^* = 96.3189$, which indeed represents an increase of almost 0.32, as expected.

5.2 Using (5.14), $\alpha_K = \frac{p_K K}{p_K K + p_L L}$, which in turns implies that

$$L = \frac{p_K(1-\alpha_K)}{\alpha_K p_L} K.$$

Substituting this identity into $u = \Phi(K,L)$ yields

$$u = A\Big(\frac{p_K(1-\alpha_K)}{\alpha_K p_L}\Big)^{\alpha_L} K^{\alpha_K+\alpha_L}$$

from which the value of K^* and then L^* can be determined. For problem 5.1(a), we have $K/(K+2L) = 1/3$, which immediately gives $K^* = L^*$.

5.3 Using (5.14), the cost share for factor M is $0.40/(0.25+0.15+0.40) = 0.50$, which means that $0.5(200) = 100$ was spent on factor M.

5.4 A direct result of Shephard's Lemma:

$$\frac{p_i \partial Q(u,p)/\partial p_i}{Q(u,p)} = \frac{p_i x_i^*}{Q(u,p)} = S_i.$$

5.5 To simplify the derivation, rewrite the minimal cost function as

$$Q(u,p) = \min \left\{ q \cdot y : \sum_i y_i^\rho \geq v \right\},$$

where $y_i := \alpha_i^{1/\rho} x_i$, $q_i := p_i/\alpha_i^{1/\rho}$ and $v := u^\rho$. First-order optimality conditions are

$$q_i = (\lambda\rho) y_i^{\rho-1},\ 1 \leq i \leq n. \tag{5.44}$$

Multiply both sides of (5.44) by y_i, sum over i, and use the fact that

$$\sum_i y_i^\rho = v \tag{5.45}$$

to conclude that

$$Q(u,p) = (\lambda\rho) v. \tag{5.46}$$

It remains to find the expression for $(\lambda\rho)$. Use (5.44) to express

$$y_i^\rho = (\lambda\rho/q_i)^{\rho/(1-\rho)},$$

then use (5.45) to conclude that

$$v = \left[\sum_i q_i^{\rho/(\rho-1)} \right] (\lambda\rho)^{\rho/(1-\rho)}.$$

Thus,

$$(\lambda\rho) = v^{(1-\rho)/\rho} \left[\sum_i q_i^{\rho/(\rho-1)} \right]^{(\rho-1)/\rho}$$

and consequently

$$Q(u,p) = u \left[\sum_i \alpha_i^{1-r} p_i^r \right]^{1/r},$$

where $r := \rho/(\rho-1)$. Note how the cost function possesses the general CES form, too.

5.6 (a) $\Phi(8,27) = [16(8)^{1/3} + 9(27)^{1/3}]^3 = 59^3 = 205,379$.

(b) The rate of technical substitution equals

$$\frac{3[16K^{1/3} + 9L^{1/3}](16/3)K^{-2/3}}{3[16K^{1/3} + 9L^{1/3}](9/3)L^{-2/3}} = \frac{16}{9}\left(\frac{L}{K}\right)^{2/3} = 4.$$

(c) If the labor input decreases by 1%, then it decreases by an absolute amount of 0.27. We know that $\Delta L \approx 4\Delta K$, which implies that $\Delta K \approx (0.25)(0.27) = 0.0675$. Thus, the capital input required to maintain the current input level increase to approximately 8.0675. (The actual value is 8.0679.)

(d) The factor price ratio equals the rate of technical substitution, which is 4.

(e) Since $p_K/p_L = (16/9)(L/K)^{2/3}$, the elasticity of substitution is the reciprocal of 2/3 or 1.5. (See Remark 2.25, p. 28.) Thus, a 2% decline in the factor price ratio results in a 3% decline in the labor-capital ratio. Hence, $L = (0.97)(27/8)K = 3.27375K$. Substituting this identity into the production function, we have $59 = 16K^{1/3} + 9(3.27375K)^{1/3} = 29.36363K^{1/3}$, which gives $K = 8.11198$ and $L = 26.5566$. (The actual values for K and L are 8.11141 and 26.55884, respectively.)

5.7 (a) For this price vector, the points in the VRS output-cost set that define its efficient frontier are $\{(0,8),(2,8),(3,10),(7,20)\}$ (in order of increasing output). The line $u = 5$ intersects this set at the midpoint of the line segment joining points $(3,10)$ and $(7,20)$, and hence $Q(5,p) = 15$ for the VRS technology. The firm with the highest output per unit of cost is firm 5, and so $Q(5,p) = (5/7)(20) = 14.28$ for the CRS technology.

(b) For this price vector, the points in the VRS output-cost set that define its efficient frontier are $\{(0,12),(3,12),(6,32),(7,38)\}$ (in order of increasing output). The line $u = 5$ intersects this set two-thirds along the line segment joining points $(3,12)$ and $(6,32)$. Hence, $Q(5,p) = 12 + (2/3)(32-12) = 25.\overline{3}$ for the VRS technology. The firm with the highest output per unit of cost is firm 2, and so $Q(5,p) = (5/3)(12) = 20$ for the CRS technology.

5.8 By definition

$$Q(u,p) = \min\big\{p_1x_1 + p_2x_2 + p_3x_3 :\ x_1/a_1 + \min(x_2/a_2, x_3/a_3) \geq u\big\}.$$

Let $u_1 := a_1x_1$ and $u_2 := \min(x_2/a_2, x_3/a_3)$. *Given* u_1 and u_2 such that $u_1 + u_2 \geq u$, the minimal cost is obtained when $x_1 = a_1p_1$, $x_2 = a_2p_2$ and $x_3 = a_3p_3$. Thus, the minimal cost optimization problem can be reformulated as

$$Q(u,p) = \min\big\{(p_1a_1)u_1 + (p_2a_2 + p_3a_3)u_2 :\ u_1 + u_2 \geq u\big\}.$$

This problem is a simple *linear program* whose optimal value is given as

$$Q(u,p) = \min\big\{(p_1a_1), (p_2a_2 + p_3a_3)\big\}.$$

(Only one of u_1 or u_2 needs to be positive—for an explanation see Remark 6.15, p. 107.)

5.9 (a) Let $u^k := \phi^k(x^k)$. *Given* the u^k, the components of each x^k should be chosen to minimize their respective cost functions

$$Q^k(u^k,p^k) = \min\{p^k \cdot x^k :\ \phi^k(x^k) \geq u^k\}.$$

Thus, the minimum cost function problem can be reformulated using the u^k as decision variables, i.e.,

$$Q(u,p) = \min\left\{\sum_k Q^k(u^k,p^k) : \sum_k u^k \geq u\right\}.$$

The final form uses the fact that each cost function is homothetic.

(b) This is a simple linear program. The solution is to pick only one u^k to be positive. The index chosen coincides with the one whose $P^k(p^k)$ is the smallest, i.e.,

$$Q(u,p) = u \min_k\{P^k(p^k)\}.$$

(c) The optimization problem here is

$$Q(u,p) = \min\left\{\sum_k P^k(p^k)(u^k)^\beta : \sum_k u_k \geq u\right\}.$$

Since $\beta > 1$, this problem is a well-behaved convex optimization problem. To ease notational burdens, let $\gamma_k := P^k(p^k)$ for each k. The first-order optimality conditions are $\gamma_k\beta(u^k)^{\beta-1} = \lambda$, $1 \leq k \leq M$. Invert these equations to express u^k in terms of λ/β and γ_k, and then use the fact that $\sum_k u^k = u$ to obtain the final solution

$$u^k = \frac{\gamma_k^{1/(1-\beta)}}{\sum_k \gamma_k^{1/(1-\beta)}}\, u.$$

Notice that since here $\beta > 1$, the convexity of the objective function ensures that *all* u^k will be positive in the optimal solution.

(d) Here, the first-order optimality conditions are $\gamma_k\beta_k(u^k)^{\beta_k-1} = \lambda$, $1 \leq k \leq M$. Inverting these equations yields

$$u^k = u^k(\lambda) := (\lambda/\gamma_k\beta_k)^{1/(\beta_k-1)}$$

for each k. Since $\sum_k u^k = u$, we seek to find a value of λ for which $\psi(\lambda) := \sum_k u^k(\lambda) = u$. Since $\psi(\cdot)$ is increasing and continuous with $\psi(0) = 0$ and $\psi(\lambda) \to \infty$ as $\lambda \to \infty$, there is a *unique* value of λ for which $\psi(\lambda) = u$. To find this value of λ one may perform, for example, a simple *bisection search*: essentially, if the current guess for λ is such that $\psi(\lambda) < u$, then increase the lower bound for λ, or if the current guess for λ is such that $\psi(\lambda) > u$, then decrease the upper bound for λ, and continue to move back-and-forth reducing the interval of uncertainty until the desired level of precision is reached.

5.10 By definition

$$Q(u,p) = \min\Big\{p_1x_1 + p_2x_2 + p_3x_3 + p_4x_4 : \max\{b_1x_1 + b_2x_2, b_3x_3 + b_4x_4\} \geq u\Big\}.$$

Let $u_1 = b_1x_1 + b_2x_2$ and $u_2 = b_3x_3 + b_4x_4$. *Given* u_1 and u_2 such that $\max\{u_1, u_2\} \geq u$, the minimal cost is obtained by choosing values for the x_i to solve the two *separable* minimal cost problems

$$\min\{p_1x_1 + p_2x_2 :\ b_1x_1 + b_2x_2 \geq u_1\},\ \min\{p_3x_3 + p_4x_4 :\ b_3x_3 + b_4x_4 \geq u_2\}.$$

Once again, these are simple linear programs whose optimal values are given by $Q_{12} := \min(p_1/b_1, p_2/b_2)$ and $Q_{34} := \min(p_3/b_3, p_4/b_4)$, respectively. Thus, the minimal cost problem can be reformulated now as

$$Q(u, p) = \min\big\{Q_{12}u_1 + Q_{34}u_2 :\ \max(u_1, u_2) \geq u\big\}.$$

The solution here is obvious: set either u_1 or u_2 to u depending on which value Q_{12} or Q_{34} is lower, i.e.,

$$Q(u, p) = u \min\{Q_{12}, Q_{34}\} = u \min\{p_1/b_1, p_2/b_2, p_3/b_3, p_4/b_4\}.$$

That is, only one of the x_i should be chosen positive, and the index chosen should correspond to the input with the lowest p/b ratio.

5.11 The Lagrangian for the minimum cost problem is

$$L(x, u, p) = p \cdot x - \lambda\{\Phi(x) - u\}.$$

The Envelope Theorem says that

$$\frac{\partial Q}{\partial u} = \frac{\partial L(x^*, \lambda^*, u, p)}{\partial u} = \lambda_i^*,$$
$$\frac{\partial Q}{\partial p_i} = \frac{\partial L(x^*, \lambda^*, u, p)}{\partial p_i} = x_i^*.$$

5.12 (a) Since $x^* \in L(u)$ it will always be the case that $p \cdot x^* \geq Q(u, p)$ for any price vector p.

(b) By definition of cost minimizer, $p_0 \cdot x_0 = Q(u, p_0)$.

(c) Since p_0 is a strictly positive solution to the optimization problem $\max\{\psi(p) : p \in \mathbb{R}^n_+\}$, the gradient of $\psi(\cdot)$ at p_0 must vanish, i.e., $\nabla\psi(p_0) = 0$, which directly implies Shephard's lemma.

5.13 (a) Since the production function is differentiable, (i) there is a unique point at which each line L^i is tangent to the isoquant of $L(u)$, and (ii) these tangent points must be different since the price vectors are not proportional. Thus, the point z_{12} cannot be either of these tangent points. By definition of tangency z_{12} cannot then belong to $L(u)$.

(b) We have

$$\frac{\lambda p_K^1 + (1-\lambda)p_K^2}{\lambda p_L^1 + (1-\lambda)p_L^2} = \frac{\lambda p_K^1}{\lambda p_L^1} \frac{\lambda p_L^1}{\lambda p_L^1 + (1-\lambda)p_L^2} + \frac{(1-\lambda)p_K^2}{(1-\lambda)p_L^2} \frac{(1-\lambda)p_L^2}{\lambda p_L^1 + (1-\lambda)p_L^2} = \mu\frac{p_K^1}{p_L^1} + (1-\mu)\frac{p_K^2}{p_L^2},$$

where $\mu \in [0,1]$. Furthermore,

$$p^\lambda \cdot z_{12} = \lambda(p^1 \cdot z_{12}) + (1-\lambda)(p^2 \cdot z_{12}) = \lambda Q(u,p^1) + (1-\lambda)Q(u,p^2) = Q^\lambda,$$

since $z_{12} \in L^i$, $i = 1,2$.

(c) The line L^λ passes through z_{12} and has slope that lies in between the slopes of L^1 and L^2.

(d) Since $z_{12} \notin L(u)$, $p^\lambda \cdot z_{12} = Q^\lambda$, and $L(u) \subset H^1 \cap H^2 \subset H^\lambda$, it must be the case that the line L^λ must be moved "up" to make it tangent to $L(u)$. Thus, $Q(u,p^\lambda) > Q^\lambda$, which is precisely the definition of concavity.

(e) The rate of technical substitution is L/K for this production function. Thus, when the price vector is p^1, we must have that $4/1 = L/K$ or $L = 4K$, which implies that $\sqrt{K(4K)} = 2$ or $K_1^* = 1$ and $L_1^* = 4$. By symmetry, when the price vector is p^2, $K_2^* = 4$ and $L_2^* = 1$. In each case the minimum cost is 8. The lines are $L = -4K + 8$ in the first case and $L = -0.25K + 2$ in the second case. The point of intersection is $(1.6, 1.6)$. When $\lambda = 0.5$, the new price vector is $(2.5, 2.5)$ and the equation of the line L^λ is $L = -K + 3.2$. When the price vector is p^λ, due to symmetry the cost minimizing input vector is easily seen to be $(2,2)$ and so $Q(u,p^\lambda) = 10$. This is less than $p^\lambda \cdot z_{12} = 8$, as it should.

5.14 (a) Here $Q(p)$ is the cost function with the value of output rate u being suppressed. It follows by definition of $Q(p)$ that $L \subset \{z :\ p \cdot z \geq Q(p)\}$ for any $p \in \mathbb{R}^n_+$. This fact immediately implies that $L \subset L^*$. Since the value of $Q^*(p)$ is obtained by minimizing "cost" on a *larger* set L^* it then directly follows that $Q^*(p) \leq Q(p)$. Conversely, if $x \in L^*$, then $p \cdot x \geq Q(p)$ for each fixed p. Hence, the infimum of all such $p \cdot x$ over p, which is $Q^*(p)$, must be at least as large as $Q(p)$.

(b) L^* is the intersection of closed halfspaces, each of which is closed, convex, and exhibits input free disposability. Each of these three properties is closed under arbitrary intersections.

(c) Let C be any closed, convex, and input free disposable set that contains L. If $x \notin C$, then by the separation theorem for convex sets, there exists a $p_0 \in \mathbb{R}^n$ and scalar α such that $p_0 \cdot x < \alpha < p_0 \cdot c$ for all $c \in C$. Since C exhibits input free disposability, all of the coordinates of p_0 must be nonnegative. Since C contains L it follows from the definition of $Q(p)$ that $p_0 \cdot x < Q(p_0)$. This in turn implies that $x \notin L^*$. Thus $L^* \subset C$, which means that L^* is indeed the smallest set possessing the three properties that contains L. In shorthand, $L^* = Conv(L)$!

(d) Under reasonable assumptions, when producers are cost-minimizing, for purposes of analysis one may just as well work with L^*, the convex hull of L, as with L. That is, the intersection of all closed halfspaces that contain a set L is the closed convex hull of L.

5.15 (a) Fix $u_i \geq 0$, $i = 1,2$, and $\lambda \in [0,1]$. By definition,

$$Q(\lambda u_1 + (1-\lambda)u_2, p) = \min\{p \cdot x :\ \Phi(x) \geq \lambda u_1 + (1-\lambda)u_2\}.$$

Let x_i^* be a cost minimizing input vector associated with the $Q(u_i, p)$ problem, and define $x = \lambda x_1^* + (1-\lambda)x_2^*$. Since $\Phi(\cdot)$ is concave,

$$\Phi(x) \geq \lambda\Phi(x_1^*) + (1-\lambda)\Phi(x_2^*) \geq \lambda u + (1-\lambda)u = u.$$

Thus, the vector x is *feasible* for the $Q(\lambda u_1 + (1-\lambda)u_2, p)$ problem. In particular, this means that

$$\begin{aligned} Q(\lambda u_1 + (1-\lambda)u_2, p) &\leq p \cdot x \\ &= \lambda(p \cdot x_1^*) + (1-\lambda)(p \cdot x_2^*) \\ &= \lambda Q(u_1, p) + (1-\lambda)Q(u_2, p), \end{aligned}$$

which establishes the convexity of $Q(\cdot, p)$.

(b) Consider $\Phi(x) = x^2$ (a one-factor technology). Here, the minimal cost function is trivially $p\sqrt{u}$, which is obviously concave in u. Notice that the concavity of the cost function in u here stems from the *convexity* or *increasing* returns-to-scale of the underlying technology.

6

Indirect Production Function

We examine profit-maximizing producers faced with a budget constraint on inputs. The indirect production function provides a dual representation of technology, as it can be used to reconstruct the production function, and hence the technology set. It is also possible, under appropriate conditions, to use the indirect production function to immediately determine the cost function, and vice-versa. While we couch the presentation in terms of production, the development applies to the consumer setting in which output is interpreted as consumer utility, and the consumer wishes to maximize his utility subject to a budget constraint on expenditures. Throughout this chapter we make the following assumption:

Assumption 1 *The derived production function $\Phi(\cdot)$ is continuous.*

Remark 6.1. As a direct application of the Theorem of the Maximum (see Appendix H), the production function will be continuous if the output correspondence $P(\cdot)$ is continuous.

6.1 Definition

Definition 6.2. *For a given price vector $p \in \mathbb{R}^n_+$ and budget $B \in \mathbb{R}_+$ the* **budget set** *is*

$$\mathcal{B}(p, B) := \{x \in \mathbb{R}^n_+ : \ p \cdot x \leq B\}.$$

Each input vector x that belongs to the budget set is **budget feasible**.

Definition 6.3. *The* **indirect production function** $\Gamma_\Phi : \mathbb{R}^n_{++} \times \mathbb{R}_{++} \longrightarrow \mathbb{R}_+$ *is*

$$\Gamma_\Phi(p, B) := \max\{\Phi(x) : \ p \cdot x \leq B\}.^1$$

[1] The budget set is compact when prices are positive, and hence an optimal solution exists by Assumption 1.

Keep in mind that the indirect production function is defined only for *positive* prices and budget.

Since the budget set remains unchanged if all prices and budget are multiplied by a positive constant, it is possible to work with **normalized prices**, namely p/B, and analyze the normalized indirect production function.

Definition 6.4. *The* **normalized indirect production function** $\Gamma_\Phi : \mathbb{R}^n_{++} \longrightarrow \mathbb{R}_+$ *is*

$$\Gamma_\Phi(p) := \Gamma_\Phi(p, 1).$$

We shall refer to a normalized indirect production function simply as an indirect production function. There will be no cause for confusion, since a normalized indirect production function has only one argument. We shall also drop the subscript Φ and write $\Gamma(\cdot)$ instead of $\Gamma_\Phi(\cdot)$.

Example 6.5. Indirect production function corresponding to the generalized Cobb-Douglas production function. Let $\Phi(x) = \prod_i x_i^{\alpha_i}$ be a generalized Cobb-Douglas production function with $A = 1$. In lieu of maximizing $\Phi(\cdot)$ directly, it will be easier to maximize the natural logarithm of $\Phi(\cdot)$ and solve for $ln\ (\Gamma(p))$. The first-order optimality conditions imply that

$$\frac{\alpha_i}{x_i} = \lambda p_i \text{ for each i,} \tag{6.1}$$

from which it directly follows that

$$\Gamma(p) = \Phi(x) = \prod_i \left(\frac{\alpha_i}{\lambda p_i}\right)^{\alpha_i}.$$

Since $\Phi(\cdot)$ is increasing, the budget constraint will be tight. Given (6.1), this in turn implies that

$$1 = p \cdot x = \sum_i p_i x_i = \left(\sum_i \alpha_i\right)/\lambda; \tag{6.2}$$

thus, $\lambda = \sum_i \alpha_i$. Consequently,

$$\Gamma(p) = a^{-a} \prod_i \left(\frac{\alpha_i}{p_i}\right)^{\alpha_i} \text{ and } \Gamma(p, B) = B^a \Gamma(p), \text{ where } a := \sum_i \alpha_i. \tag{6.3}$$

6.2 Properties

It follows from its definition that the indirect production function is nonincreasing in prices p and nondecreasing in budget B. As the next theorem shows, $\Gamma(\cdot)$ is quasiconvex. A function $f(\cdot)$ is quasiconvex if $f(\lambda x + (1-\lambda)z) \leq \max\{f(x), f(z)\}$. Equivalently, f is quasiconvex if each of its *lower* level sets $L_f^{\leq}(\alpha) := \{x : \ f(x) \leq \alpha\}$ is convex.

Proposition 6.6. *$\Gamma(\cdot)$ is quasiconvex under Assumption 1.*

Proof. Pick $p_i > 0$, $i = 1, 2$, pick $\lambda \in [0, 1]$, and let $u_i := \Gamma(p_i)$. Without loss of generality assume that $u_1 \leq u_2$. We must show that $\Gamma(\lambda p_1 + (1-\lambda)p_2) \leq u_2$. From its definition,

$$\Gamma(\lambda p_1 + (1-\lambda)p_2) = \max\{\Phi(x) : \ (\lambda p_1 + (1-\lambda)p_2) \cdot x \leq 1 \ \}. \tag{6.4}$$

Pick a v larger than u_2. If $\Phi(x) \geq v$, then $p_i \cdot x > 1$, $i = 1, 2$; otherwise, $\Gamma(p_i) > v > u_i$, an obvious contradiction. Consequently, if $\Phi(x) \geq v$, then x is not budget feasible in (6.4), which implies that $\Gamma(\lambda p_1 + (1-\lambda)p_2) < v$. Since v was chosen arbitrarily, the result follows. □

Remark 6.7. The assumption of quasiconcavity of $\Phi(\cdot)$ was *not* used in the proof. The indirect production function will *always* be quasiconvex even if $\Phi(\cdot)$ is not quasiconcave.

Remark 6.8. An alternative proof of Proposition 6.6 uses separation theory, as follows. Let S denote the intersection of all open halfspaces that contain $L_\Gamma^{\leq}(u)$. S is obviously convex and $L_\Gamma^{\leq}(u) \subseteq S$. It remains to establish the reverse inclusion $S \subseteq L_\Gamma^{\leq}(u)$. To this end, pick a $q \geq 0$ such that $q \notin L_\Gamma^{\leq}(u)$. By definition of $\Gamma(\cdot)$, there must exist an x for which $q \cdot x \leq 1$ and $\Phi(x) > u$. Consequently, $L_\Gamma^{\leq}(u)$ is contained in the open halfspace $\{p > 0 : \ x \cdot p > 1\}$, which obviously does not contain q. Thus, $q \notin S$, as required. □

6.3 Duality between the Cost and Indirect Production Functions

Two additional assumptions will be used for the results of this section.

Assumption 2 *The production function $\Phi(\cdot)$ is increasing: if $x \geq y$ and $x \neq y$, then $\Phi(x) > \Phi(y)$.*

Assumption 3 *The production function $\Phi(\cdot)$ is strictly quasiconcave: $\Phi(\lambda x + (1-\lambda)z) > \min\{\Phi(x), \Phi(z)\}$ for each $x, z \in \mathbb{R}^n_+$ and $\lambda \in (0, 1)$.*

To solve for $\Gamma(p)$, one can solve the first-order optimality conditions as we did in Example 6.5. Under Assumptions 1 and 2, it turns out that knowledge of the functional form for the cost function $Q(u, p)$ will enable one to directly solve for $\Gamma(p)$ and vice-versa.

Proposition 6.9. *Under Assumptions 1 and 2,*

a) $Q(\Gamma(p, B), p) = B$.
b) $\Gamma(p, Q(u, p)) = u$.

Proof. Part (a). Pick a price $p > 0$, a budget $B > 0$ and let $v := \Gamma(p, B)$. A solution to the producer's optimization problem necessarily costs no more than B and achieves at least output rate v. Clearly, then, $Q(v, p) \leq B$. It remains to rule out the possibility that $Q(v, p) < B$. Suppose, instead, there exists a z for which $p \cdot z < B$ and $\Phi(z) \geq v$. Let $\lambda := B/(p \cdot z)$ and define $\hat{z} := \lambda z$. Clearly, $p \cdot \hat{z} = B$, and so $\hat{z}$ is budget feasible. Since $\lambda > 1$ and $\Phi(\cdot)$ is increasing, it follows that $\Phi(\hat{z}) > \Phi(z) = v$. Consequently, $\Gamma(p, B) > v$, an obvious contradiction of the definition of v.

Part (b). Pick a $p > 0$, a $u > 0$ and let $E := Q(u, p)$. A solution to the cost minimization problem $\min\{p \cdot x : \ \Phi(x) \geq u\}$ achieves output rate u and costs E. Clearly, then, $\Gamma(p, E) \geq u$. It remains to rule out the possibility that $\Gamma(p, E) > u$. Suppose, instead, there exists a z for which $p \cdot z \leq E$ and $\Phi(z) > u$. Since $\Phi(\cdot)$ is continuous, it is possible to find a $\lambda < 1$ for which $\Phi(\lambda z) > u$. Define $\hat{z} = \lambda z$. We have that $\Phi(\hat{z}) > u$ and $p \cdot \hat{z} < E = Q(u, p)$, an obvious contradiction of the definition of minimal cost. □

Remark 6.10. Here is a geometrical interpretation why these identities hold. When $\Phi(\cdot)$ is continuous and increasing, the hyperplane $\{x \geq 0 : \ p \cdot x = B\}$ must *support* the input possibility set $L_\Phi(\Gamma(p, B))$. Our analysis of the cost function showed that the hyperplane $\{x \geq 0 : \ p \cdot x = Q(u, p)\}$ supports $L_\Phi(u)$, too. There is no "gap" between these hyperplanes because the isoquant is not "thick."

Example 6.11. Suppose $\Phi(\cdot)$ is a general Cobb-Douglas form. The functional form for $\Gamma(p, B)$ was obtained in Example 6.5 using the first-order optimality conditions. Determining $\Gamma(p, B)$ is trivial with the use of the dual identities of Proposition 6.9: merely substitute $\Gamma(p, B)$ for u in the previously derived cost function formula (5.17) and invert it to obtain

$$\Gamma(p, B) = \frac{A}{\prod_i (p_i/\alpha_i)^{\alpha_i}} \left[\frac{B}{\sum_i \alpha_i} \right]^{\sum_i \alpha_i}.$$

Compare with (6.3). If, on the other hand, a formula for $\Gamma(p, B)$ had been derived, then one may recover $Q(u, p)$ by substituting it for B in this formula.

6.4 Reconstructing the Technology

For profit-maximizing producers faced with a budget constraint on inputs, it turns out that knowledge of how their output choice varies with respect to prices implicitly reveals the underlying technology, too.

Theorem 6.12. *The following identity holds under Assumption 1:*

$$\Phi^*(x) := \inf\{\Gamma(p) : p \cdot x \leq 1\} = \Phi(x).^2$$

[2] Keep in mind that the minimization problem defined here is with respect to the price vector p and *not* x.

Proof. Pick an $x \geq 0$. Since $\Gamma(p) \geq \Phi(x)$ for each feasible p, the definition of $\Phi^*(x)$ implies that $\Phi(x) \leq \Phi^*(x)$. It remains to show the reverse inequality $\Phi(x) \geq \Phi^*(x)$. To this end, pick an arbitrary value of v larger than $\Phi(x)$. Since $x \notin L_\Phi(v)$, a nonempty, closed and convex set, it is possible to strictly separate x from $L_\Phi(v)$. That is, there exists a $q \geq 0$ and $\alpha \geq 0$ for which

$$q \cdot z > \alpha > q \cdot x \text{ for all } z \in L_\Phi(v). \tag{6.5}$$

Since $\Phi(\cdot)$ is nondecreasing, each component of q must be nonnegative. If need be, it is possible to perturb q so that each of its components is positive and (6.5) remains valid. Define $p := q/\alpha$. Clearly, $L_\Phi(v) \cap \{z \geq 0 : \ p \cdot z \leq 1\}$ is empty, and so $\Gamma(p) < v$. Obviously $p \cdot x \leq 1$ and p is positive; consequently, it follows that $\Phi^*(x) < v$. As v was chosen arbitrarily, the result follows. □

Proposition 6.6 establishes that each of the lower level sets of $\Gamma(\cdot)$ is convex. Moreover, each lower level set exhibits input free disposability, and the family of lower level sets defined by $\Gamma(\cdot)$ is nested in the obvious way. Direct arguments or the Theorem of the Maximum (see Appendix H.1) can be used to show that $\Gamma(\cdot)$ is continuous and hence each lower level set is closed. Consequently, this family of lower level sets defines a well-behaved technology in the *price* space. In this sense, the indirect production function is the dual to the production function.

6.5 Revealed Preference

We now show how to directly apply Theorem 6.12 to the estimation of $\Phi(x)$ for a given input vector x. Let

$$\mathcal{D} = \Big\{(p_1, B_1, \Gamma(p_1, B_1)), (p_2, B_2, \Gamma(p_2, B_2)), \ldots, (p_N, B_N, \Gamma(p_N, B_N))\Big\}$$

denote observed data on N firms. Assume the observed prices and budgets are all positive, and, without changing notation, assume prices have been normalized by their respective budgets. Let $I(x) := \{i : \ p_i \cdot x \leq 1\}$. As an immediate consequence of Theorem 6.12,

$$\Phi(x) \leq \min_{i \in I(x)} \Gamma(p_i) := \Phi^a(x). \tag{6.6}$$

Without additional data, one may approximate $\Phi(x)$ with $\Phi^a(x)$.

The economic argument for (6.6), often referred to as a *Revealed Preference* argument, is this: if $p_i \cdot x \leq 1$, then x was budget feasible when the normalized prices were p_i. By the very definition of $\Gamma(p_i)$, it cannot be the case that $\Phi(x) > \Gamma(p_i)$, which leads directly to (6.6). As more data become revealed, the estimate $\Phi^a(x)$ in (6.6) will converge to the true output value from above.

6.6 Nonparametric Estimation

As in the previous section, let $\mathcal{D}$ denote the observed data set, and assume the observed prices and budgets are all positive and that the prices have been normalized by their respective budgets. Let $u_i := \Gamma(p_i, B_i)$.

We begin with the HR technology (see p. 60).

Proposition 6.13. *Let*

$$u^*(p) := \max\{u_i : \ p \cdot x_i \leq 1\}. \tag{6.7}$$

If $\Phi(\cdot)$ corresponds to the HR technology generated from $\mathcal{D}$, then

$$\Gamma_\Phi(p) = u^*(p).$$

Proof. It is immediate from their definitions that $\Gamma_\Phi(p) \geq u^*(p)$. We must show the reverse inequality $\Gamma_\Phi(p) \leq u^*(p)$. Pick a $p > 0$ and let $u := \Gamma_\Phi(p)$. By definition of the indirect production function,

$$L^{HR}(u) \cap \{x : \ p \cdot x \leq 1\} \neq \emptyset. \tag{6.8}$$

The input possibility set $L^{HR}(u)$ for the HR technology is the convex, input free disposable hull of those x_i for which $u_i \geq u$. Consequently, it follows from (6.8) that there is some x_i for which $u_i \geq u$ and $p \cdot x_i \leq 1$. This immediately implies that $u^*(p) \geq u = \Gamma_\Phi(p)$, as required. □

As we demonstrated with the cost function, it is possible to graphically determine $\Gamma(p)$ for the CRS and VRS nonparametric technologies. Recall the definition (5.33) of the output-cost set $\mathcal{OC}^{\mathcal{T}}(p)$: it represents the collection of all output-cost pairs that are technologically feasible when cost is measured at prices p. It follows from the definition of the indirect production function and the output-cost set that

$$\Gamma_\Phi(p) = \max\Big\{u : \ (u, c) \in \mathcal{OC}^{\mathcal{T}}(p) \cap \{(u, c) : \ c \leq 1\}\Big\},$$

which is easily determined from the output-cost set.

Example 6.14. Recall Example 5.10 and the data given in Table 5.1. The output-cost sets for the VRS and CRS technologies are depicted in Figures 5.2 and 5.3. As shown in Figure 6.1, the value for the indirect production function for the VRS and CRS technologies are 5.18 and 8.00, respectively. For the HR technology, the value for the indirect production function is only 4.

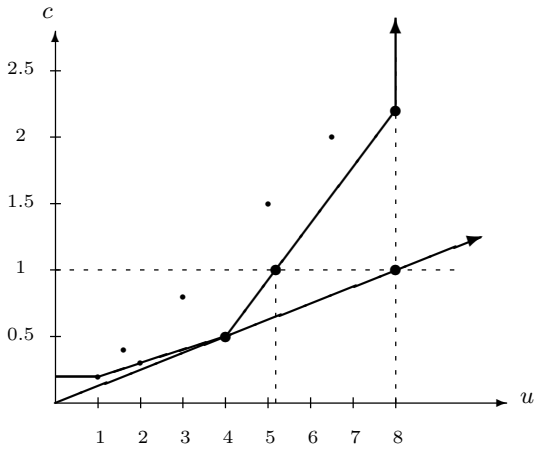


Fig. 6.1. Computation of the indirect production function from the output-cost set.

6.7 Exercises

6.1. Derive the indirect production function for the simple Leontief technology $\Phi(x) = \min_i x_i/a_i$ where a is strictly positive.

6.2. Let $\bar{x} \in \mathbb{R}^n_{++}$ and consider the *Stone-Geary* production function $\Phi(x) = \prod_{i=1}^n (x_i - \bar{x}_i)^{\beta_i}$ defined on $\{x \in \mathbb{R}^n_+ : \ x \geq \bar{x}\}$.

(a) Interpret $\bar{x}$.
(b) Derive the indirect production function.
(c) What interpretation do the β_i have?

6.3. Consider the CES production function $\Phi(x) = [\sum_i \alpha_i x_i^\rho]^{1/\rho}$.

(a) Derive $\Gamma(p, B)$ using the method of Lagrange multipliers.
(b) Derive the minimum cost function using the duality relationship between it and the indirect production function.

6.4. Suppose $\Gamma(p, B) = B/(p \cdot a)$ where a is strictly positive.

(a) Derive the minimum cost function using the duality relationship between it and the indirect production function.
(b) Derive the direct production function using the duality relationship between it and the indirect production function.

6.5. For the input-output data of Exercise 5.7 given in Table 5.2 on p. 87:

(a) Suppose the price vector $p = (2, 1)$ and the budget $B = 9$. Determine the maximal output for the HR, VRS and CRS technologies using the output-cost set.
(b) Answer (a) when the price vector $p = (1, 4)$ and $B = 35$.

6.6. Assume $\Phi(\cdot)$ is strictly quasiconcave, increasing, and differentiable. Let $x(p, B)$ denote the unique solution to the producer's maximization problem and assume the function x is differentiable. Prove *Roy's identity*, which states that

$$x_i(p, B) = -\frac{\frac{\partial \Gamma}{\partial p_i}}{\frac{\partial \Gamma}{\partial B}}.$$

6.7. The dual identities in Proposition 6.9 will hold for parametric models of technology, since these forms are continuous. Provide an example of a well-behaved two-input technology for which (i) $Q(\Gamma(p, B), p) < B$ and (ii) $\Gamma(p, Q(u, p)) > u$.

6.8. Let $\Phi : \mathbb{R}^n \to \mathbb{R}$ be finite-valued, upper semicontinuous and let $C \subset \mathbb{R}^n$ be compact. Prove that the sup $\{\Phi(x) : \ x \in C\}$ is achieved.

6.9. Let A be an $n \times m$ matrix, $b \in \mathbb{R}^m$ and let $C := \{x \in \mathbb{R}^n : \ Ax \leq b\}$. If C is compact and $\Phi : C \to \mathbb{R}$ is finite-valued, upper semicontinuous and concave, prove that

$$\nu(b) := \max\{\Phi(x) : \ x \in C\} \tag{6.9}$$

is concave.

6.8 Bibliographical Notes

Shephard [1970], Varian [1992], Jehle and Reny [2001], Mas-Colell et. al. [1995] and Chambers [1988] provide extensive developments of the theory.

6.9 Solutions to Exercises

6.1 By definition

$$\Gamma(p) = \max\{\Phi(x) : \ p \cdot x \leq 1\}.$$

Let $u = \Gamma(p)$. For the simple Leontief production function $x = u \cdot a$ will be the most efficient way to achieve u. The budget constraint $p \cdot (ua) \leq 1$ implies that $u \leq 1/(p \cdot a)$. Thus, $\Gamma(p)$ can be equivalently expressed as

$$\max\{u : \ u \leq 1/(p \cdot a)\}.$$

This yields $\Gamma(p) = 1/(p \cdot a)$ and $\Gamma(p, B) = B/(p \cdot a)$.

6.2 (a) Each coordinate of $\bar{x}$ represents a minimal amount of input necessary for any output to emerge.

(b) By definition

$$\begin{aligned}\Gamma(p, B) &= \max_{x \geq \bar{x}} \Big\{ \prod_i (x_i - \bar{x}_i)^{\beta_i} : \ p \cdot x \leq B \Big\} \\ &= \max_{y \geq 0} \Big\{ \prod_i y_i^{\beta_i} : \ p \cdot y \leq B - p \cdot \bar{x} \Big\},\end{aligned}$$

where $y_i := x_i - \bar{x}_i$ for each i. (If B is less than or equal to the cost of the minimum input vector $\bar{x}$, then the output is, of course, zero.) This maximization problem is identical to the one associated with a Cobb-Douglas production function provided in Example 6.11, except that B is replaced with $B - p \cdot \bar{x}$, which represents the discretionary budget.

(c) After netting out the cost of the minimum input vector $\bar{x}$ from the budget B, both the indirect production function and cost function here are identical to the ones associated with the Cobb-Douglas production function. It follows from the fundamental property (5.14), p. 75, that the β_i represent that portion of the discretionary budget devoted to the respective factor input.

6.3 (a) To simplify the derivation rewrite the indirect production function as

$$\Gamma(p, B) = \Gamma(\hat{p}) = \Big[\max \Big\{ \sum_i y_i^{\rho} : \ q \cdot y \leq 1 \Big\} \Big]^{1/\rho},$$

where $y_i := \alpha_i^{1/\rho} x_i$, $q_i := \hat{p}_i / \alpha_i^{1/\rho}$, and $\hat{p} = p/B$ represents normalized prices. First-order optimality conditions are

$$\rho y_i^{\rho - 1} = \lambda q_i, \ 1 \leq i \leq n. \tag{6.10}$$

Multiply both sides of (6.10) by y_i, sum over i, and use the fact that

$$q \cdot y = 1 \tag{6.11}$$

to conclude that

$$\Gamma(p) = (\lambda/\rho)^{1/\rho}. \tag{6.12}$$

It remains to find the expression for (λ/ρ). Use (6.10) to express

$$y_i = (\lambda/\rho)^{1/(\rho-1)} q_i^{1/(\rho-1)},$$

then use (6.11) to conclude that

$$1 = q \cdot y = \sum_i q_i y_i = \left[\sum_i q_i^{\rho/(\rho-1)}\right] (\lambda/\rho)^{1/(\rho-1)}.$$

Thus,

$$(\lambda/\rho) = \left[\sum_i q_i^{\rho/(\rho-1)}\right]^{1-\rho}$$

and consequently

$$\Gamma(p, B) = B \left[\sum_i \alpha_i^{1-r} p_i^r\right]^{-1/r},$$

where $r := \rho/(\rho - 1)$. Note how the cost function possesses the general CES form, too.

(b) The duality relationship $u = \Gamma(p, Q(u, p))$ and the solution to part (a) yields

$$Q(u, p) = u \left[\sum_i \alpha_i^{1-r} p_i^r\right]^{1/r}.$$

This is identical to the solution obtained using the Lagrange multiplier method of Exercise 5.5, p. 86.

6.4 (a) We have $\Gamma(p, Q(u, p)) = u$, which immediately yields $Q(u, p) = u(p \cdot a)$.

(b) The duality relationship implies that

$$\begin{aligned}
\Phi(x) &= \min\{\Gamma(p) : \ p \cdot x \leq 1\} \\
&= \Big[\max\{p \cdot a : \ p \cdot x \leq 1\}\Big]^{-1} \\
&= \Big[\max\big\{\sum_i q_i; \ \sum_i y_i q_i \leq 1\big\}\Big]^{-1},
\end{aligned}$$

where $q_i = p_i a_i$ and $y_i = x_i/a_i$ for each i. Let i^* be an index for which $y_{i^*} = \min_i y_i$. To achieve the maximum sum of the q_i, it is best to simply set $q_{i^*} = 1/y_{i^*}$ and the rest of the q_i equal to zero. This immediately implies that $\Phi(x) = \min_i x_i/a_i$, a simple Leontief production technology.

Remark 6.15. The maximization problem is an example of a *linear program.* Since there is only one constraint, only one variable can be positive. To see this for this specific problem, consider the case when there are only two inputs, and we seek to maximize $q_1 + q_2$ subject to the constraint that $y_1 q_1 + y_2 q_2 \leq 1$. The *feasible region* is the set of all points lying inside the triangle formed by the intersection of the lines $q_1 \geq 0$, $q_2 \geq 0$ and $q_2 \leq -(y_1/y_2)q_1 + 1/y_2$. The optimal solution (q_1^*, q_2^*) lies on the line given by $q_2 = -q_1 + \Phi(x)$. To solve this problem geometrically, draw a line in the (q_1, q_2) space with slope equal to -1 and "push it" in the northeasterly direction until it is "tangent" to the feasible region. The value of the intercept will be $\Phi(x)$. You will see that as long as $y \neq 1$, then the tangent point will coincide with one of the two vertices (also known as extreme points of the feasible region) given by $(0, 1/y_2)$ and $(1/y_1, 0)$. (If $y = 1$, then all points on the line segment joining these two vertices will be optimal, but this does not negate the claim.) Now suppose $n > 2$. If the claim were not true, then there must exist at least two coordinates of q that are positive such that the corresponding y values do not both equal one. By restricting attention to these two coordinates, you can use the previous argument to show a contradiction.

6.5 We shall use the solution to Exercise 5.7, p. 92.

(a) With a budget of 9 the maximal output for the HR technology is 2. For this price vector, the line $c = 9$ intersects the output-cost set for the VRS technology at the midpoint of the line segment joining points $(2, 8)$ and $(3, 10)$. Hence, the maximal output $\Gamma(p, 9) = 2.5$. The boundary of the output-cost set for the CRS technology is determined by the line $c = (20/7)u$. The intersection of this line with the line $c = 9$ occurs at $((9/20)7, 9) = (3.15, 9)$, and so the maximal output $\Gamma(p, 9) = 3.15$.

(b) With a budget of 35 the maximal output for the HR technology is 6. For this price vector, the line $c = 35$ intersects the output-cost set for the VRS technology at the midpoint of the line segment joining points $(6, 32)$ and $(7, 38)$. Hence, the maximal output $\Gamma(p, 35) = 6.5$. The boundary of the output-cost set for the CRS technology is determined by the line $c = (12/3)u$. The intersection of this line with the line $c = 35$ occurs at $((35/12)3, 35) = (8.75, 35)$, and so the maximal output $\Gamma(p, 9) = 8.75$.

6.6 The Lagrangian for the producer's maximization problem is

$$L(x, p, B) = \Phi(x) \; - \; \lambda(p \cdot x - B).$$

As a direct application of the Theorem of the Maximum F.2, p. 492,

$$\frac{\partial \Gamma}{\partial p_i} = \frac{\partial L(x^*, p, B)}{\partial p_i} = -\lambda x_i(p, B),$$
$$\frac{\partial \Gamma}{\partial B} = \frac{\partial L(x^*, p, B)}{\partial B} = \lambda.$$

Now use the fact that $\lambda > 0$ since $\Phi(\cdot)$ is increasing.

6.7 (a) Consider a two-input, single-output technology described by two input possibility sets given by $L(1) = \mathcal{IFDH}((1,1)) = (1,1) + \mathbb{R}^2_+$ and $L(2) = \mathcal{IFDH}((2,2)) = (2,2) + \mathbb{R}^2_+$. Set the price vector $p = (1,1)$ and a budget $B = 1.5$. Here, $\Gamma(p,B) = 1$ but $Q(1,p) = 1 < 1.5$.

(b) Consider a two-input, single-output technology described by two input possibility sets $L(1) = \mathcal{IFDH}(Conv\{(1,3),(3,1)\}) = Conv\{(1,3),(3,1)\} + \mathbb{R}^2_+$ and $L(2) = \mathcal{IFDH}((2,2)) = (2,2) + \mathbb{R}^2_+$. (The input possibility set $L(1)$ corresponds to the unit input possibility set associated with a general Leontief technology with two "atoms" or intensity vectors given by $a^1 = (1,3)$ and $a^2 = (3,1)$.) Set the price vector $p = (1,1)$ and an output rate $u = 1$. Here, $Q(1,p) = 4$ but $\Gamma(p,4) = 2 > 1$.

6.8 Let $s = \sup_{x \in C} \Phi(x)$. We shall first show that $s < \infty$. If not, then it is possible to find $x_n \in C$ such that $\Phi(x_n) \geq n$ for all $n = 1, 2, \ldots$. Since the x_n belong to a compact set, it is possible to extract a convergent subsequence. Without changing notation for the subsequence, we have $x_n \to x$. Pick a positive real number u. Eventually $\Phi(x_n) \geq u$, and since $L^{\geq}_{\Phi}(u)$ is closed, it then follows that $x \in L^{\geq}_{\Phi}(u)$ or that $\Phi(x) \geq u$. As u was chosen arbitrarily, this in turn implies that $\Phi(x) = \infty$, contradicting the finiteness of $\Phi(\cdot)$. By definition of supremum and the fact that $s < \infty$, there exist $z_n \in C$ such that $\Phi(z_n) \geq s(1 - 1/n)$, $n = 1, 2, \ldots$. Extract a convergent sequence and let $z \in C$ denote the limit point. Pick $\epsilon > 0$. Eventually $\Phi(z_n) \geq s(1-\epsilon)$. Once again, since $L^{\geq}_{\Phi}(u(1-\epsilon))$ is closed, it follows that $z \in L^{\geq}_{\Phi}(u(1-\epsilon))$, too, or that $\Phi(z) \geq u(1-\epsilon)$. As ϵ was chosen arbitrarily, it follows that $\Phi(z) \geq u$. This implies that $\Phi(z) = u$, and the claim has been established.

6.9 The previous exercise establishes that for each b there exists an $x \in C$ such that $\nu(b) = \Phi(x) < \infty$. Pick b_1, b_2, $\lambda \in [0,1]$ and pick x_i, $i = 1, 2$, such that $\Phi(x_i) = \nu(b_i)$. Since $Ax_i \leq b_i$ for each i, $A(\lambda x_1 + (1-\lambda)x_2) \leq \lambda b_1 + (1-\lambda)b_2$. Thus, $\lambda x_1 + (1-\lambda)x_2$ is feasible for the problem defined by $\nu(\lambda b_1 + (1-\lambda)b_2)$. Given this and the concavity of $\Phi(\cdot)$,

$$\begin{aligned} \nu(\lambda b_1 + (1-\lambda)b_2) &\geq \Phi(\lambda x_1 + (1-\lambda)x_2) \\ &\geq \lambda \Phi(x_1) + (1-\lambda)\Phi(x_2) \\ &= \nu(b_1) + \nu(b_2). \end{aligned}$$

This establishes the concavity of $\nu(\cdot)$, as required.

7

Distance Functions

A distance function is a remarkably simple yet powerful concept. This function underpins nonparametric efficiency analysis, and is also the basis for a nonparametric approach for assessing productivity that we will discuss in Part III. The distance and cost function are linked via two, symmetric identities; knowledge of one function is sufficient to determine the other one.

7.1 Definition

7.1.1 Input Distance Function

Let $\mathcal{R}(x) := \{sx : s \geq 0\}$ denote the ray emanating from the origin and passing through the point $x \in \mathbb{R}^k_+$.

Definition 7.1. *The* **input distance function** $\mathcal{D} : \mathbb{R}^n_+ \times \mathbb{R}^m_+ \to \mathbb{R}_+$ *associated with a technology set* $\mathcal{T}$ *is*

$$\mathcal{D}(x,y) := \begin{cases} +\infty & \text{if } y = 0, \\ \max\{s : x/s \in L(y)\} & \text{if } \mathcal{R}(x) \cap L(y) \neq \emptyset \text{ and } y \neq 0, \\ 0 & \text{if } \mathcal{R}(x) \cap L(y) = \emptyset. \end{cases} \quad (7.1)$$

For notational convenience, we suppress the functional dependence of $\mathcal{D}(\cdot,\cdot)$ on the technology set $\mathcal{T}$. The input distance function measures how much x has to be scaled down (or up) to place it on the boundary of the input possibility set $L(y)$.[1] This interpretation only makes sense when the ray $\mathcal{R}(x)$ actually intersects $L(y)$.[2] The distance is set to zero when the ray does not intersect $L(y)$. The distance is set to $+\infty$ when $y = 0$ since $L(0) = \mathbb{R}^n_+$.

[1] A point lies on the boundary of a set if each neighborhood intersects the set and its complement.

[2] The maximum is achieved since $L(y)$ is closed.

Example 7.2. In Figure 7.1, the distance $\mathcal{D}(x_1, u) = 1.25$ and the distance $\mathcal{D}(x_2, u) = 0.75$. For the point $x_3 = (1, 0)$ (not shown) the distance would be $\mathcal{D}(x_3, u) = 0$.

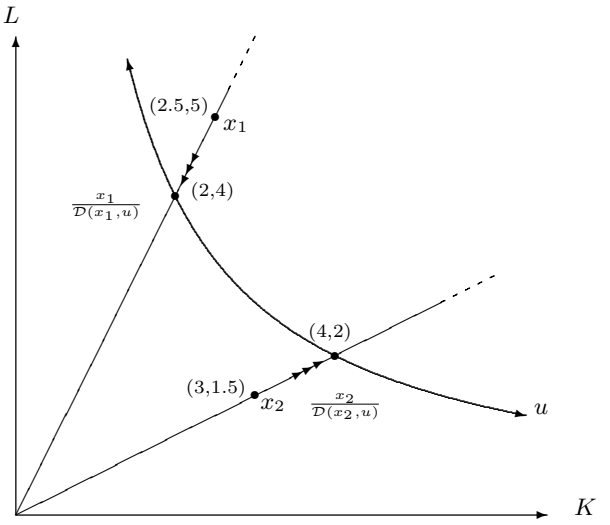


Fig. 7.1. Calculation of the input distance.

Keep in mind the definition of distance applies to the multi-output setting. Since the vector y merely identifies the input possibility set, we shall sometimes drop the symbol "y" and simply write $\mathcal{D}(x)$.

The input distance function completely characterizes the technology. First, it directly follows from its definition that

$$L(y) = \{x : \mathcal{D}(x, y) \geq 1\}.^3 \tag{7.2}$$

Next, it characterizes the boundary of each $L(y)$ since the boundary is $\{x : \mathcal{D}(x, y) = 1\}$. Finally, $x_0 \notin L(y_0)$ if and only if $\mathcal{D}(x_0, y_0) < 1$.

7.1.2 Output Distance Function

Definition 7.3. *The* **output distance function** $\mathcal{O} : \mathbb{R}^n_+ \times \mathbb{R}^m_+ \to \mathbb{R}_+$ *associated with a technology set* $\mathcal{T}$ *is*

[3] In terms of the input distance function, $P(x) = \{y : \mathcal{D}(x, y) \geq 1\}$ and $\mathcal{T} = \{(x, y) : \mathcal{D}(x, y) \geq 1\}$.

$$\mathcal{O}(x,y) := \begin{cases} +\infty & \text{if } x = 0, \\ \min\{s : y/s \in P(x)\} & \text{if } \mathcal{R}(y) \cap P(x) \neq \{0\} \text{ and } x \neq 0, \\ 0 & \text{if } P(x) = \{0\}. \end{cases} \tag{7.3}$$

For notational convenience, we suppress the functional dependence of $\mathcal{O}(\cdot,\cdot)$ on the technology set $\mathcal{T}$. The output distance function measures how much y has to be scaled down (or up) to place it on the boundary of the output possibility set $P(x)$. This interpretation only makes sense when the ray $\mathcal{R}(y)$ non-trivially intersects $P(x)$.[4] The distance is set to zero when $\mathcal{R}(y) \cap P(x) = \{0\}$. The distance is set to $+\infty$ when $x = 0$ since $P(0) = \mathbb{R}^m_+$.

Remark 7.4. In the single-output setting, the scaled output $u/\mathcal{O}(x,u)$ represents the maximum output the input vector x can achieve.

The output distance function completely characterizes the technology. First, it follows from its definition that

$$P(x) = \{y : \mathcal{O}(x,y) \leq 1\}.^5 \tag{7.4}$$

Next, it characterizes the boundary of each $P(x)$ since the boundary is $\{y : \mathcal{O}(x,y) = 1\}$. Finally, $y_0 \notin P(x_0)$ if and only if $\mathcal{O}(x_0,y_0) > 1$.

Example 7.5. Figure 7.2 graphically portrays the calculation of the output distance for the two-output case. Since $y_1 \notin P(x)$, the output distance $\mathcal{O}(x,y_1) > 1$, whereas the output distance $\mathcal{O}(x,y_2) < 1$ since $y_2 \in P(x)$. Notice how the boundary of the output possibility set $P(x)$, namely, its *Efficient Frontier*, is defined by a *concave* function.

7.2 Properties

Proposition 7.6. *The input distance function satisfies the following properties:*

a) Linearly homogeneous in x.
b) Nondecreasing in x.
c) Super-additive in x: $\mathcal{D}(x_1 + x_2) \geq \mathcal{D}(x_1) + \mathcal{D}(x_2)$.
d) Concave in x.
e) Continuous in x.

[4] The minimum is achieved since $P(x)$ is closed.
[5] In terms of the output distance function, $L(y) = \{x : \mathcal{O}(x,y) \leq 1\}$ and $\mathcal{T} = \{(x,y) : \mathcal{O}(x,y) \leq 1\}$.

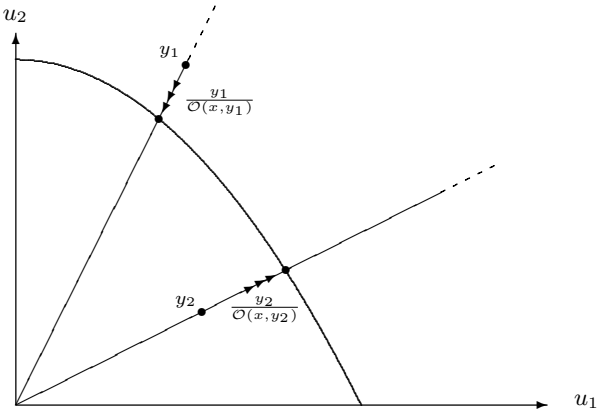


Fig. 7.2. Calculation of the output distance.

Proof. Parts (a) and (b) are immediate consequences of the definition. As for part (c), if either $\mathcal{D}(x_1)$ or $\mathcal{D}(x_2) = 0$, then super-additivity follows from part (b). Suppose then that both $\mathcal{D}(x_1)$ and $\mathcal{D}(x_2)$ are positive. $L(1)$ is convex, and so

$$z := \frac{\mathcal{D}(x_1)}{\mathcal{D}(x_1) + \mathcal{D}(x_2)} \frac{x_1}{\mathcal{D}(x_1)} + \frac{\mathcal{D}(x_2)}{\mathcal{D}(x_1) + \mathcal{D}(x_2)} \frac{x_2}{\mathcal{D}(x_2)} \in L(1).$$

Thus,

$$1 \leq \mathcal{D}(z) = \mathcal{D}\Big(\frac{x_1 + x_2}{\mathcal{D}(x_1) + \mathcal{D}(x_2)}\Big),$$

and the result now follows from part (a). Part (d) is a direct consequence of parts (a) and (c). Part (e) is proved in the Appendix to this Chapter. □

Proposition 7.7. *The output distance function satisfies the following properties.*

a) Linearly homogeneous in y.
b) Nondecreasing in y.
c) Sub-additive in y: $\mathcal{O}(y_1 + y_2) \leq \mathcal{O}(y_1) + \mathcal{O}(y_2)$.
d) Convex in y.
e) Continuous in y.

The proof mirrors the proof of Proposition 7.6 and is omitted.

7.3 Efficiency and Cost

Distance functions are intimately associated with the most common measures of efficiency. Suppose a firm uses input vector x_0 to produce output rate u_0. In

Chapter 2, we discussed how the production function can be used to measure radial input efficiency as

$$\theta^* := \min\{\theta : \Phi(\theta x_0) \geq u_0\},$$

which can be equivalently expressed as

$$\theta^* := \min\{\theta : \mathcal{D}(\theta x_0, u_0) \geq 1\}.$$

Since $\mathcal{D}(\cdot, u_0)$ is linearly homogenous, it immediately follows that

$$\theta^* = \min\{\theta : \ \mathcal{D}(x_0, u_0) \geq 1/\theta\} = \min\{\theta : \ \theta \geq \mathcal{D}(x_0, u_0)^{-1}\} = \mathcal{D}(x_0, u_0)^{-1}.$$

Consequently, $\mathcal{D}(x_0, u_0)$ is the reciprocal of the radial input efficiency for an $x_0 \in L(u_0)$.

Radial input efficiency can be measured when there are *multiple* outputs. That is, it is possible to define $\mathcal{RI}(x, y) := \min\{\theta : \theta x \in L(y)\}$. It only makes sense to define radial input efficiency when the ray $\mathcal{R}(x) \cap L(u) \neq \emptyset$. By the same reasoning as above, $\mathcal{RI}(x, y) = \mathcal{D}(x, y)^{-1}$. Analogously, the obvious definition of radial output efficiency is given by the output distance.

Remark 7.8. Keep in mind the input distance can be less than one and the output distance can be greater than one. While the reciprocal of the input or output distance in this case can not be interpreted as an input or output efficiency, we shall use these distance functions in Chapter 14 to assess a firm's input or output productivity change over time.

Remark 7.9. We have shown that the distance function coincides with the radial input efficiency measure. Continuity of the distance function ensures that efficiency measurement will not exhibit "jumps," which would be most undesirable!

The distance function is also intimately related to the cost function. At the end of Chapter 5, we showed how the cost function could be applied to the question of how to determine whether $x_0 \notin L(u_0)$. Suppose each input vector $x_i \in \mathbb{R}^n_+$, $i = 1, 2, \ldots, N$, produces at least output rate u_0. Assuming the HR technology, and that $x_0 > 0$, the optimal value of the linear program

$$\mu^*(x_0) := \min\{p \cdot x_0 : \ p \cdot x_i \geq 1 \text{ for all i}, \ p \geq 0\} \tag{7.5}$$

was shown to be less than one if and only if $x_0 \notin L(u_0)$. Using the concept of the distance function, we know that $x_0 \notin L(u_0)$ if and only if $\mathcal{D}(x_0, u_0) < 1$. Since both $\mu^*(\cdot)$ and $\mathcal{D}(\cdot, u_0)$ are linearly homogeneous, it follows that $\mu^* = \mathcal{D}(x_0, u_0)$. Indeed, the dual linear program to (7.5) is[6]

[6] Given a primal linear program expressed as $\min\{c \cdot x : \ Ax \geq b, \ x \geq 0\}$, its dual linear program is $\max\{b \cdot y : \ y^T A \leq c^T, \ y \geq 0\}$ Linear programming duality states that if both of these two linear programs are feasible, then they both have optimal solutions and their optimal values coincide.

$$\mu^*(x_0) = \max\left\{\sum_i \mu_i : \sum_i \mu_i x_i \le x_0,\ \mu_i \ge 0 \text{ for all } i\right\}. \tag{7.6}$$

Let $\theta := (\sum_i \mu_i)^{-1}$ and let $\lambda_i = \mu_i / \sum_i \mu_i$ for each i.[7] Since

$$\sum_i \Big(\frac{\mu_i}{\sum_i \mu_i}\Big) x_i \le \Big(\frac{1}{\sum_i \mu_i}\Big) x_0,$$

it follows that

$$\begin{aligned}
\mu^*(x_0) &= \left[\min\{\theta : \sum_i \lambda_i x_i \le \theta x_0,\ \sum_i \lambda_i = 1,\ \lambda_i \ge 0 \text{ for all } i\}\right]^{-1} \\
&= [\min\{\theta : \theta x_0 \in L(u_0)\}]^{-1} \\
&= [\min\{\theta : \mathcal{D}(\theta x_0, u_0) \ge 1\}]^{-1} \\
&= [\min\{\theta : \mathcal{D}(x_0, u_0) \ge 1/\theta\}]^{-1} \\
&= \mathcal{D}(x_0, u_0).
\end{aligned}$$

We see that the cost and distance function perspectives generate two dual linear programs to answer the same question. There is a deeper connection between the cost and distance functions, to which we now turn.

7.4 Reconstructing the Input Distance Function from the Cost Function

Since both $\mathcal{D}(x, y)$ and $Q(y, p)$ characterize the boundary of $L(y)$, perhaps it should not be too surprising that these functions are closely related. (The previous section already demonstrated a connection via the dual linear programs.) As R.W. Shephard originally demonstrated, the distance and cost functions determine each other through the following *symmetric* identities:

$$Q(y, p) = \min_{x \ge 0}\{p \cdot x : \mathcal{D}(x, y) \ge 1\} \tag{7.7}$$

$$\mathcal{D}(x, y) = \inf_{p \ge 0}\{p \cdot x : Q(y, p) \ge 1\}. \tag{7.8}$$

The first identity is a direct consequence of the definitions of the distance and cost functions. It is the second identity, first proposed and proved by Shephard, that requires proof.

Here is Shephard's interpretation of the symmetric identities. Let

$$L_Q(y) := \{p : Q(y, p) \ge 1\},$$

and let $\mathcal{F}_Q$ denote the family of upper level sets

[7] The sum of the μ_i will not be zero since $x_0 > 0$.

$$\mathcal{F}_Q := \{L_Q(y),\ y \geq 0\}.$$

Shephard showed that for each y the function $Q(y, \cdot)$ has the same properties (linearly homogeneous, nondecreasing, concave, and continuous) as the distance function. Interpreting $Q(y, p)$ as a distance function, $\mathcal{F}_Q$ is a well-behaved technology in the *price space*. Naturally, the technology $\mathcal{F}_Q$ will have its own cost function. Interpreting the symbol x as a price vector for the price space, Shephard's duality theorem states that the distance function for the original technology $\mathcal{T}$ (in the input space) is the *cost function* for the technology $\mathcal{F}_Q$ (in the price space). Most important, knowledge of either $\mathcal{T}$ or $\mathcal{F}_Q$ is sufficient to specify *both* technologies. Technology $\mathcal{F}_Q$ in the price space is referred to as the *dual technology*. Under suitable conditions, therefore, the *"dual technology of the dual technology is the original or primal technology."*

Shephard's dual identity (7.8) is largely a consequence of the geometry of closed, convex and input free disposable subsets of $\mathbb{R}^n_+$. In fact, the dual identity will be an immediate consequence of two Propositions whose geometry we motivate below.[8]

Let L denote a closed, convex, and input free disposable subset of $\mathbb{R}^n_+$. We begin by examining the nature of supporting hyperplanes to points that lie on the boundary of L.

Proposition 7.10. *Let x be a point on the boundary of L. For every $\delta > 0$, there exists a $p^\delta \in \mathbb{R}^n_+$ for which (i) $L \subset H^{\geq}(p^\delta, 1)$ and (ii) $p^\delta \cdot x < 1 + \delta$.*

Here is a geometrical interpretation. The well-known supporting hyperplane theorem guarantees that each point x on the boundary of L has a supporting hyperplane $H(p, Q)$. The input free disposability of L implies that both p and Q are nonnegative. Now suppose Q is positive and let $\hat{p} := p/Q$. It readily follows that $L \subset H^{\geq}(\hat{p}, 1)$ and $\hat{p} \cdot x = 1$.[9] Hence,

$$\inf_{\{p \geq 0:\ L \subset H^{\geq}(p,1)\}} p \cdot x = 1, \tag{7.9}$$

since $p \cdot x \geq 1$ whenever $L \subset H^{\geq}(p, 1)$.

The assumption that $Q > 0$ is equivalent to assuming that the hyperplane $H(p, Q)$ separates L from the origin. It is, however, possible that a point x on the boundary of L has *no* supporting hyperplane that separates L from the origin.

Example 7.11. Take L to be the smallest closed, convex, and input free disposable set containing the points $x_n := (1 - 1/n, 1/2^n)$, $n = 2, 3, \ldots$, i.e., $L := \mathcal{IFDH}(\{x_n\})$. Since L is closed the limit point $(1, 0)$ lies in L. The *only* support for $(1, 0)$ is of the form $((0, p_2), 0)$.

Note that the vector x in Example 7.11 has at least one zero component; that is, the index set $\mathcal{I}(x) := \{i :\ x_i = 0\}$ is nonempty. Further, for each $\epsilon > 0$ and

[8] The proofs are somewhat technical—see the Appendix to this Chapter.

[9] Recall that $H^{\geq}(p, \alpha) := \{z :\ p \cdot z \geq \alpha\}$—see Definition C.3, p. 462.

$e := (\epsilon, \epsilon, \ldots, \epsilon) \in \mathbb{R}^n$, the point $(x+e)$ lies in the interior of L. The distance function ensures that a positive constant c exists for which $c(x+e)$ lies on the boundary of L. Since $c(x+e)$ is a strictly positive vector, a supporting hyperplane of L at $c(x+e)$ is of the form $H(p,1)$. If ϵ is sufficiently small, then $p \cdot x \approx 1$. Proposition 7.10 shows that even if x cannot be supported directly with a hyperplane that separates L from the origin, the identity (7.9) is still true.

We now turn our attention to those points x for which the ray $\mathcal{R}(x)$ does not intersect L.

Proposition 7.12. *If $\mathcal{R}(x) \cap L$ is empty, then*

$$\inf_{\{p \geq 0:\ L \subset H^{\geq}(p,1)\}} p \cdot x = 0. \tag{7.10}$$

A geometrical interpretation to Proposition 7.12 is provided by the following two examples.

Example 7.13. Consider the point $x = (0,1) \in \mathbb{R}^2$ and suppose that $L = (1,1) + \mathbb{R}^2_+$. Identity (7.10) holds since $p = (1,0)$ is feasible and obviously $p \cdot x = 0$.

Example 7.14. Consider the point $x = (0,1) \in \mathbb{R}^2$ and suppose that L corresponds to an input possibility set of the Cobb-Douglas technology $\Phi(K,L) = \sqrt{KL}$. If $L \subset H^{\geq}(p,1)$, then both prices p_K and p_L must be positive and so $p \cdot x \neq 0$. (If, for example, $p_L = 0$, then $L \subset H^{\geq}(p,1)$ implies that $K \geq 1/p_K$ for all $(K,L) \in L$, which clearly does not hold.) However, there do exist hyperplanes $H(p^n,1)$ for which $L \subset H^{\geq}(p^n,1)$ and $p^n \cdot x \to 0$, which is what the Proposition guarantees.

Let $\mathcal{D}^*(x,y)$ denote the value obtained on the right-hand side of (7.8).

Theorem 7.15. Shephard's Duality $\mathcal{D}^*(x,y) = \mathcal{D}(x,y)$.

Proof. Fix $x \in \mathbb{R}^n_+$ and $y \in \mathbb{R}^m_+$, $y \neq 0$. The identity follows immediately from Propositions (7.10) and (7.12), since

(i) $Q(y,p) \geq 1$ if and only if $L(y) \subset H^{\geq}(p,1)$, and
(ii) if $\mathcal{R}(x) \cap L(y) \neq \emptyset$, then $x/\mathcal{D}(x,y)$ lies on the boundary of $L(y)$. □

Remark 7.16. Identity (7.9) establishes Shephard's dual identity (7.8) when the input vector $x > 0$ and $\mathcal{R}(x) \cap L(y) \neq \emptyset$, since in this case $x/\mathcal{D}(x,y)$ will belong to the boundary of $L(y)$. A supporting hyperplane of $L(y)$ at $x/\mathcal{D}(x,y)$ necessarily has a positive Q value.

7.5 Application to Homothetic Technologies

It turns out that whenever the cost (or distance) function *factors*, that is, it is multiplicatively separable as defined in Remark 5.6, p. 74, then the production function must be homothetic. *Thus, the property of factorability is equivalent to the property of homotheticity.* We now establish this important fact using Shephard's duality theorem.

We begin by assuming the cost function $Q(u, p)$ can be represented as $f(u)P(p)$. (Keep in mind $u \in \mathbb{R}_+$.) Here, the function $f(\cdot)$ is the inverse of a transform $F(\cdot)$ with its assumed properties, i.e., $f(\cdot) = F^{-1}(\cdot)$. Using Shephard's dual identity (7.8) and the factorability of the cost function, the distance function factors, too, as shown by the following chain of equalities:

$$\begin{aligned}
\mathcal{D}(x, u) &= \inf\{p \cdot x : \ P(p)f(u) \geq 1\} \\
&= \left(\frac{1}{f(u)}\right) \ \inf\Big\{(f(u)p) \cdot x : \ P(f(u)p) \geq 1\Big\} \\
&= \left(\frac{1}{f(u)}\right) \ \inf\{\hat{p} \cdot x : \ P(\hat{p}) \geq 1\} \\
&= \frac{f(1)}{f(u)} \, \mathcal{D}(x, 1). \qquad (7.11)
\end{aligned}$$

Now using (7.11) and the relationship between $\mathcal{D}(\cdot, \cdot)$ and $\Phi(\cdot)$, the following identities are established:

$$\begin{aligned}
\Phi(x) &= \max\{u : \ \mathcal{D}(x, u) \geq 1\} \\
&= \max\{u : \ f(1)\mathcal{D}(x, 1) \geq f(u)\} \\
&= \max\{u : \ F(f(1)\mathcal{D}(x, 1)) \geq u\} \\
&= F(f(1)\mathcal{D}(x, 1)) \\
&= F(\phi(x)).
\end{aligned}$$

Conversely, if the production function is homothetic, then the cost function must factor; we leave the proof of this fact as an exercise.

As a direct consequence of (7.11),

$$\begin{aligned}
L(u) &= \{x : \ \mathcal{D}(x, u) \geq 1\} \\
&= \frac{f(u)}{f(1)} \ \left\{\frac{f(1)}{f(u)} \, x : \ \mathcal{D}\left(\frac{f(1)}{f(u)} \, x, 1\right) \geq 1\right\} \\
&= \frac{f(u)}{f(1)} \, L(1).
\end{aligned}$$

This shows that the scaling properties are independent of x, which we have previously established for homothetic technologies.

7.6 Appendix

Proof of continuity of the distance function

We begin by establishing the following lemma.

Lemma 7.17. *Let $x \in \mathbb{R}^n_+$ be non-zero. For each positive integer m define $y^m \in \mathbb{R}^n_+$ by $y^m_i = x_i + 1/m$ for each i. For each positive integer m and each i define $z^m \in \mathbb{R}^n_+$ by $z^m_i = x_i - 1/m$ if $x_i > 0$ or $z^m_i = 0$ if $x_i = 0$.*[10]

a) If $\mathcal{D}(x) = 0$, then $\mathcal{D}(y^m) \to 0$.
b) If $\mathcal{D}(x) > 0$, then for each $\alpha > 0$ eventually $\mathcal{D}(y^m) < \mathcal{D}(x)(1+\alpha)$.
c) If $\mathcal{D}(x) > 0$, then for each $\alpha > 0$ eventually $\mathcal{D}(z^m) > \mathcal{D}(x)(1-\alpha)$.

Proof. Let m^* denote the first integer m for which z^m is nonnegative. It is understood below that only those z^m's for which $m \geq m^*$ will be considered.

Part (a). If the claim is false, then a positive ϵ and a subsequence $\{y^{m_k}\}$ exists for which $\mathcal{D}(y^{n_k}) \geq \epsilon$. Since $y^{n_k} \leq y^1$ for each k and $\mathcal{D}$ is nondecreasing, $\{\mathcal{D}(y^{n_k})\} \subset [\epsilon, \mathcal{D}(y^1)]$, a compact set. Thus, it is possible to extract a convergent subsequence $\mathcal{D}(y^{n_k}) \to \rho$. (We will not change notation for the subsequence.) Obviously, ρ is positive. We then have

$$\frac{y^{n_k}}{\mathcal{D}(y^{n_k})} \to x/\rho. \tag{7.12}$$

Since each $y^{n_k}/\mathcal{D}(y^{n_k}) \in L(y)$ and $L(y)$ is closed, $x/\rho \in L(y)$, which immediately implies that $\mathcal{D}(x)$ is positive, a contradiction.

Part (b). By similar reasoning as in the proof of part (a), if the claim is false, it will be possible to extract a convergent subsequence $\{y^{n_k}\}$ for which (7.12) holds with $\rho \geq \mathcal{D}(x)(1+\alpha)$. Since $x/\rho \in L(y)$ it follows that $\rho \geq \mathcal{D}(x)$, a contradiction.

Part (c). The definition of $\mathcal{D}(\cdot)$ ensures that $\mathcal{D}(z^{m^*})$ is positive. The proof now mirrors the proof of part (b). □

Now on to the proof of continuity.

Proof. We need to show for each $\epsilon > 0$ there exists a $\delta > 0$ for which

$$|\mathcal{D}(h) - \mathcal{D}(x)| < \epsilon \text{ whenever } ||h - x|| < \delta. \tag{7.13}$$

When $\mathcal{D}(x) = 0$, (7.13) follows from Lemma 7.17(a) and the fact that $||y^n - x|| = 1/n$.

Now suppose $\mathcal{D}(x) > 0$. Fix $\epsilon > 0$. Set $\alpha = \epsilon/\mathcal{D}(x)$. From Lemma 7.17(b, c), we know we can find an m large enough, say M, for which $\mathcal{D}(x)(1-\alpha) < \mathcal{D}(z^M)$ and $\mathcal{D}(y^M) < \mathcal{D}(x)(1+\alpha)$. If $||h - x|| \leq 1/M$, then $z^M \leq h \leq y^M$, from which it follows that $\mathcal{D}(z^M) \leq \mathcal{D}(h) \leq \mathcal{D}(y^M)$. Thus, $|\mathcal{D}(h) - \mathcal{D}(x)| < \alpha\mathcal{D}(x) = \epsilon$ whenever $||h - x|| \leq 1/M$, and so the result follows by setting $\delta = 1/M$. □

[10] We suppress the functional dependence of the y^m and the z^m on x.

Remark 7.18. We showed that $\mathcal{D}(\cdot, u)$ is concave. It is well known that a concave function is continuous on the interior of its domain. Here, we established $\mathcal{D}(\cdot, u)$ is continuous on all of $\mathbb{R}^n_+$ (minus the origin). Note the proof of continuity did not use the convexity of the input possibility set.

Remark 7.19. We established continuity of $\mathcal{D}(\cdot, y)$ when y is *fixed*. It would be desirable to ensure that $\mathcal{D}(\cdot, \cdot)$ is jointly continuous in *both* y and x. To make this notion precise requires a definition of how one measures the degree of closeness of one technology to another. (Recall that a technology can be represented as a *family* of sets.) Formally, one needs to impose a suitable topology on a suitable space of sets. Under a few reasonable assumptions, $\mathcal{D}(x, y)$ will be jointly continuous under the *topology of closed convergence*.

Proof of Proposition 7.10

For each $x \in \mathbb{R}^n_+$ let $\mathcal{I}(x) := \{i : \ x_i = 0\}$.

Proof. Fix $y \in \mathbb{R}^m_+$, $y \neq 0$. Since $\mathcal{D}(\cdot)$ is continuous, for each $\gamma > 0$ an $\epsilon > 0$ exists for which

$$\hat{x} := \frac{x + e}{\mathcal{D}(x + e)} \in B_\gamma(x),$$

the open ball of radius γ about x. As $\hat{x}$ is strictly positive a $p^\epsilon \in \mathbb{R}^n_+$ exists for which $0 < p^\epsilon \cdot \hat{x} = 1$. We have

$$1 = \sum_i p^\epsilon_i \cdot \hat{x}_i \geq \sum_{i \notin \mathcal{I}(x)} p^\epsilon_i \cdot \hat{x}_i \geq \sum_{i \notin \mathcal{I}(x)} p^\epsilon_i \cdot \frac{\min_{i \notin \mathcal{I}(x)} x_i + \epsilon}{\mathcal{D}(x + e)} \ . \tag{7.14}$$

Let

$$f(x, \epsilon) := \frac{\mathcal{D}(x + e)}{\min_{i \notin \mathcal{I}(x)} x_i + \epsilon} \ .$$

It follows from (7.14) that

$$p^\epsilon \cdot (x - \hat{x}) \leq \sum_{i \notin \mathcal{I}(x)} p^\epsilon_i (x_i - \hat{x}_i) \leq f(x, \epsilon)\gamma \ . \tag{7.15}$$

Since $\epsilon \to 0$ as $\gamma \to 0$, a sufficiently small positive $\epsilon(\delta)$ exists for which $p^{\epsilon(\delta)} \cdot x < 1 + \delta$, as required. □

Proof of Proposition 7.12

Proof. Define $p^n \in \mathbb{R}^n_+$ by setting $p^n_i = n$ if $i \in \mathcal{I}(x)$ or by setting $p^n_i = (\sum_{i \notin \mathcal{I}(x)} x_i)^{-1}$ otherwise. Note that $p^n \cdot x = 1$ for each n. Since p^n is strictly positive a z^n exists for which $(p^n, p^n \cdot z^n)$ supports L at z^n.

We claim that $p^n \cdot z^n \to \infty$. If this were not the case, then $z^n_i \to 0$ if $i \in \mathcal{I}(x)$. Moreover, $\{\sum_{i \notin \mathcal{I}(x)} z^n_i\}$ would be bounded. This in turn implies

that the sequence of z^n's is bounded, and therefore possesses a convergent subsequence. Let $\hat{z}$ denote a limit point. Since L is closed, $\hat{z} \in L$. In addition, $\hat{z}_i = 0$ if $i \in \mathcal{I}(x)$. Thus, for s sufficiently large, $sx \geq \hat{z}$, which would imply that $R(x) \cap L$ is nonempty, a contradiction.

Now set $\hat{p}^n := p^n/(p^n \cdot z^n)$, and observe that $L \subset H^{\geq}(\hat{p}^n, 1)$ and $\hat{p}^n \cdot z \to 0$, as required. □

7.7 Exercises

7.1. Consider the input-output data set depicted in Figure 4.4 on p. 61. Let $x = (1, 3)$.

(a) Determine $\mathcal{D}^{HR}(x, 20)$ for the HR technology.
(b) Determine $\mathcal{D}^{CRS}(x, 20)$ for the CRS model of technology.
(c) Determine $\mathcal{D}^{VRS}(x, 20)$ for the VRS model of technology.
(d) What is the relationship between the answers in (a)-(c)? Explain why this is so.

7.2. Explicitly define the p_n described in Example 7.14.

7.3. Show that if the production function $\Phi(\cdot)$ is homothetic, then the cost function factors.

7.4. Suppose the cost function is $Q(u, p) = u(a_1 p_1 + a_2 p_2 + A\sqrt{p_1 p_2})$.

(a) Use Shephard's duality theorem to show that $\Phi(x) = \mathcal{D}(x, 1)$.
(b) Use Shephard's duality theorem to obtain a closed-form solution for $\mathcal{D}(x, 1)$.

7.5. The input distance function $\mathcal{D}(\cdot, u)$ is concave in x for each fixed u. Either prove that the input distance function is *jointly* concave in both x and u or provide a concrete counter-example.

7.8 Bibliographical Notes

Shephard's [1970] monograph contains a full chapter on the distance function. The proof of the general duality between the distance and the cost function is based on Hackman [1986].

7.9 Solutions to Exercises

7.1 (a) By definition $x/\mathcal{D}^{HR}(x,20)$ lies on the boundary of the input possibility set $L^{HR}(20)$, which is depicted in Figure 4.5, p. 62. The point x lies on the line $L = 3K$. This line intersects the line $L = -(1/3)K + 19/3$ joining points (1, 6) and (4, 5). The point of intersection is $(1.9, 5.7)$. Thus $\mathcal{D}^{HR}(x,20) = 1/1.9 = 0.526$.

(b) By definition $x/\mathcal{D}^{CRS}(x,20)$ lies on the boundary of the input possibility set $L^{CRS}(20)$, which is depicted in Figure 4.7, p. 64. As described in Example 4.18, p. 64, $L^{CRS}(20)$ is the convex, input free disposable hull of the points $\hat{x}_1, \ldots, \hat{x}_6$. The point x lies on the line $L = 3K$. This line intersects the line $L = -(5/11)K + 50/11$ joining points $\hat{x}_3 = (5/6, 25/6)$ and $\hat{x}_6 = (8/3, 10/3)$. The point of intersection is $(25/19, 75/19)$. Thus $\mathcal{D}^{CRS}(x,20) = 1/(25/19) = 0.76$.

(c) By definition $x/\mathcal{D}^{VRS}(x,20)$ lies on the boundary of the input possibility set $L^{VRS}(20)$, which is depicted in Figure 4.8, p. 65. As described in Example 4.19, p. 64, $L^{VRS}(20)$ is the convex, input free disposable hull of the points x_5, x_6 and the points in $G(20)$. The point x lies on the line $L = 3K$. This line intersects the line $L = -(43/19)K + 157/19$ joining points $x_5 = (1, 6)$ and $\hat{x}^{20}_{3,6} = (37/18, 65/18)$. The point of intersection is $(1.57, 4.71)$. Thus $\mathcal{D}^{VRS}(x,20) = 1/(1.57) = 0.637$.

(d) We have $\mathcal{D}^{CRS}(x,20) \leq \mathcal{D}^{VRS}(x,20) \leq \mathcal{D}^{HR}(x,20)$. This must be the case, as $L^{HR}(20) \subset L^{VRS}(20) \subset L^{CRS}(20)$.

7.2 We wish to find p_n such that (i) $L \subset H^{\geq}(p^n, 1)$ and (ii) $p^n \cdot x \to 0$ as $n \to \infty$. To meet the first requirement both prices must be positive. To meet the second requirement $p^n_L \to 0$ as $n \to \infty$. Since the first coordinate of x is 0, the value of p^n_K is irrelevant, which gives a degree of freedom. Since the price of labor here will be low, the amount of labor purchased will be high. Accordingly, pick $x^n = (1/n, n)$ as the cost-minimum input vectors corresponding to the (as-yet chosen) p^n. The rate of technical substitution is $-\Phi_K/\Phi_L = L/K$ for this technology, and it coincides with the ratio of the prices p^n_K/p^n_L at the cost minimum input vector. Thus, the choice $p^n = (n/2, 1/(2n))$ for the price vector will do the job. (The constant of 1/2 normalizes the price vector so that the minimum cost equals one.)

7.3 Let $\Phi(x) = F(\phi(x))$ where $F(\cdot)$ is a transform and $\phi(\cdot)$ is homogeneous of degree one. Let $f(\cdot) = F^{-1}(\cdot)$. We have

$$\begin{aligned}
Q(u,p) &= \min\{p \cdot x : \ F(\phi(x)) \geq u\} \\
&= \min\{p \cdot x : \ \phi(x) \geq f(u)\} \\
&= \min\left\{p \cdot x : \ \phi\left(\frac{x}{f(u)}\right) \geq 1\right\} \\
&= f(u)\min\{p \cdot y : \ \phi(y) \geq 1\} \\
&:= f(u)P(p).
\end{aligned}$$

7.4 The structure of the cost function implies that $\Phi(x) = 0$ if any coordinate of x is zero. (If not, it would be possible to find a price vector p such that $p \cdot x = 0$, which would imply that Q(u, p) = 0 for a positive u.) Assume then that x is strictly positive. (a) By definition of the input distance function

$$\Phi(x) = \max\{u : \mathcal{D}(x, u) \geq 1\}.$$

By Shephard's duality theorem,

$$\begin{aligned}\mathcal{D}(x, u) &= \inf\{p \cdot x : Q(u, p) \geq 1\} \\ &= \inf\{p \cdot x : u(a_1 p_1 + a_2 p_2 + A\sqrt{p_1 p_2}) \geq 1\} \\ &= (1/u)\inf\{p \cdot x : a_1 p_1 + a_2 p_2 + A\sqrt{p_1 p_2} \geq 1\} \\ &= (1/u)\mathcal{D}(x, 1).\end{aligned}$$

Thus,

$$\Phi(x) = \max\{u : \mathcal{D}(x, u) \geq 1\} = \max\{u : \mathcal{D}(x, 1) \geq u\} = \mathcal{D}(x, 1).$$

Remark 7.20. The cost function factors as $f(u)P(p)$ and so the technology is homothetic. Here, the function $f(u) = u$. For general homothetic technologies, we established on p. 117 that $\Phi(x) = F(f(1)\mathcal{D}(x, 1))$, which equals $\mathcal{D}(x, 1)$ for this particular $f(\cdot)$.

(b) We have

$$\mathcal{D}(x, 1) = \inf\{p \cdot x : a_1 p_1 + a_2 p_2 + A\sqrt{p_1 p_2} \geq 1\}.$$

First-order optimality conditions imply that $x_1 = \lambda(a_1 + A\sqrt{p_1/p_2})$ and $x_2 = \lambda(a_2 + A\sqrt{p_2/p_1})$. This in turn implies that $x_1(a_2 + A/\theta) = x_2(a_1 + A\theta)$ where $\theta := \sqrt{p_1/p_2}$. It then follows that

$$0 = (Ax_2)\theta^2 + (a_1 x_2 - a_2 x_1)\theta - Ax_1,$$

from which we may conclude that

$$\theta = \theta(x_1, x_2) = \frac{(a_1 x_2 - a_2 x_1) + \sqrt{(a_1 x_2 - a_2 x_1)^2 + 4A^2 x_1 x_2}}{2Ax_2}.$$

Since $p_1 = p_2\theta^2$ and $a_1 p_1 + a_2 p_2 + A\sqrt{p_1 p_2} = 1$, we have that $(a_1\theta^2 + a_2 + A\theta)p_2 = 1$ or

$$p_2 = p_2(\theta) = \frac{1}{(a_1\theta^2 + a_2 + A\theta)}, \quad p_1 = p_1(\theta) = \frac{\theta^2}{(a_1\theta^2 + a_2 + A\theta)}.$$

Finally, we have

$$\Phi(x) = \mathcal{D}(x, 1) = p_1 x_1 + p_2 x_2 = p_1(\theta(x_1, x_2))x_1 + p_2(\theta(x_1, x_2))x_2.$$

7.5 Consider the technology described by the following production function:

$$\Phi(x) = \begin{cases} x, & 0 \leq x < 2, \\ x + 1, & 2 \leq x < \infty. \end{cases}$$

The definition of input distance implies that $\mathcal{D}(1,1) = \mathcal{D}(3,4) = 1$. Let $\lambda = 0.6$ and consider the point in the (x,u) space given by $0.6(1,1) + 0.4(3,4) = (1.8, 2.2)$. To achieve output rate 2.2, it is necessary to achieve at least 3 with a minimal input of 2. Thus, $\mathcal{D}(1.8, 2.2) = 0.9$ (since $1.8/0.9 = 2$). We have provided an example in which

$$\mathcal{D}(\lambda x_1 + (1-\lambda)x_2, \lambda u_1 + (1-\lambda)u_2) < \lambda \mathcal{D}(x_1, u_1) + (1-\lambda)\mathcal{D}(x_2, u_2),$$

which obviously contradicts the defining property of concavity.

8

Nonconvex Models of Technology

As we have seen in earlier chapters, convexity is essential to the microeconomic theory of production. From a modeling perspective, convexity ensures that a particular kind of mixing of feasible actions is still feasible, which is often a reasonable assumption. Generally, convexity (in some form) is assumed to exploit its convenient (e.g. separation) properties. There have been efforts to generalize the concept of convexity, i.e., convex sets and quasiconcave/quasiconvex functions. These efforts can be categorized into two approaches. The first approach allows a separation of a point to the set by something other than a hyperplane. The second approach allows two points in the set to be connected by a more general path than a line segment or, from a modeling perspective, to allow a more general kind of mixing of feasible actions.

The first three sections of this chapter describe explicit examples of nonconvex models that arise in resource allocation, producer budgeting and Data Envelopment Analysis. The final section discusses a generalization of convexity, called *projective-convexity*, that encompasses the nonconvex models described herein and shares the benefits of both approaches to generalize convexity described above.

8.1 Resource Allocation

Consider the question of how to represent and measure technology of a sector (such as agriculture) when output is produced by more than one technique. To simplify the analysis, we consider only two techniques, whose technologies are modeled by production functions $f_1(K_1, L_1)$ and $f_2(K_2, L_2)$ with capital (K) and labor (L) as the factors of production. Both production functions are nondecreasing, continuous and quasiconcave. We assume that simultaneous production by both techniques is possible.

8.1.1 Aggregate Production Function

A producer must decide how to allocate aggregate capital, K, and aggregate labor, L, to respective techniques to maximize overall output, which we denote by $\Phi(K, L)$. The **aggregate production function** $\Phi(\cdot)$ is obtained as the value of the following allocation optimization problem:

$$\Phi(K, L) := \max\{f_1(K_1, L_1) + f_2(K_2, L_2) : \ K_1 + K_2 \leq K, \ L_1 + L_2 \leq L\}. \tag{8.1}$$

It is understood in (8.1) that the decision variables must be nonnegative.

Proposition 8.1. *If $f_1(\cdot)$ and $f_2(\cdot)$ exhibit constant returns-to-scale, then the aggregate production function $\Phi(\cdot)$ defined in (8.1) is quasiconcave.*

Proof. Clearly, $\Phi(\cdot)$ exhibits constant returns-to-scale, too, and so its technology is completely characterized by the unit input possibility set $L^{\geq}_{\Phi}(1)$. We shall show this set is convex by establishing that

$$L^{\geq}_{\Phi}(1) = Conv\left(L^{\geq}_{f_1}(1) \cup L^{\geq}_{f_2}(1)\right). \tag{8.2}$$

To this end, pick $(\hat{K}_i, \hat{L}_i) \in L^{\geq}_{f_i}(1)$, $i = 1, 2$, $\lambda \in [0, 1]$ and let

$$(K, L) = \lambda(\hat{K}_1, \hat{L}_1) + (1 - \lambda)(\hat{K}_2, \hat{L}_2).$$

The allocations $\hat{K}_1, \hat{K}_2, \hat{L}_1, \hat{L}_2$ are obviously feasible for (8.1) for this given choice of (K, L). Consequently,

$$\begin{aligned} \Phi(K, L) &\geq f_1(\lambda(\hat{K}_1, \hat{L}_1)) + f_2((1 - \lambda)(\hat{K}_2, \hat{L}_2)) \\ &= \lambda f_1(\hat{K}_1, \hat{L}_1) + (1 - \lambda) f_2(\hat{K}_2, \hat{L}_2) \\ &\geq \lambda + (1 - \lambda) = 1. \end{aligned}$$

The second line above follows since each $f_i(\cdot)$ is linearly homogeneous. Since $(K, L) \in L^{\geq}_{\Phi}(1)$, we have shown that

$$Conv\left(L^{\geq}_{f_1}(1) \cup L^{\geq}_{f_2}(1)\right) \subset L^{\geq}_{\Phi}(1). \tag{8.3}$$

To show the reverse inclusion

$$L^{\geq}_{\Phi}(1) \subset Conv\left(L^{\geq}_{f_1}(1) \cup L^{\geq}_{f_2}(1)\right), \tag{8.4}$$

pick a (K, L) for which $\Phi(K, L) = 1$, and let K_1, K_2, L_1, L_2 denote a feasible set of allocations such that

$$\Phi(K, L) = f_1(K_1, L_1) + f_2(K_2, L_2).$$

Let $u_i = f_i(K_i, L_i)$. First assume both u_i are positive. Let $(\hat{K}_i, \hat{L}_i) = (K_i, L_i)/u_i$, $i = 1, 2$. Clearly, $(K, L) = \sum_i u_i(\hat{K}_i, \hat{L}_i)$. Since $u_i \in (0, 1]$, $u_1 + u_2 = 1$ and $(\hat{K}_i, \hat{L}_i) \in L^{\geq}_{f_i}(1)$, it follows that

$$(K, L) \in Conv\Big(L^{\geq}_{f_1}(1) \cup L^{\geq}_{f_2}(1)\Big). \tag{8.5}$$

If either of the u_i is zero, then (8.5) follows immediately. We have therefore shown that the isoquant of $L^{\geq}_{\Phi}(1)$ is contained within $Conv(L^{\geq}_{f_1}(1) \cup L^{\geq}_{f_2}(1))$. The reverse inclusion (8.4) now follows, since $\Phi(\cdot)$ is homogeneous of degree one. □

There is another common setting in which $\Phi(\cdot)$ is guaranteed to be quasiconcave. When both production functions $f_1(\cdot)$, $f_2(\cdot)$ are *concave*, then it is easy to establish that $\Phi(\cdot)$ is concave, too, since the constraint set is linear. Is $\Phi(\cdot)$ always quasiconcave? The answer is no, as the following counter-example demonstrates.

8.1.2 Counter-Example to Quasiconcavity

We examine the following concrete example in which $f_1(K_1, L_1) = \sqrt{K_1}L_1$ and $f_2(K_2, L_2) = K_2\sqrt{L_2}$. Both production functions are Cobb-Douglas, exhibit *increasing* returns-to-scale and are symmetric in that their marginal returns on capital and labor are reversed but otherwise equal.

Fix (K, L) and let (K_i, L_i), $i = 1, 2$, denote optimal allocations to (8.1). Since the marginal return on capital is infinite for the first production function when $K_1 = 0$, and the marginal return on labor is infinite for the second production function when $L_2 = 0$, it would appear, at first blush, that it would always be optimal to allocate at least some positive amounts of capital and labor to both techniques. In fact, this intuition is false: it will always be the case that *only* one of the two techniques will be employed; that is, either $(K_1, L_1) = (0, 0)$ or $(K_2, L_2) = (0, 0)$. This will in turn imply that the input possibility sets of $\Phi(\cdot)$ are not convex, and hence $\Phi(\cdot)$ is not quasiconcave.

Suppose instead that all optimal allocations are positive. (It would never pay to allocate a positive amount of one input and a zero amount of the other input to a technique.) Let μ_K and μ_L denote the Lagrange multipliers for the capital and labor constraints, respectively. The first-order optimality conditions are

$$\frac{L_1}{2\sqrt{K_1}} = \mu_K = \sqrt{L_2},$$
$$\sqrt{K_1} = \mu_L = \frac{K_2}{2\sqrt{L_2}},$$

which imply that $K_2 = L_1$.

We shall first examine the case when $K \geq L$ (the analysis of the reverse inequality is symmetric). The allocation problem (8.1) reduces to the following one-dimensional optimization problem

$$\max_{0<L_1\leq L} \sqrt{K - L_1}\, L_1 + L_1\sqrt{L - L_1}. \tag{8.6}$$

Let $L_1 := \alpha L$ and $K := \alpha_0 L$. Note that $\alpha \in [0,1]$ and $\alpha_0 \geq 1$. Expressed in terms of α, the maximization problem in (8.6) becomes

$$M := L^{3/2} \max_{0<\alpha\leq 1} \xi(\alpha) \tag{8.7}$$

where

$$\xi(\alpha) := \alpha\Big(\sqrt{\alpha_0 - \alpha} + \sqrt{1-\alpha}\Big).$$

Let $g(x) := \sqrt{x}$. Since $g(\cdot)$ is strictly concave,

$$\sqrt{x} < (1+x)/2,\ x \neq 1.^1 \tag{8.8}$$

Since α is constrained to be positive, we may apply (8.8) to $\xi(\alpha)$ to establish that

$$q(\alpha) := \Big[\alpha(\frac{3+\alpha_0}{2} - \alpha)\Big] \ > \ \xi(\alpha).$$

The quadratic form $q(\cdot)$ is concave and achieves its *unconstrained* maximum at $\alpha^* := (3+\alpha_0)/4$. The value for α^* is at least as large as 1 since $\alpha_0 \geq 1$. The function $q(\cdot)$ is an increasing function on $[0, \alpha^*]$, and so the constrained maximum of q on $[0,1]$ is achieved at 1. As a direct consequence,

$$M < L^{3/2} q(1) = L^{3/2}\ \frac{1+\alpha_0}{2} \leq \alpha_0 L^{3/2} = K\sqrt{L}.$$

The value $K\sqrt{L}$ is obtained by allocating all resources to the second technique, which cannot be larger than M. Our original supposition, namely that all allocations were positive, must be false. We conclude, therefore, that when $K \geq L$ the optimal allocation of aggregate resources will be to allocate all resources to the second technique.

When $L \geq K$, one defines α and α_0 as $K_1 = \alpha K$ and $\alpha_0 = L/K$, and the arguments above (with $K^{3/2}$ replacing $L^{3/2}$) show that when $L \geq K$, the optimal allocation of aggregate resources will be to allocate all resources to the first technique.

In sum, we have shown that $\Phi(K,L) = f_2(K,L)$ when $K \geq L$ and $\Phi(K,L) = f_1(K,L)$ when $K \leq L$. In particular,

$$\Phi(K,L) = \max\{f_1(K,L),\ f_2(K,L)\}.$$

In terms of the unit input possibility sets,

$$L_\Phi(1) = \bigcup_i L_{f_i}(1).$$

It is the *union* of two convex sets, which is definitely *not* convex (for these particular convex sets).

[1] Strict concavity implies that $g(x) < g(1) + g'(1)(x-1)$ holds for all $x \neq 1$. Geometrically, the line $y(x) := (1+x)/2$ is tangent to the hypograph of $g(\cdot)$ at the point $(1,1)$.

8.2 Producer Budgeting

8.2.1 Multi-Dimensional Indirect Production Function

Suppose the set of inputs I naturally partition into m subsets $I_1, \ldots, I_m$. The set I_i could represent inputs (i) available in the i^{th} period, (ii) whose availability is contingent on the i^{th} state of the world, or (iii) with related physical characteristics.

Let $x = (x_1, \ldots, x_m) \in \prod_{i=1}^{m} \mathbb{R}_+^{n_i}$ denote the vector of quantities of inputs consumed in each category, let $p_i \in \mathbb{R}_{++}^{n_i}$ denote the vector of (positive) prices of the inputs in set C_i, and let b_i denote the allocation of the overall budget $B = \sum_i b_i$ to the i^{th} category. The production function $\Phi(\cdot)$ is assumed *continuous*.

Definition 8.2. *The* **multi-dimensional indirect production function** $\Gamma : \prod_{i=1}^{m}(X_i \times \mathbb{R}) \to \mathbb{R}$ *is*

$$\Gamma((p_1, b_1), \ldots, (p_m, b_m)) := \max\{\Phi(x) : \ p_i \cdot x_i \leq b_i, \ i = 1, 2, \ldots, m\}.$$

As we did with the indirect production function, the price vectors, $p_1, \ldots, p_m$ can be normalized by their respective budget allocations, $b_1, \ldots, b_m$, and we write $\Gamma(p_1, \ldots, p_m)$ in lieu of $\Gamma((p_1/b_1, 1), \ldots, (p_m/b_m, 1))$.

The budget allocations b_i, $i = 1, 2, \ldots, m$, are *optimal* when

$$\Gamma((p_1, b_1), \ldots, (p_m, b_m)) = \Gamma_\Phi((p_1, \ldots, p_m), B).$$

Viewed as a function of prices, the indirect production function $\Gamma_\Phi((p_1, \ldots, p_m), B)$ is *always* quasiconvex. If, however, the budget allocations are *not* optimal, then nonconvexities may arise, as we now demonstrate with the following counter-example.

8.2.2 Counter-Example to Quasiconvexity

Let $\Phi : \mathbb{R}_+^2 \to \mathbb{R}$ be a restricted Cobb-Douglas function given by

$$\Phi(x_1, x_2) = \sqrt{g(x_1)x_2} \tag{8.9}$$

where $g(x) = \min(x, 1)$. It follows that

$$\Gamma(p_1, p_2) = \max\{\sqrt{g(x_1)x_2} : \ p_1x_1 \leq 1, \ p_2x_2 \leq 1\}.$$

In this simple setting, it is straightforward to show that

$$\Gamma(p_1, p_2) = \sqrt{g(1/p_1) \cdot 1/p_2} \ = \ \begin{cases} 1/p_1p_2 & \text{if} \quad p_1 \geq 1, \\ 1/p_2 & \text{if} \quad p_1 < 1. \end{cases}$$

Consider the lower level set $L_\Gamma^{\leq}(1)$. It is the union of the convex set $\{(p_1, p_2) : p_1p_2 \geq 1\}$, which is identical to a Cobb-Douglas unit input possibility set, with the convex set $\{(p_1, p_2) : p_2 \geq 1\}$. This union is most definitely *not* convex. Consequently, $\Gamma(p_1, p_2)$ is not quasiconvex.

8.3 Data Envelopment Analysis with Lower Bounds

8.3.1 Fixed-Charge Technology

Given a set of observed data

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$$

on N Decision-Making Units (DMUs), the constant returns-to-scale (CRS) DEA model of technology is

$$T^{CRS} = \{(x, y) : \ x \geq \sum_j \lambda_j x_j, \ y \leq \sum_j \lambda_j y_j \text{ for some } \lambda \geq 0\}. \tag{8.10}$$

Conceptually, T^{CRS} is the smallest convex, free disposable, constant return-to-scale technology that contains the data set $\mathcal{D}$.

Let

$$R_j := \{(x, y) = \lambda_j (x_j, y_j) \ \text{ for some } \ \lambda_j \geq 0\}$$

denote the ray emanating from the origin that passes through the observed data point (x_j, y_j). Let $\mathcal{FDH}(S)$ denote the free disposable hull of the set S (see Definition 3.19, p. 42). The next proposition follows directly from (8.10).

Proposition 8.3. $T^{CRS} = \mathcal{FDH}(\sum_j R_j)$.

Proposition (8.3) shows that points in the CRS technology are generated by all points obtained as a *sum of operations*, each of whom represents a *scaled* version of an observed operation.

The accuracy of an efficiency rating and its acceptance by management is critically dependent on how well the constructed efficiency frontier corresponds to the true efficiency frontier. In the application of efficiency analysis to the warehouse and distribution industry (see Chapter 12), quite often a scaled component warehouse in the composite warehouse was smaller than any actual warehouse in the data set. For this industry, using such low intensities in the construction of a composite DMU is certainly not realistic from either an economic or modeling perspective. Equally as important, managers often question the meaning of convex combinations that involve what they perceive to be irrelevant DMUs.

Points in the *Fixed-Charge DEA* model of technology are also generated by all points obtained as a sum of operations; however, each operation represents an *appropriately* scaled version of an observed operation. That is, the FC technology eliminates all points in R_j whose λ_j is positive but too small. Let

$$\Lambda_j(\ell_j) := \{0\} \cup [\ell_j, \infty)$$

and let

$$T_j(\ell_j) := \{(x, y) = \lambda_j (x_j, y_j), \ \lambda_j \in \Lambda_j(\ell_j)\}.$$

Definition 8.4. *The* **Fixed-Charge DEA model of technology (FC technology)** *is*

$$T^{FC} := \mathcal{FDH}\left(\sum_j T_j(\ell_j)\right),$$

where $\ell_j \in [0,1]$, $1 \le j \le N$.[2]

The FC technology is algebraically equivalent to

$$T^{FC} = \{(x,y) : \; x \ge \sum_j \lambda_j x_j, \; y \le \sum_j \lambda_j y_j \text{ for some } \lambda \in \prod_j \Lambda_j(\ell_j)\}. \quad (8.11)$$

We insist that $\ell_j \le 1$ to ensure that the fixed-charge technology T^{FC} contains the observed data.

The FC technology is so named because it requires the addition of a *fixed-charge* integer constraint to the usual linear program. The FC technology focuses on lower bounds for two reasons. First, scaling upwards can often be justifiable using the replication argument, and so it is less egregious. Second, upper bounds are unnecessary for the VRS models. For certain applications, most notably in which resources are easily scalable (e.g. labor), adding lower bound restrictions is unnecessary. For other applications, most notably those that involve capital investments that only make sense above a minimum level, adding lower bound restrictions is a simple way to avoid unreasonable composites.

We illustrate the technology construction and inherent nonconvexities of the FC technology with the following example.

Example 8.5. Table 8.1 lists the input and output of three DMUs: DMU_1, DMU_2 and the reference DMU_0. Figure 8.1 depicts $L^{FC}(20)$ when the lower bounds are set to 0.00, 0.25, 0.33, 0.60, and 0.75, respectively. (For simplicity the lower bounds are identical for each DMU.)

Table 8.1. Data for geometrical constructions.

DMU	Input 1	Input 2	Output
DMU_0	3	2.5	20
DMU_1	2	3	40
DMU_2	3	1	30

The graphs demonstrate the nonconvexity of the input possibility set—the left-most and right-most portions of a line segment joining two boundary

[2] It is understood that the definition of the FC technology depends on the particular choice for the ℓ_j.

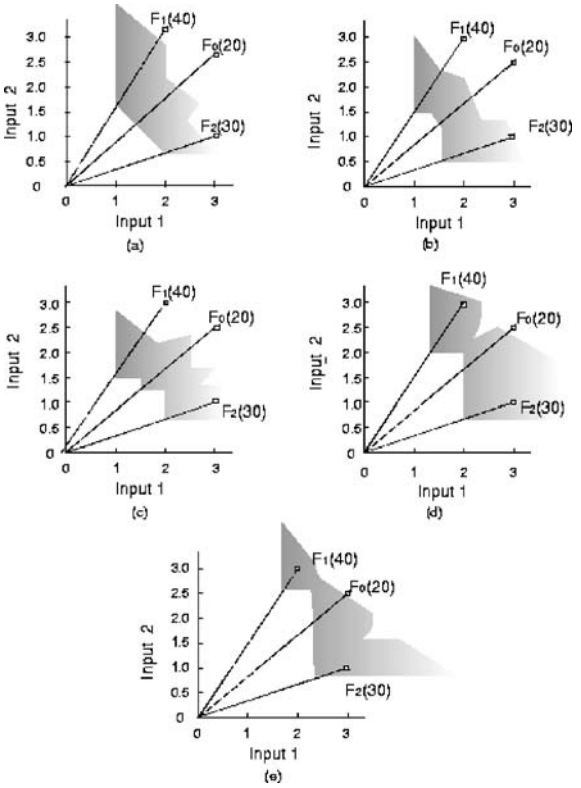


Fig. 8.1. $L^{FC}(20)$ for different lower bounds.

points will no longer belong to the input possibility set, the degree of which depends entirely on the size of the lower bounds. The input possibility sets $L^{FC}(20)$ move to the upper right as the lower bounds increase until $\ell = 1.00$.

8.3.2 Nonconvex Geometry of the Fixed-Charge Technology

We restrict attention to the case of two inputs and scalar output, so that we may show how the input possibility sets in Figure 8.1 are obtained.

The input possibility set $L^{FC}(y)$ is defined as

$$L^{FC}(y) := \left\{ x : \; x \geq \sum_j \lambda_j x_j, \; y \leq \sum_j \lambda_j y_j, \quad \text{for some } \lambda \in \prod_j \Lambda_j \right\}.$$

For each nonempty subset $K \subset N$ define the **level-subset** $L_K^{FC}(y)$ as

$$L_K^{FC}(y) := \left\{ x : \; x \geq \sum_{j \in K} \lambda_j x_j, \; y \leq \sum_{j \in K} \lambda_j y_j, \; \lambda_j \geq \ell_j, \text{ for each } j \in K \right\}. \tag{8.12}$$

To show how to geometrically construct $L^{FC}(y)$, it is sufficient to show how to geometrically construct each $L_K^{FC}(y)$, since

$$L^{FC}(y) = \bigcup_{K \subset N} L_K^{FC}(y).$$

Define

$$\hat{x}_K(\ell) := \sum_{j \in K} \ell_j x_j, \;\; \hat{y}_K(\ell) := \sum_{j \in K} \ell_j y_j, \;\; \tilde{x} := x - \hat{x}_K(\ell), \;\; \hat{\lambda}_j := \lambda_j - \ell_j.$$

The level-subset $L_K^{FC}(y)$ is equivalent to:

$$L_K^{FC}(y) = \hat{x}_K(\ell) + \{\tilde{x} : \; \tilde{x} \geq \sum_{j \in K} \lambda_j x_j, \; y - \hat{y}_K(l) \leq \sum_{j \in K} \lambda_j y_j, \; \lambda_j \geq 0\}. \tag{8.13}$$

The set defined in brackets in (8.13) is precisely the definition of the input possibility set corresponding to output level $y - \hat{y}_K(\ell)$ of the CRS technology restricted to those firms in the data set K. We shall denote this set as $L_K^{CRS}(y - \hat{y}_K(\ell))$, and it is understood that this set coincides with $\mathbb{R}_+^m$ when $y \leq \hat{y}_K(\ell)$. Thus,

$$L_K^{FC}(y) = \hat{x}_K(\ell) \; + \; L_K^{CRS}(y - \hat{y}_K(\ell)). \tag{8.14}$$

Table 8.2 summarizes the essential data required to graphically depict the input possibility sets. Recall that to construct $L_K^{CRS}(u)$ for any output $u > 0$, one calculates the *generator* points $x_j^* = (u/y_j)x_j$, for $j \in K$, forms their convex hull, and extends it by adding the nonnegative orthant to each x_j^*. Here, the generator points are simply

$$s_j^*(y) = \begin{cases} \hat{x}_K(\ell)\,, & y \leq \hat{y}_K(\ell), \\ \hat{x}_K(\ell) + \frac{y - \hat{y}_K(\ell)}{y_j} \; x_j\,, & y > \hat{y}_K(\ell). \end{cases} \tag{8.15}$$

8.3.3 The Low Intensity Phenomenon

We explain why the "low intensity" phenomenon occurs in the CRS technology. We first establish a property of *partial efficiency scores*.

Let $\theta_k^*(S_i, N_j)$ denote the optimum solution of the familiar linear program to determine radial input efficiency:

$$P(S_i, N_j) : \min \left\{ \theta_k : \sum_{s \in N_j} \lambda_s x_s \leq \theta_k x_k, \sum_{s \in N_j} \lambda_s y_{s\ell} \geq y_{k\ell}, \ell \in S_i \right\}, \tag{8.16}$$

except that here the output set is restricted to a subset S_i and the set of DMUs is restricted to a subset N_j. Let S and N denote, respectively, the sets of all outputs and DMUs.

Table 8.2. Construction of $L^{FC}(20)$ for different lower bound values.

i	**Subset** K	$\hat{x}_K(\ell)$		$\hat{y}_K(l)$	$x^*(\ell, K)$		s_i^*	
		$\hat{x_1}$	$\hat{x_2}$		x_1^*	x_2^*	s_{i1}^*	s_{i2}^*
$\ell = 0.25$								
(1)	$\{\text{DMU}_1\}$	0.50	0.75	10.00	0.50	0.75	1.00	1.50
(2)	$\{\text{DMU}_2\}$	0.75	0.25	7.50	1.25	0.42	2.00	0.67
(3)	$\{\text{DMU}_1, \text{DMU}_2\}$	1.25	1.00	17.50	0.13	0.19	1.38	1.19
(4)					0.25	0.08	1.50	1.08
(5)	$\{\text{DMU}_0\}$	0.75	0.63	5.00	2.25	1.88	3.00	2.50
$\ell = 0.33$								
(1)	$\{\text{DMU}_1\}$	0.67	1.00	13.33	0.33	0.50	1.00	1.50
(2)	$\{\text{DMU}_2\}$	1.00	0.33	10.00	1.00	0.33	2.00	0.66
(3)	$\{\text{DMU}_1, \text{DMU}_2\}$	1.67	1.33	23.33	–	–	1.67	1.33
(4)	$\{\text{DMU}_0\}$	1.00	0.83	6.66	2.00	1.67	3.00	2.50
$\ell = 0.6$								
(1)	$\{\text{DMU}_1\}$	1.20	1.80	24.00	–	–	1.20	1.80
(2)	$\{\text{DMU}_2\}$	1.80	0.60	18.00	0.20	0.07	2.00	0.67
(3)	$\{\text{DMU}_1, \text{DMU}_2\}$	3.00	2.40	42.00	–	–	3.00	2.40
(4)	$\{\text{DMU}_0\}$	1.80	1.50	12.00	1.20	1.00	3.00	2.50
$\ell = 0.75$								
(1)	$\{\text{DMU}_1\}$	1.50	2.25	30.00	–	–	1.50	2.25
(2)	$\{\text{DMU}_2\}$	2.25	0.75	22.50	–	–	2.25	0.75
(3)	$\{\text{DMU}_1, \text{DMU}_2\}$	3.75	3.00	52.50	–	–	3.75	3.00
(4)	$\{\text{DMU}_0\}$	2.25	1.88	15.00	0.75	0.63	3.00	2.50

Proposition 8.6. *For each partition $S_1, \ldots, S_K$ of the set of outputs S and collection of subsets $N_1, N_2 \ldots N_L$ of DMUs N,*

$$\sum_{i,j} \theta_k^*(S_i, N_j) \geq \theta_k^*(S, N).$$

Proof. Let $\lambda^*(S_i, N_j)$ denote an optimal solution to $P(S_i, N_j)$. Extend $\lambda^*(S_i, N_j)$ to $\mathbb{R}^{N+1}$ by adding the requisite zeroes in the obvious way, and let $\hat{\lambda}(S_i, N_j)$ denote the extended vector. Define $\tilde{\lambda} = \sum_{i,j} \hat{\lambda}(S_i, N_j)$ and $\tilde{\theta}_k = \sum_{i,j} \theta_k^*(S_i, N_j)$. The result follows since $(\tilde{\lambda}, \tilde{\theta}_k)$ is a feasible solution for the original CRS problem involving all outputs and DMUs. □

Example 8.7. Table 8.3 lists the normalized inputs and outputs of three hypothetical DMUs. We chose DMU_0 as the reference unit and used its input-output values to normalize the data. Using the CRS model, the composite unit is composed by 0.02 of DMU_2 and 1.21 of DMU_1. The small intensity associated with DMU_2 is the result of two effects: a *scale effect* and *complement effect*. We now explain each effect.

If DMU_0 is compared separately to DMU_1 or to DMU_2, it will be assessed as efficient. Stated differently, $\theta_0^*(\{1,2,3,4\}, \{j\}) = 100\%$, $j = 1, 2$. (Its $\mathcal{FDH}$

efficiency score is 1.) However, when both DMU_1 and DMU_2 are considered, DMU_0 is only 13% efficient. Why does this happen? The main problem is that DMU_1 produces very little output 1. If output 1 were not in the model so that $S_2 = \{2, 3, 4\}$, and if DMU_0 is compared only to DMU_1 so that $N_1 = \{1\}$, then the efficiency score becomes $\theta_0^*(S_2, N_1) = 100 \times (0.09/0.68) = 13.2\%$. On the other hand, DMU_2 produces a disproportional amount of output 1 per unit of input. So, if $S_1 = \{1\}$ and $N_2 = \{2\}$, then the efficiency score becomes $\theta_0^*(S_1, N_2) = 100 \times (1.27/51.6) = 2.5\%$ and not 100% as obtained by the $\mathcal{FDH}$ model. DMU_2 complements DMU_1, and since DMU_2 is exceptionally productive with respect to output 1, DMU_1 can afford to "buy" only a small portion.

Table 8.3. Normalized data for the example.

	Normalized Inputs		Normalized Outputs			
	$\overline{I}_1$	$\overline{I}_2$	$\overline{O}_1$	$\overline{O}_2$	$\overline{O}_3$	$\overline{O}_4$
DMU_0	1.00	1.00	1.00	1.00	1.00	1.00
DMU_1	**0.09**	0.04	0.03	0.87	0.82	**0.68**
DMU_2	**1.27**	0.53	**51.60**	2.50	0.36	75.40

In this example, DMU_1 can reveal DMU_0 to be extremely inefficient, if only S_1 is considered; similarly, DMU_2 can reveal DMU_0 to be extremely inefficient, if only S_2, the complementary set of outputs, is considered. To reveal DMU_0 efficient *all* outputs must be considered. It turns out that when all outputs are considered DMU_0 is still extremely inefficient. As Proposition 8.6 shows, this result is not an artifact of the particular numbers chosen. The inefficiencies of DMU_0 computed when the output set is restricted to two complementary subsets can be used to bound the inefficiency of DMU_0 when all outputs are considered, namely,

$$\hat{\theta}_0 = \theta_0(S_2, N_1) + \theta_0(S_1, N_2) = 15.7\% > 13\% = \theta_0^*.$$

8.4 Projective-Convexity

The defining property of convexity can be formulated in terms of paths. A set C is convex if for each two points x and y in C, there exists a continuous path π whose initial point $\pi(0)$ is x, whose terminal point $\pi(1)$ is y and whose coordinate functions π_i are given by $\pi_i(\lambda) = (1 - \lambda)x_i + \lambda y_i$. A natural generalization of convexity is to allow a *different* parameter λ for each coordinate i. This is the essential idea behind projective-convexity.

8.4.1 Definitions and characterizations

Let X_i, $i = 1, 2, \ldots, N$, denote finite-dimensional Euclidean spaces.

Definition 8.8. *A subset $C \subset \prod_{i=1}^{N} X_i$ is* **projectively-convex** *(P-convex) if for each two points x, y in C and each coordinate i and for all $\lambda \in [0,1]$, there exist $\lambda_1, \lambda_2, \ldots, \lambda_{i-1}, \lambda_{i+1}, \ldots, \lambda_N \in [0,1]$ such that the point*

$$((1-\lambda_1)x_1 + \lambda_1 y_1, \ldots, (1-\lambda)x_i + \lambda y_i, \ldots, (1-\lambda_N)x_N + \lambda_N y_N)$$

belongs to C.

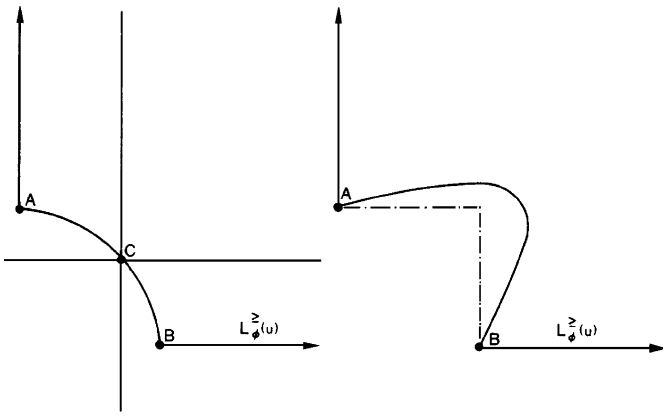


Fig. 8.2. Graphical illustration of projective-convexity.

As in the case of convex sets, the defining property of projective-convexity can be reformulated in terms of (more general) paths. Let $[a, b]$ denote the line segment joining points a and b.

Definition 8.9. *For each two points x, $y \in \prod_{i=1}^{N} X_i$, the* **rectangle joining** x **and** y, *$R[x, y]$, is*

$$R[x, y] := [x_1, y_1] \times [x_2, y_2] \times \ldots \times [x_N, y_N].$$

A set C is projectively-convex if for each two points x and y in C, there exists a path π (not necessarily continuous) for which the image of π lies in $R[x, y] \cap C$ and the image of each π_i is $[x_i, y_i]$. See Figure 8.2(A, B).

The reason why these sets are called projectively-convex is explained by the following proposition, which characterizes them in terms of the projection maps $P^k : \prod_{i=1}^N X_i \longrightarrow X_k$, $k = 1, 2, \ldots, N$. Let Ω denote the set of all convex subsets $C \subset \prod_{i=1}^N X_i$ expressible as products $\prod_{i=1}^N C_i$ of convex subsets $C_i \subset X_i$, $i = 1, 2, \ldots, N$. See Figure 8.3(C, D).

Theorem 8.10. *A set $S \subset \prod_{i=1}^N X_i$ is projectively-convex (P-convex) if and only if it satisfies the following projection property: for every $C \subset \Omega$ each projection $P^k(S \cap C)$, $k = 1, 2, \ldots, N$, is convex.*

Proof. Suppose S satisfies the projection property. The rectangle $R[x, y] \in \Omega$ for each $x,\ y \in S$. The projection property implies that each $P^k(S \cap R[x, y])$ is convex, which is equivalent to the defining property of projective-convexity. Conversely, the intersection of a projectively-convex set with a set in Ω is projectively-convex. Since the projection of a projectively-convex set is convex, a projectively-convex set possesses the projection property. □

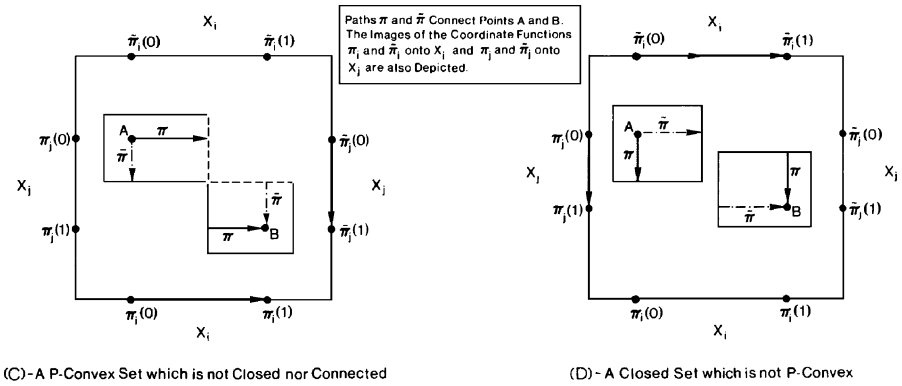


Fig. 8.3. Topological properties of projectively-convex sets.

Next, we establish an alternate characterization of closed projectively-convex sets, which will be required in the sequel. For notational convenience, we take $N = 2$.

Theorem 8.11. *Let $C \subset X_1 \times X_2$ be a closed set. Then C is projectively-convex if and only if $C \cap R[x, y]$ is connected for all $x,\ y \in C$.*

Proof. If C is not projectively-convex, then, without loss of generality, there exist vectors $x = (x^1, x^2),\ y = (y^1, y^2) \in C$ and a $\lambda_0 \in (0, 1)$ for which

$$\left(\{(1 - \lambda_0)x^1 + \lambda_0 y^1\} \times [x^2, y^2] \right) \cap C = \emptyset.$$

Since both x and y belong to $C \cap R[x,y]$, the set $C \cap R[x,y]$ is disconnected, since it can be written as a nonempty, disjoint union of two relatively open subsets.

Conversely, suppose there exist $x,\ y \in C$ for which $C \cap R[x,y]$ is disconnected. Since $C \cap R[x,y]$ is also compact, it can be written as the disjoint union of two compact subsets, say A and B. Pick an $a \in A$ and $b \in B$ that achieves the minimum Euclidean distance between A and B. Obviously, $a \neq b$ and by construction $R[x,y] \cap C = \{a,b\}$. This immediately shows that C is not projectively-convex. □

We now turn to functions. A function is quasiconcave (quasiconvex) if each of its upper (lower) level sets are convex.

Definition 8.12. *A function* $f : \prod_{i=1}^{N} X_i \longrightarrow I\!R$ *is* **projectively-concave (projectively-convex)** *if each of its upper (lower) level sets is projectively-convex.*

In practice, the domain of a projectively-concave function or set is determined by the variables of a particular problem. We note the following facts.

- Every quasiconcave (quasiconvex) function is projectively-concave (projectively-convex).
- Every projectively-concave (projectively-convex) function is quasiconcave (quasiconvex) in each of its arguments.

A more general relationship between quasiconcavity and projective-concavity is recorded in the following proposition, whose proof is an immediate consequence of the definitions.

Proposition 8.13. *Suppose* $\psi : I\!R^N \longrightarrow I\!R$ *is nondecreasing (nonincreasing) in each of its arguments, and each function* $u_i : X_i \to I\!R$ *is quasiconcave (quasiconvex) on its domain* X_i, $i = 1, 2, \ldots, N$. *The function* $\Phi : \prod_{i=1}^{N} X_i \to I\!R$ *defined by*

$$\Phi(x_1, x_2, \ldots, x_N) := \psi(u_1(x_1), u_2(x_2), \ldots, u_N(x_N))$$

is projectively-concave (projectively-convex).

Example 8.14. If the $u_i(\cdot)$ are quasiconcave (quasiconvex), then the functions $\Phi(x_1, x_2, \ldots, x_N) = \max\{u_1(x_1), \ldots, u_N(x_N)\}$ and $\Phi(x_1, \ldots, x_N) = \min\{u_1(x_1), \ldots, u_N(x_N)\}$ are projectively-concave (projectively-convex).

Example 8.15. If the $u_i(\cdot)$ are quasiconcave (quasiconvex) and positive, then $\Phi(x_1, x_2, \ldots, x_N) = \prod_i u_i(x_i)$ is projectively-concave (projectively-convex).

Example 8.16. If the $u_i(\cdot)$ are quasiconcave (quasiconvex) and positive, then $\Phi(x_1, x_2, \ldots, x_N) = \sum_i u_i(x_i)$ is projectively-concave (projectively-convex). It is fundamental result, first established by G. Debreu, that the sum of quasiconcave functions *cannot* be quasiconcave if at least two of the functions are quasiconcave but not concave.

8.4.2 Separation Properties

For the sake of presentation we restrict attention to the space $X_1 \times X_2 = \mathbb{R}^n \times \mathbb{R}^m$. The symbols $P^n(A)$ and $P^m(A)$ denote the projections of the set $A \in \mathbb{R}^n \times \mathbb{R}^m$ onto $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. For each $a \in \mathbb{R}^k$, let $H(a)$ denote the closed cone $\{x \in \mathbb{R}^k : \ a \cdot x \geq 0\}$.

Definition 8.17. *The* **quadrant generated by** $(a, b) \in \mathbb{R}^n \times \mathbb{R}^m$ *is the closed cone*

$$Q[a, b] := \{(\hat{x}, \hat{y}) \in I\!R^n \times I\!R^m : \ a \cdot \hat{x} \geq 0, \ b \cdot \hat{y} \geq 0\}.$$

The translation of the quadrant $Q[a, b]$ to the point (x, y) is the set $(x, y) + Q[a, b]$. The quadrant $Q[a, b]$ **separates** *(x, y)* **from** *$S \subset \mathbb{R}^n \times \mathbb{R}^m$ if*

$$(\ (x, y) + Q[a, b]\ ) \cap S = \emptyset.$$

A quadrant is **nontrivial** *if either a or b is not zero.*

Theorem 8.18. Separation by Quadrant *Let $S \subset I\!R^n \times I\!R^m$ be a closed, projectively-convex set. If $(x, y) \notin S$, then there exists a nontrivial quadrant $Q[a, b]$ that separates (x, y) from S.*

Proof. Since translations preserve the properties of closure and projective-convexity, we may assume, without loss of generality, that $(x, y) = (0, 0)$.

Since $(0, 0) \notin S$ there is a positive ϵ for which

$$S \cap \ (B^1_\epsilon(0) \times B^2_\epsilon(0)) = \emptyset, \tag{8.17}$$

where we let $B^i_\epsilon(0)$ denote the open ball of radius ϵ about $0 \in X_i$, $i = 1, 2$. Let

$$A := \{x \in \mathbb{R}^n : \text{ there exists a } \ y \ \text{ such that } \ (x, y) \in S \text{ and } ||y|| \leq \epsilon\}. \tag{8.18}$$

Assume first that A is not empty. Since the set A can be expressed as the projection

$$P^n(S \cap (\mathbb{R}^n \times B^2_\epsilon(0))),$$

it is convex by Theorem 8.10. It is immediate from (8.17) and (8.18) that $A \cap B^1_\epsilon(0)$ is empty. Since 0 does not belong to the closure of A, a nontrivial hyperplane exists that strictly separates it from A, i.e., there exists a nonzero $a \in X_1$ such that

$$H(a) \cap A = \emptyset. \tag{8.19}$$

If A itself is empty, then pick an arbitrary nonzero $a \in \mathbb{R}^n$ and note that (8.19) holds trivially.

If $(x, y) \in S$ such that $a \cdot x \geq 0$, then (8.19) and the definition of A (8.18) imply that $||y|| > \epsilon$. Thus the set

$$B := \{y \in \mathbb{R}^m : \text{ there exists an } x \text{ such that } (x, y) \in S \text{ and } a \cdot x \geq 0\}. \tag{8.20}$$

is disjoint from $B_\epsilon^2(0)$. Assume first that B is not empty. Since the set B can be expressed as the projection

$$P^m(S \cap (H(a) \times \mathbb{R}^m)),$$

it is convex by Theorem 8.10. Since 0 does not belong to the closure of B, a nontrivial hyperplane exists that strictly separates it from B, i.e., there exists a nonzero $b \in X_2$ for which

$$H(b) \cap B = \emptyset. \tag{8.21}$$

If B is empty, then pick an arbitrary nonzero $b \in \mathbb{R}^m$ and note that (8.21) holds trivially.

Now pick an arbitrary $(x, y) \in S$. If $a \cdot x \geq 0$, then by definition (8.20) $y \in B$. But then $y \notin H(b)$, and thus $(x, y) \notin Q[a, b]$. We have found a separating quadrant, as required. □

Corollary 8.19. *A closed projectively-convex set is the intersection of all complements of quadrants that contain it.*

Convex sets have supporting hyperplanes. Analogously, projectively-convex sets have supporting quadrants.

Definition 8.20. *The quadrant* $(x, y) + Q[a, b]$ *is said to* **support** $S \subset \mathbb{R}^n \times \mathbb{R}^m$ *at* $(x, y) \in S$ *if the interior of* $(x, y) + Q[a, b]$ *is disjoint from* S.

Corollary 8.21. *Let* $S \subset \mathbb{R}^n \times \mathbb{R}^m$ *be a closed projectively-convex set. If* (x, y) *lies on the boundary of* S*, then there exists a supporting quadrant to* S *at* (x, y).

Remark 8.22. Every projectively-convex set $S \subset \mathbb{R}^n \times \mathbb{R}^m$ is projectively-convex when viewed as a subset of $\mathbb{R}^{n+m}$. Consequently, at least one of the 2^{n+m} orthants generated by a point x not in a closed projectively-convex set S must be disjoint from S.

Remark 8.23. With the obvious modifications the separating quadrant theorem is valid for locally convex topological vector spaces. The quadrants are generated by the intersection of halfspaces determined by the continuous linear functionals.

We now turn to the differentiable setting. For a differentiable quasiconcave function $\theta : \mathbb{R}^k \to \mathbb{R}$

$$\nabla\theta(x_1) \cdot (x_2 - x_1) < 0 \Rightarrow \theta(x_2) < \theta(x_1) \text{ holds for all } x_1,\ x_2 \in \mathbb{R}^k. \tag{8.22}$$

Property (8.22) has a geometrical interpretation: if $\nabla\theta(x) \neq 0$ and x lies on the boundary of $L_\theta^{\geq}(\theta(x))$, then the gradient $\nabla\theta(x)$ generates a supporting hyperplane

$$\{z \in \mathbb{R}^k : \nabla\theta(x) \cdot z = \nabla\theta(x) \cdot x\}$$

to $L^{\geq}_{\theta}(\theta(x))$ at x. A differentiable projectively-concave function possesses a similar geometrical property: if $\nabla_x f(x,y) \neq 0$, $\nabla_y f(x,y) \neq 0$ and (x,y) lies on the boundary of $L^{\geq}_{f}(f(x,y))$, then the quadrant $Q[-\nabla_x f(x,y),\ -\nabla_y f(x,y)]$ supports $L^{\geq}_{f}(f(x,y))$ at (x,y). The following theorem establishes this geometrical property.

Theorem 8.24. *Let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ be differentiable and projectively-concave. Then*

$$\left.\begin{array}{l} \nabla_x f(x_1,y_1) \cdot (x_2 - x_1) < 0 \\ \nabla_y f(x_1,y_1) \cdot (y_2 - y_1) < 0 \end{array}\right\} \Rightarrow f(x_2,y_2) < f(x_1,y_1).$$

Proof. Suppose to the contrary there exist vectors (x_1,y_1), (x_2,y_2) for which $f(x_2,y_2) \geq f(x_1,y_1)$ but both $\nabla_x f(x_1,y_1)\cdot(x_2-x_1)$ and $\nabla_y f(x_1,y_1)\cdot(y_2-y_1)$ are negative. Since

$$L^{\geq}_{f}(f(x_1,y_1)) \cap R[(x_1,y_1),(x_2,y_2)]$$

is connected (Theorem 8.11), there exists an infinite sequence of points

$$(x_k,y_k) \in L^{\geq}_{f}(f(x_1,y_1)) \cap R[(x_1,y_1),(x_2,y_2)],$$

$k = 1,2,\ldots$, not equal to (x_1,y_1) that converges to (x_1,y_1). We may write each (x_k,y_k) as $(x_1,y_1) + d_k$ where $d_k := (\alpha_k(x_2 - x_1),\ \beta_k(y_2 - y_1))$ and $\alpha_k,\ \beta_k \in (0,1)$. Note that $\alpha_k \to 0$, $\beta_k \to 0$ and $||d_k|| \to 0$.[3] By construction,

$$\begin{aligned} 0 &\leq f(x_k,y_k) - f(x_1,y_1) \\ &= \alpha_k \nabla_x f(x_1,y_1) \cdot (x_2 - x_1) + \beta_k \nabla_y f(x_1,y_1) \cdot (y_2 - y_1) + o(||d_k||), \end{aligned} \tag{8.23}$$

where $\lim_{||d_k||\to 0} o(||d_k||)/||d_k|| = 0$. Passing to a subsequence if necessary, we may assume, without loss of generality, that $\alpha_k \geq \beta_k$ for all k. Thus, $||d_k|| \leq \alpha_k||d||$ where $d = (x_2 - x_1, y_2 - y_1)$. Dividing both sides of (8.23) by $\alpha_k||d||$ and taking limits produces the desired contradiction. □

8.4.3 Dual Characterization

The duality between the indirect and direct production functions in the quasiconcave setting extends in a natural way to the projectively-concave setting. Once again, we shall develop the duality in the special case of $X_1 \times X_2 = \mathbb{R}^n \times \mathbb{R}^m$, although it holds generally.

Define

$$\Phi^*(x_1,x_2) := \inf\{\Gamma(p_1,p_2) : \ p_i \cdot x_i \leq 1,\ i = 1,2\}.$$

[3] The symbol $||\cdot||$ denotes the Euclidean norm in $\mathbb{R}^{n+m}$.

Theorem 8.25. *If $\Phi(\cdot,\cdot)$ is continuous, nondecreasing and projectively-concave, then $\Phi(\cdot,\cdot) = \Phi^*(\cdot,\cdot)$.*

Proof. The definitions of $\Phi^*(x_1,x_2)$ and $\Gamma(p_1,p_2)$ imply that $\Phi^*(x_1,x_2) \geq \Phi(x_1,x_2)$. It remains to show the reverse inequality.

Pick a $v > \Phi(x_1,x_2)$ such that $L_\Phi^{\geq}(v)$ is not empty. (If no such v exists, then $\Phi^*(x_1,x_2) \leq \Phi(x_1,x_2)$, and the result follows.) Since $L_\Phi^{\geq}(v)$ is closed and does not contain (x_1,x_2), there exists a positive ϵ for which $(\hat{x}_1,\hat{x}_2) \notin L_\Phi^{\geq}(v)$, where

$$\hat{x}_1 := x_1 + \epsilon(1,1,\ldots,1)$$
$$\hat{x}_2 := x_2 + \epsilon(1,1,\ldots,1).$$

Since $\Phi(\cdot,\cdot)$ is projectively-concave, the separation by quadrant Theorem 8.18 guarantees existence of a nontrivial quadrant $Q[a,b]$ such that

$$\Big((\hat{x}_1,\hat{x}_2) + Q[a,b]\Big) \cap L_\Phi^{\geq}(v) = \emptyset.$$

Consequently, if $(\bar{x}_1,\bar{x}_2) \in L_\Phi^{\geq}(v)$, then either

$$a \cdot \bar{x}_1 > a \cdot \hat{x}_1 = a \cdot x_1 + \epsilon \sum_{i=1}^{n} a_i, \tag{8.24}$$

or

$$b \cdot \bar{x}_2 > b \cdot \hat{x}_2 = b \cdot x_2 + \epsilon \sum_{i=1}^{m} b_i. \tag{8.25}$$

Since $\Phi(\cdot,\cdot)$ is nondecreasing, each component of a and b must be nonnegative. It follows from (8.24) and (8.25) that, if necessary, it is possible to perturb both a and b so that (i) each vector is positive and (ii) if $(\bar{x}_1,\bar{x}_2) \in L_\Phi^{\geq}(v)$, then either

$$a \cdot \bar{x}_1 > a \cdot x_1 \tag{8.26}$$

or

$$b \cdot \bar{x}_2 > b \cdot x_2. \tag{8.27}$$

Let $\hat{p}_1 := a/(a \cdot x_1)$ and $\hat{p}_2 := b/(b \cdot x_2)$. By construction, both $\hat{p}_1$ and $\hat{p}_2$ are positive and $\hat{p}_i \cdot x_i \leq 1$, $i = 1,2$. Moreover, $\Gamma(\hat{p}_1,\hat{p}_2) < v$ in light of (8.26) and (8.27). We have therefore established that

$$\Phi^*(x_1,x_2) \leq \Gamma(\hat{p}_1,\hat{p}_2) < v.$$

As v was chosen arbitrarily, $\Phi^*(x_1,x_2) \leq \Phi(x_1,x_2)$, as required. □

8.5 Exercises

8.1. Show that an input free disposable set in $\mathbb{R}^2_+$ is always projectively-convex.

8.2. Prove Corollary 8.19.

8.3. Prove Corollary 8.21.

8.4. Let A be an $m \times n$ matrix. Consider the following two statements:

(1) There exists an $x \in \mathbb{R}^n$ such that $Ax < 0$.
(2) There exists a nonzero $y \geq 0$ such that $y^T A = 0$.

Prove that exactly one of these two statements must hold.

8.5. A set $S \subset \mathbb{R}^n$ is a *cone* if $x \in S$ implies that $\lambda x \in S$ for all $\lambda \geq 0$. If S is also convex, then it is called a *convex cone*. The *polar cone* of a (not necessarily convex) set $S \subset \mathbb{R}^n$ is

$$S^* := \{p \in \mathbb{R}^n : \ p \cdot x \leq 0 \text{ for all } x \in S\}.$$

(If S is empty, then $S^* = \mathbb{R}^n$.)

(a) Let $S \subset \mathbb{R}^n$ be nonempty. Show that S^* is a closed, convex cone.
(b) Let $S \subset \mathbb{R}^n$ be a nonempty, closed convex cone. Show that $S = (S^*)^*$.

8.6. Prove that every closed, connected, evenly-separable subset of $\mathbb{R}^2$ must also be projectively-convex.

8.6 Bibliographical Notes

Data Envelopment Analysis with lower bounds is fully developed in Bouhnik et. al. [2001] (see also Vlatsa [1995]). Projective-convexity is introduced and analyzed in Hackman and Passy [1988], and its properties are further developed in First et. al. [1990, 1992]. General duality theory for projective-convexity is developed in First et. al. [1993], which also provides sufficient conditions that guarantee that an evenly-separable set will also be projectively-convex. Multi-dimensional indirect production functions are discussed in Blackorby et. al. [1978]. Petersen [1990] develops an interesting nonconvex model of technology.

8.7 Solutions to Exercises

8.1 Let $S \subset \mathbb{R}^2_+$ be input freely disposable, and pick $(K_i, L_i) \in S$, $i = 1, 2$. If either $(K_2, L_2) \geq (K_1, L_1)$ or $(K_1, L_1) \geq (K_2, L_2)$, then the line segment joining these two point also belongs to S. Obviously, this line segment defines an appropriate path connecting the two points. It remains to consider the case, without loss of generality, when $K_2 \geq K_1$ but $L_2 \leq L_1$. In this setting, the collections of points $\{(K, L) : K_1 \leq K \leq K_2,\ L = L_1\}$ and $\{(K, L) : K = K_2,\ L_1 \leq L \leq L_2\}$ also belong to S. These points define an appropriate path connecting the two points.

8.2 Let S be a closed projectively-convex set, and let $\mathcal{I}$ denote the intersection of all complements of quadrants that contain S. By definition, $S \subset \mathcal{I}$. To show the reverse inclusion, pick an $(x, y) \notin S$. By Theorem 8.18, there exists a quadrant $Q[a, b]$ that separates (x, y) from S. Clearly, $\mathcal{I} \subset \mathbb{R}^n \times \mathbb{R}^m \setminus ((x, y) + Q[a, b])$, and so $(x, y) \notin \mathcal{I}$.

8.3 Without loss of generality, assume that $(x, y) = (0, 0)$. Pick an infinite sequence $(x_i, y_i) \to (0, 0)$ for which $(x_i, y_i) \notin S$ for each $i = 1, 2, \ldots$. By Theorem 8.18, it is possible to find a nontrivial quadrant $Q[a_i, b_i]$ that separates (x_i, y_i) from S. Let $\bar{a}_i := a_i/||a_i||$ and $\bar{b}_i := b_i/||b_i||$. Since the boundary of the unit ball is compact, we may extract convergent subsequences from $\{\bar{a}_i\}$ and $\{\bar{b}_i\}$; let a and b denote the nonzero limit points. We claim that $Q[a, b]$ supports S at $(0, 0)$. To establish this, pick an arbitrary $(x_0, y_0) \in S$. Since $(x_0, y_0) \notin (x_i, y_i) + Q[a_i, b_i]$,

$$\min \left\{ \frac{a_i \cdot x_0 - a_i \cdot x_i}{||a_i||},\ \frac{b_i \cdot y_0 - b_i \cdot y_i}{||b_i||} \right\} < 0 \text{ for each } i. \tag{8.28}$$

Taking limits in (8.28) shows that $\min\{a \cdot x_0,\ b \cdot y_0\} \leq 0$, which implies that (x_0, y_0) does not belong to the interior of $Q[a, b]$, as required.

8.4 First, we show that both statements cannot simultaneously hold. Suppose both (1) and (2) hold. It then follows that $0 > y^T(Ax) = (y^T A)x = 0$, an obvious contradiction. Suppose (1) does not hold. Let $V := \{v \in \mathbb{R}^m :\ v = Ax\}$ and let $Z := \mathbb{R}^m_{--} = \{z \in \mathbb{R}^m :\ z < 0\}$. Clearly, $V \cap Z$ is empty. By the separation theorem for convex sets, there exists a $p \in \mathbb{R}^m$ such that $p \cdot v \geq p \cdot z$ for all $v \in V$ and $z \in Z$. Since all the components of z are negative, this inequality can only hold if all the components of p are non-negative. Moreover, since the origin lies on the boundary of Z, this inequality also implies that $p \cdot v \geq 0$ for all $v \in V$. This in turn implies that $p \cdot v = p^T(Ax) = (p^T A)x \geq 0$ for all $x \in \mathbb{R}^m$, which can only hold true if $p^T A = 0$. The result follows.

8.5 (a) Obviously, $0 \in S^*$ and so S^* is nonempty. Pick $p_1,\ p_2 \in S^*$ and $\lambda \in [0, 1]$. For each $x \in S$,

$$(\lambda p_1 + (1 - \lambda)p_2) \cdot x = \lambda(p_1 \cdot x) + (1 - \lambda)(p_2 \cdot x) \leq 0,$$

which implies that $\lambda p_1 + (1-\lambda)p_2 \in S^*$. Thus, S^* is convex. As for closure, pick an infinite sequence of points $p_1, p_2, \ldots$ in S^* such that $p_n \to p$. Pick $x \in S$. Since $p_n \cdot x \leq 0$ for each n and $p_n \to p$, it follows by the continuity of the inner product that $p \cdot x \leq 0$, too. This shows that $p \in S^*$, which establishes closure.

(b) By definition,

$$(S^*)^* = \{y : y \cdot p \leq 0 \text{ for all } p \in S^*\}.$$

Clearly, $S \subset (S^*)^*$. To show the reverse inclusion, we show that if $x \notin S$ then $x \notin (S^*)^*$. Pick $x \notin S$. By the separation theorem for convex sets, there exists a p such that $p \cdot x > p \cdot z$ for all $z \in S$. Since S is a cone, this inequality can only hold if $p \cdot z \leq 0$ for all $z \in S$. This in turn implies that $p \in S^*$. Moreover, since S is closed, the origin belongs to S, and so $p \cdot x > 0$, too. Consequently, $x \notin (S^*)^*$, as required.

8.6 Let $S \subset \mathbb{R}^2_+$ be closed, connected, and evenly-separable. If S is not projectively-convex, there exists two points $x = (K_1, L_1)$, $y = (K_2, L_2)$ in S such that the rectangle $R[x, y] = \{x, y\}$. Since translations preserve the properties of closure, connectedness and even-separability, we may assume, without loss of generality, that $x = 0$ and $y > 0$. We now argue that

$$S \cap \{(K, L) : K \leq 0,\ L > 0\} = \emptyset.$$

Suppose, to the contrary, there exists a $z = (K, L) \in S$ such $K \leq 0$ and $L > 0$. Pick L' such that $L' < \min\{L, L_2\}$. The point $(0, L') \notin S$. Each of the four quadrants defined by translating $(0, L')$ to the origin contains either x, y, or z. Consequently, it is not possible to separate $(0, L')$ from S with a quadrant, thus contradicting the evenly-separable property of S. A similar argument shows that

$$S \cap \{(K, L) : K \geq 0,\ L < 0\} = \emptyset$$

and

$$S \cap \left(\mathbb{R}^2_+ \ - \ \{(K, L) : K \leq K_2,\ L \geq L_2\}\right) = \emptyset,$$

too. For each $\epsilon > 0$, define the sets

$$S_1(\epsilon) := \{(K, L) :\ K \leq \epsilon,\ L \leq \epsilon\},$$
$$S_2(\epsilon) := \{(K, L) :\ K \geq K_2 - \epsilon,\ L \geq L_2 - \epsilon\}.$$

We have established that

$$S \subset S_1(\epsilon) \cup S_2(\epsilon) \tag{8.29}$$

for each $\epsilon > 0$. Since $y > 0$, it is possible to find an ϵ such that

$$S_1(\epsilon) \cap S_2(\epsilon) = \emptyset. \tag{8.30}$$

Since the sets $S_1(\epsilon)$ and $S_1(\epsilon)$ are obviously open, (8.29) and (8.30) show that S can be written as the disjoint union of two open sets, which implies that S is disconnected, a clear contradiction. The result follows.

Part II

Efficiency Measurement

9

Efficiency Analysis

At its core, a measure of efficiency compares an observed input-output pair (x, y) to its projection $(\hat{x}, \hat{y})$ onto the boundary of a technology set $\mathcal{T}$. Obviously, there are several ways to construct a technology $\mathcal{T}$ set from a data set $\mathcal{D}$, and there are several ways to project a point onto a boundary of a set. In this chapter, we formalize these notions. In what follows, we suppress the functional dependence of a particular efficiency measure on $\mathcal{T}$.

9.1 Input and Output Efficiency

We begin with the most commonly used measures for input or output efficiency.

Definition 9.1. *The* **radial measure of input efficiency of an input-output pair** $(x, y) \in \mathcal{T}$ *is*

$$\mathcal{RI}(x, y) := \min\{\theta : (\theta x, y) \in \mathcal{T}\}. \tag{9.1}$$

Definition 9.2. *The* **radial measure of output efficiency of an input-output pair** $(x, y) \in \mathcal{T}$ *is*

$$\mathcal{RO}(x, y) := \min\{\theta : (x, y/\theta) \in \mathcal{T}\}. \tag{9.2}$$

Both measures are well-defined under the standard assumptions of closed input and output possibility sets and the assumption that a given input vector cannot achieve unlimited output.

Radial measures of input and output efficiency possess the following simple properties. First, since $(x, y) \in \mathcal{T}$, both measures of efficiency are bounded above by one. Next, as previously discussed in Chapter 7, the input and output distance functions and the radial measures of input and output efficiency are related via the following identities:

$$\mathcal{RI}(x,y) = \mathcal{D}(x,y)^{-1}$$
$$\mathcal{RO}(x,y) = \mathcal{O}(x,y).$$

Finally, radial measures of input and output efficiency coincide for constant returns-to-scale technologies.

Proposition 9.3. *If $\mathcal{T}$ exhibits constant returns-to-scale*[1]*, then $\mathcal{RI}(x,y) = \mathcal{RO}(x,y)$.*

Proof. An immediate consequence of the fact that $(\theta x, y) \in \mathcal{T}$ if and only if $(x, y/\theta) \in \mathcal{T}$. □

Radial measures have two distinct benefits. First, they are conservative. Each seeks an *equiproportionate* projection: *reduction* (in the case of input) or *expansion* (in the case of output). A firm that is rated 75% input efficient is being told it is possible to reduce each of inputs by *at least* 25% and still achieve the same outputs. Similarly, a firm that is rated 80% output efficient is being told it is possible to expand each of outputs by *at least* 25% using the same level of inputs. Second, radial measures are *independent* of the unit of measurement for each of the inputs and outputs.

Radial measures have one serious deficiency. It is possible for $\mathcal{RI}(x,y) = 1$, suggesting the firm is technically input efficient, when it is also possible to find an $x' \lneqq x$ for which $(x', y) \in \mathcal{T}$. The firm will prefer to use (x', y) instead of (x, y). Similarly, it is possible for $\mathcal{RO}(x,y) = 1$, suggesting the firm is technically output efficient, when it is also possible to find a $y' \gneqq y$ for which $(x, y') \in \mathcal{T}$. The firm will prefer to use (x, y') instead of (x, y). Formally, it is possible that the radial measures of efficiency will *not* project onto the Efficient Frontier of the technology; informally, slacks can be present.

The deficiency of the radial measure of efficiency is overcome with the following notion of efficiency.

Definition 9.4. *A* **linear measure of input efficiency of an input-output pair** (x,y) *is*

$$\mathcal{LI}(x,y) := \min\Big\{ \sum_i w_i\theta_i : ((\theta_1 x_1, \theta_2 x_2, \ldots, \theta_n x_n), y) \in \mathcal{T} \Big\}, \tag{9.3}$$

where the weights w_i are nonnegative and sum to one.

An example of a linear measure of input efficiency is the Russell measure defined in Chapter 1—see (1.7), p. 6.

Definition 9.5. *A* **linear measure of output efficiency of an input-output pair** (x,y) *is*

$$\mathcal{LO}(x,y) := \min\Big\{ \sum_i w_i\theta_i : (x, (y_1/\theta_1, y_2/\theta_2, \ldots, y_n/\theta_n)) \in \mathcal{T} \Big\}, \tag{9.4}$$

where the weights w_i are nonnegative and sum to one.

[1] This means that $(x,y) \in \mathcal{T}$ if and only if $s(x,y) \in \mathcal{T}$ for all $s \geq 0$.

Linear measures of input or output efficiency cannot be higher than their radial measure counterparts, since the radial efficiency can be defined as a linear measure with two additional restrictions in either (9.3) or (9.4), namely, the w_i are equal and the θ_i are constrained to be equal, too.

When the constraints that characterize a technology are linear, the linear measures of input and output efficiency can be calculated via linear programming. The power of computing now makes solving reasonably-sized convex programming problems tractable. A natural generalization of the linear measure is a weighted measure.

Definition 9.6. *A convex function $f(\cdot)$ that maps $[0,1]$ onto $[0,1]$ such that $f(0) = 0$ and $f(1) = 1$ is an* **efficiency weighting function**.

Remark 9.7. An efficiency weighting function is necessarily nondecreasing.

Definition 9.8. *A* **weighted measure of input efficiency of an input-output pair** (x, y) *is*

$$\mathcal{WI}(x,y) := \min\Big\{\sum_i w_i f_i(\theta_i) : ((\theta_1 x_1, \theta_2 x_2, \ldots, \theta_n x_n), y) \in \mathcal{T}\Big\}, \qquad (9.5)$$

where the weights w_i are nonnegative and sum to one and each $f_i(\cdot)$ is an efficiency weighting function.

Definition 9.9. *A* **weighted measure of output efficiency of an input-output pair** (x, y) *is*

$$\mathcal{WO}(x,y) := \min\Big\{\sum_i w_i f_i(\theta_i) : (x, (y_1/\theta_1, y_2/\theta_2, \ldots, y_n/\theta_n)) \in \mathcal{T}\Big\}, \quad (9.6)$$

where the weights w_i are nonnegative and sum to one and each $f_i(\cdot)$ is an efficiency weighting function.

Remark 9.10. A weighted measure of input or output efficiency is additively-separable in each input or output. It is a special case of a convex measure of input or output efficiency defined by replacing $\sum_i w_i f_i(\theta_i)$ in (9.5) or (9.6) with a general convex function of the θ_i.

9.2 Scale Efficiency

Let

$$\mathcal{RI}^{CRS}(x,y),\ \mathcal{RO}^{CRS}(x,y),\ \mathcal{RI}^{VRS}(x,y),\ \mathcal{RO}^{VRS}(x,y)$$

denote the radial measures of input and output efficiency corresponding to the CRS and VRS technologies. Since $\mathcal{T}^{VRS} \subset \mathcal{T}^{CRS}$, it follows that

$$\mathcal{RI}^{VRS}(x,y) \leq \mathcal{RI}^{CRS}(x,y) \text{ and } \mathcal{RO}^{VRS}(x,y) \leq \mathcal{RO}^{CRS}(x,y).$$

By Proposition 9.3, $\mathcal{RI}^{CRS}(x,y) = \mathcal{RO}^{CRS}(x,y)$ but, in general, $\mathcal{RI}^{VRS}(x,y) \neq \mathcal{RO}^{VRS}(x,y)$.

Definition 9.11. *The* **input-based scale efficiency** *is the ratio*

$$\mathcal{SI}(x,y) := \frac{\mathcal{RI}^{CRS}(x,y)}{\mathcal{RI}^{VRS}(x,y)},$$

and the **output-based scale efficiency** *is the ratio*

$$\mathcal{SO}(x,y) := \frac{\mathcal{RO}^{CRS}(x,y)}{\mathcal{RO}^{VRS}(x,y)}.$$

It is useful to decompose overall input or output efficiency into the product of its pure technical efficiency component, $\mathcal{RI}^{VRS}(x,y)$ or $\mathcal{RO}^{VRS}(x,y)$, *when no assumption about returns-to-scale is made*, and its scale component, $\mathcal{SI}(x,y)$ or $\mathcal{SO}(x,y)$:

$$\mathcal{RI}^{CRS}(x,y) = \mathcal{RI}^{VRS}(x,y)\,\mathcal{SI}(x,y), \tag{9.7}$$

$$\mathcal{RO}^{CRS}(x,y) = \mathcal{RO}^{VRS}(x,y)\,\mathcal{SO}(x,y). \tag{9.8}$$

This decomposition defines overall input or output efficiency via the CRS technology.

Example 9.12. Suppose $\mathcal{RI}^{CRS}(x,y) = 0.54$ and $\mathcal{RI}^{VRS}(x,y) = 0.90$. Then $\mathcal{SI}(x,y) = 0.60$. In this case, the lion's share of input inefficiency is due to inappropriate scale, as opposed to pure technical input efficiency. Suppose, instead, that $\mathcal{RI}^{CRS}(x,y) = 0.54$ and $\mathcal{RI}^{VRS}(x,y) = 0.60$. Then $\mathcal{SI}(x,y) = 0.90$. In this case, the lion's share of input inefficiency is due to pure technical input inefficiency as opposed to inappropriate scale.

9.3 Cost Efficiency

We turn now to discussing concepts and measures of efficiency pertaining to cost when data

$$\mathcal{D} = \{(x_1,y_1,p_1),(x_2,y_2,p_2),\ldots,(x_N,y_N,p_N)\},$$

on inputs, outputs and prices for N firms are given.

A natural measure for *cost efficiency* is to take the ratio of minimum to actual cost.

Definition 9.13. Cost efficiency *is the ratio*

$$\mathcal{C}(x_i,y_i,p_i) := \frac{Q(y_i,p_i)}{p_i \cdot x_i}.$$

Cost inefficiencies arise from two sources: (i) inefficient use of input and (ii) incorrect choice of input mix based on factor prices. The first source of cost inefficiency is measured by $\mathcal{RI}(x,y)$. To isolate the second source of cost inefficiency, a natural approach is to take the ratio of the minimum cost to the cost of the technically input efficient vector $x/\mathcal{D}(x,y)$. Either technology $\mathcal{T}^{CRS}$ or $\mathcal{T}^{VRS}$ can be used to determine $\mathcal{RI}(x,y) = \mathcal{D}(x,y)^{-1}$.

Definition 9.14. Allocative efficiency *is the ratio*

$$\mathcal{A}(x_i, y_i, p_i) := \frac{Q(y_i, p_i)}{p_i \cdot \hat{x}_i},$$

where $\hat{x}_i := x_i/\mathcal{D}(x_i, y_i)$ *for each* i.

Note that $\mathcal{A}(x_i, y_i, p_i) = \mathcal{C}(x_i, y_i, p_i)\mathcal{D}(y_i, x_i)$.

It is useful to decompose cost efficiency into the product of its allocative component, $\mathcal{A}(x_i, y_i, p_i)$, and its input efficient component, $\mathcal{RI}(x, y) = \mathcal{D}(x, y)^{-1}$:

$$\mathcal{C}(x_i, y_i, p_i) = \mathcal{A}(x_i, y_i, p_i)\ \mathcal{D}(x, y)^{-1}.$$

Example 9.15. Figure 9.1 illustrates the decomposition of cost efficiency into its allocative and technical efficiency components. The point $\hat{x}$ is technically efficient but not allocatively efficient. The degree of allocative inefficiency is measured by the gap between the dotted line and the isocost line associated with minimum cost.

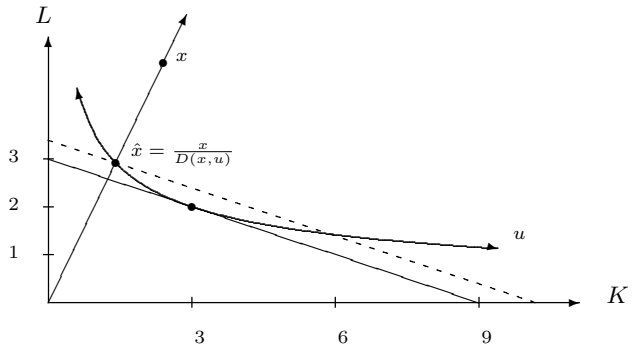


Fig. 9.1. Decomposition of cost efficiency into its allocative and technically efficient components.

9.4 Joint Input-Output Efficiency

We briefly mention the possibilities to define a joint input-output measure. Here is but one example:

Definition 9.16. Joint input-output efficiency *is*

$$\mathcal{H}^{\mathcal{T}}(x, y) := \min\{\theta :\ (\theta x, y/\theta) \in \mathcal{T}\}.$$

9.5 Computing Input Efficiency

We are given a data set $\mathcal{D}$ that contains input, output and price data for N firms. There are n inputs and m outputs. All factor prices are assumed positive. Some of the components of the input-output pair (x_i, y_i) can be zero, but no input or output vector is identically equal to zero. Following convention, (x_0, y_0) denotes the input-output pair for which a particular efficiency is being computed. It is commonly referred to as the **reference firm**. Keep in mind that (x_0, y_0) corresponds to (x_i, y_i) for some index i, $1 \leq i \leq N$.

For notational convenience, let

$$x(\lambda) := \sum_i \lambda_i x_i,\ y(\lambda) := \sum_i \lambda_i y_i,\ e := (1, 1, \ldots, 1) \in \mathbb{R}^N.$$

With this notation, the constraints $\sum_{i=1}^N \lambda_i x_i \leq \theta x_0$, $\sum_{i=1}^N \lambda_i y_i \geq y_0$, and $\sum_{i=1}^N \lambda_i = 1$ can be respectively expressed in compact form as $x(\lambda) \leq \theta x_0$, $y(\lambda) \geq y_0$ and $e(\lambda) = 1$. For the remainder of this chapter, all vectors z will be represented as *columns*. The *transpose* of vector z, denoted by z^T, is z expressed as a row vector.

9.5.1 *CRS* Technology

The radial measure of input efficiency, $\mathcal{RI}^{CRS}(x_0, y_0)$, can be computed via the linear program

$$(P^{CRS}): \qquad \min_{(\theta,\lambda)\geq 0} \{\theta :\ x(\lambda) \leq \theta x_0,\ y(\lambda) \geq y_0\}.$$

(P) can be represented in the canonical form

$$\min_{z\geq 0} \{c^T z :\ Az \geq b\}, \tag{9.9}$$

where

$$z^T := (\theta, \lambda_1, \lambda_2, \ldots, \lambda_N) \in \mathbb{R}_+^{N+1},$$

$$c^T := (1, 0, 0, \ldots, 0) \in \mathbb{R}_+^{N+1},$$

$$b := \begin{pmatrix} 0 \\ y_0 \end{pmatrix} \in \mathbb{R}^{n+m},$$

$$A := \begin{bmatrix} x_0 & -x_1 & -x_2 & \ldots & -x_N \\ 0 & y_1 & y_2 & \ldots & y_N \end{bmatrix}.$$

A is an $(n + m) \times (N + 1)$.

Let $\mu \in \mathbb{R}^n_+$ denote the dual variables associated with the input constraints, $\theta x_0 - x(\lambda) \geq 0$, let $\nu \in \mathbb{R}^m_+$ denote the dual variables associated with the output constraints, $y(\lambda) \geq y_0$, and let $\zeta^T := (\mu, \nu)$ denote the vector of dual variables. The dual linear program to the primal linear program (9.9) is

$$\max_{\zeta \geq 0} \{b^T \zeta : \ \zeta^T A \leq c^T\}. \tag{9.10}$$

Accordingly, the dual linear program of (P^{CRS}) is

$$(D^{CRS}): \qquad \max_{\mu \geq 0,\ \nu \geq 0} \{\nu^T y_0 : \ \nu^T y_i - \mu^T x_i \leq 0 \text{ for all } i, \ \mu^T x_0 = 1\}.^2 \tag{9.11}$$

There is a natural economic interpretation to (D^{CRS}). The components of the dual vector ν can be thought of as the prices (per unit revenues) of the outputs, whereas the components of the dual vector μ can be thought of as the prices (per unit cost) of the inputs. Consequently, the expression $\nu^T y_i - \mu^T x_i$ is the *profit* of firm i *given* this pricing system. The constraints of (D^{CRS}) simply say that for a pricing system associated with a constant returns-to-scale technology to be valid, the economic profit of each firm cannot be positive. (If the profit were positive, then the firm would scale its input-output vector arbitrarily high to make infinite profits.) Given a pricing system, the profit equation is unaffected by multiplying all prices by a positive scale, i.e., it is homogeneous of degree zero, and so, without loss of generality, prices can be normalized by setting the cost of the reference firm, $\mu^T x_0$, to one. The revenue of the reference firm will then correspond to its input efficiency.

The dual linear program (9.11) is equivalent to the following *linear fractional programming problem*

$$(LF^{CRS}): \qquad \max_{\mu \geq 0,\ \nu \geq 0} \left\{ \frac{\nu^T y_0}{\mu^T x_0} : \ \frac{\nu^T y_i}{\mu^T x_i} \leq 1 \text{ for all } i \right\}. \tag{9.12}$$

The interpretation of (LF^{CRS}) is related to scoring in multi-criteria decision-making, as follows. Efficiency is measured as the ratio of aggregate output to aggregate input. To (linearly) aggregate the vectors of input and output, one needs weights (the dual variables). Suppose the manager of a firm is permitted to *choose* the weights that will make his firm look as efficient as possible. However, the weights he chooses must ensure that the efficiency ratios of each firm is bounded above by one. The bound of one reflects the physical axiom that the transformation of input to output for any engineering system can result in a possible loss but cannot create a positive gain. The dual variables are referred to as **multipliers** or **weights**, and the dual formulation is referred to as the **multiplier formulation**.

[2] Since the dual variable of the last dual constraint $\mu^T x_0$ is θ, which is necessarily positive, complementary slackness implies this dual constraint must be tight.

Remark 9.17. Historically, Problem (LF^{CRS}) was defined first by Charnes, Cooper and Rhodes [1978] as the vehicle for defining input efficiency. Charnes and Cooper [1961] pioneered linear fractional programming, and so they knew how to transform problem (LF^{CRS}) into (D^{CRS}). Since an efficiency ratio remains unaffected if the dual variables are scaled (it is homogeneous of degree zero), the denominator $\mu^T x_0$ can be set to one, without loss of generality. Next, they took the dual to (D^{CRS}) to arrive at (P^{CRS}), and then interpreted the constraints as defining the smallest convex cone containing the original input-output data, namely, T^{CRS}.

9.5.2 *VRS* Technology

The radial measure of input efficiency, $\mathcal{RI}^{VRS}(x_0, y_0)$, can be computed via the linear program

$$(P^{VRS}_{input}): \qquad \min_{\theta\geq 0,\ \lambda\geq 0} \{\theta :\ x(\lambda) \leq \theta x_0,\ y(\lambda) \geq y_0,\ e(\lambda) = 1\}.$$

As before, let μ and ν denote the dual variables associated with the input and output constraints, respectively, and let η denote the dual variable associated with the constraint $e(\lambda) = \sum_i \lambda_i = 1$. The dual linear program of (P^{VRS}_{input}) is

$$(D^{VRS}_{input}): \qquad \max_{\mu\geq 0,\ \nu\geq 0} \{\nu^T y_0 + \eta :\ \nu^T y_i + \eta - \mu^T x_i \leq 0 \text{ for all } i,\ \mu^T x_0 = 1\}. \tag{9.13}$$

Since the dual variable η is associated with an *equality* constraint, there are no constraints on its sign in the dual formulation. It turns out that the sign of the optimal dual variable η^* can be used to characterize the returns-to-scale for an input-output pair that is both radially input and output efficient.

9.5.3 *HR* Technology

The radial measure of input efficiency, $\mathcal{RI}^{HR}(x_0, y_0)$, can be computed by solving the linear program

$$(P^{HR}_{input}): \qquad \min_{\theta\geq 0,\ \lambda\geq 0} \{\theta : \sum_{i\in\mathcal{I}(y_0)} \lambda_i x_i \leq \theta x_0,\ e(\lambda) = 1\}, \tag{9.14}$$

where $\mathcal{I}(y_0) := \{i :\ y_i \geq y_0\}$. Once again, μ and η respectively denote the dual variables associated with the input and equality constraints. The dual linear program of (P^{HR}_{input}) is

$$(D^{VRS}_{input}) \qquad \max_{\mu\geq 0}\{\eta :\ \mu^T x_0 \leq 1,\ -\mu^T x_i + \eta \leq 0 \text{ for each } i \in \mathcal{I}(y_0)\}, \tag{9.15}$$

which is equivalent to[3]

[3] Set $p = \mu/\eta$. (The dual variable η must be positive since the efficiency cannot be zero.) The first constraint can be expressed as $p^T x_0 \leq 1/\eta$. Maximizing η is equivalent to minimizing $p^T x_0$.

$$\min_{p\geq 0}\{p^T x_0 :\ p^T x_i \geq 1\}. \tag{9.16}$$

Recall that this dual linear program has been previously interpreted—see (5.40), p. 84 and (7.5), p. 113.

9.6 Computing Output Efficiency

CRS technology. As we have noted several times, the radial measures of input and output efficiency for a constant returns-to-scale technology are equal.

VRS technology. The output efficiency, $\mathcal{RO}^{VRS}(x_0, y_0)$, is the following linear program:

$$(P^{VRS}_{output}) \qquad \max_{\gamma\geq 0,\ \lambda\geq 0} \{\gamma :\ x(\lambda) \leq x_0,\ y(\lambda) \geq \gamma y_0,\ e(\lambda) = 1\}.^4 \tag{9.17}$$

HR technology. In the single-output setting, output efficiency is simply the ratio of the observed to maximum output, namely, $y_0/\Phi(x_0)$. Consequently, the computation of output efficiency is synonymous with computing the production function. For the *HR* technology, it is *not* possible to formulate a *single* linear program to compute $\Phi(x_0)$. Fortunately, a sequence of linear programs can be solved to compute $\Phi(x_0)$.

As we previously discussed in Section 7.1.1, the input distance function characterizes the technology, since

$$(x, y) \in \mathcal{T} \Longleftrightarrow \mathcal{D}(x, y) \geq 1.$$

Since $(x, y) \in \mathcal{T}$ if and only if $\Phi(x) \geq y$, it follows that

$$\Phi(x_0) = \max\{y_i :\ \mathcal{D}(x_i, y_i) \geq 1\}. \tag{9.18}$$

The input distance equals the reciprocal of the radial measure of input efficiency, which can be computed via the linear program (P^{HR}_{input}). Thus, at most N linear programs need to be solved to compute $\Phi(x_0)$.

Remark 9.18. For each x, the function $\mathcal{D}(x, \cdot)$ is a decreasing function of y, and so a bisection search algorithm can be used to solve for $\Phi(x_0)$. Computationally, roughly ln N (instead of N) linear programs will have to be solved. Keep in mind that one has to solve this problem for *each* firm in the database. Given the size of the typical data and the speed of today's computers, this gain in computational efficiency may be unnecessary.

Remark 9.19. It is possible to formulate a single *mathematical programming* problem to solve for $\Phi(x_0)$. Let $\mathcal{B}^N := \{0, 1\}^N$ denote the collection of all vectors in $\mathbb{R}^N$ whose components are either zero or one (i.e. binary). Assume

[4] The nonlinear program represented by minimizing θ is converted to a linear program by simply maximizing $\gamma := \theta^{-1}$.

the labels have been arranged so that $y_1 \leq y_2 \leq \ldots \leq y_N$. The following *binary-linear program* can be solved to compute $\Phi(x_0)$:

$$\max_{\lambda \geq 0,\ z \in \mathcal{B}^N} \{\sum_i z_i y_i :\ x(\lambda) \leq x_0,\ e(\lambda) = 1,\ \sum_i z_i = 1,\ \lambda_i \leq \sum_{j=1}^{i} z_j \text{ for each } i\}. \tag{9.19}$$

To see why the formulation (9.19) works, first note that there will be exactly one i for which $z_i^* = 1$, say i^*. The constraint $\lambda_i \leq \sum_{j=1}^{i} z_j$ ensures that $\lambda_i = 0$ for all $i < i^*$. Thus, $x(\lambda^*)$ will be a convex combination of input vectors that achieve at least output rate y_{i^*}. If some of the y_i are equal, it is always possible to set the z_i so that i^* will correspond to the lowest index, thus ensuring that $L^{HR}(y_{i^*})$ will include *all* input vectors x_i that achieve output rate y_{i^*} as required by the HR technology.

Given the size of the typical data and the speed of today's computers, it is possible to solve this binary-linear program very quickly, thus obviating the need to solve a sequence of linear programs.

9.7 Computing Cost Efficiency

The computation of cost efficiency boils down to solving for the minimal cost

$$Q(y_i, p_i) = \min\{p_i x :\ (x, y_i) \in \mathcal{T}\}.$$

The constraint set $(x, y_i) \in \mathcal{T}$ is linear for each of the CRS, VRS, and HR technologies, and so one may calculate $Q(y_i, p_i)$ by solving an appropriate linear program.

Remark 9.20. In the single-output setting, we showed in Chapter 5 how to graphically compute $Q(y, p)$ for the CRS and VRS technologies.

9.8 Computing Joint Input-Output Efficiency

For the computational models to follow in this section, let $\mu := \theta\lambda$, $\gamma := \theta^2$, and note that $\theta x(\lambda) = x(\mu)$ and $\theta y(\lambda) = y(\mu)$.

Consider first the CRS technology. The hyperbolic measure of efficiency, $\mathcal{H}^{CRS}(x_0, y_0)$, can be computed by solving the following nonlinear programming problem

$$\min_{\lambda \geq 0, \theta \geq 0} \{\theta :\ x(\lambda) \leq \theta x_0,\ y(\lambda) \geq y_0/\theta\}. \tag{9.20}$$

The optimal value of Problem (9.20) is the *square root* of the optimal value of the following *linear* program

$$\min_{\mu \geq 0, \gamma \geq 0} \{\gamma :\ x(\mu) \leq \gamma x_0,\ y(\mu) \geq y_0\}. \tag{9.21}$$

Remark 9.21. A comparison of (9.21) with (9.9) shows that the *hyperbolic measure of efficiency for the CRS technology is simply the square root of the radial measure of input or output efficiency.*

As for the VRS technology, the hyperbolic measure of efficiency, $\mathcal{H}^{VRS}(x_0, y_0)$, can be computed by solving the nonlinear programming problem

$$\min_{\lambda \geq 0, \theta \geq 0} \{\theta : \ x(\lambda) \leq \theta x_0, \ y(\lambda) \geq y_0/\theta, \ e(\lambda) = 1\}. \tag{9.22}$$

This is equivalent to the nonlinear program

$$\min_{\mu \geq 0, \theta \geq 0} \{\gamma : \ x(\mu) \leq \gamma x_0, \ y(\mu) \geq y_0, \ \sum_i \mu_i = \sqrt{\gamma}\}. \tag{9.23}$$

We now turn to the HR technology.

Proposition 9.22. $\mathcal{H}^{HR}(x_0, y_0) = \min_{1 \leq i \leq N} \{\max\{\mathcal{D}(x_0, y_i)^{-1}, \ y_0/y_i\}\}$.

Proof. Let

$$\rho_i := \max\{\mathcal{D}(x_0, y_i)^{-1}, \ y_0/y_i\}. \tag{9.24}$$

The definitions of ρ_i and $\mathcal{D}(x_0, y_i)$ imply that $(\rho_i x_0, y_0/\rho_i) \in \mathcal{T}^{HR}$. Consequently, $\mathcal{H}(x_0, y_0) \leq \rho_i$ for each index i, and so $\mathcal{H}(x_0, y_0) \leq \min_i \rho_i$. It remains to show the reverse inequality $\mathcal{H}(x_0, y_0) \geq \min_i \rho_i$.

To this end, pick a θ for which $(\theta x_0, y_0/\theta \in \mathcal{T}^{HR})$. Obviously, (i) $\theta x_0 \in L^{HR}(y_0/\theta)$. The way in which the HR technology is defined implies that (ii) $L^{HR}(y_0/\theta) = L(y_i)$ for some index i. By (i) and (ii), and the definition of the input distance function, it follows that $\theta \geq \mathcal{D}(x_0, y_i)^{-1}$. By (ii) alone, $y_0/\theta \geq y_i$ or, equivalently, $\theta \geq y_0/y_i$. Clearly, then, $\theta \geq \rho_i$ and so $\theta \geq \min_{1 \leq i \leq N} \rho_i$. Since θ was chosen arbitrarily, the result follows. □

9.9 Exercises

The exercises use the input-output data displayed in Table 9.1.

9.1. Assume the HR model of technology. For Firm 5:

(a) Determine the radial measure of input efficiency.
(b) Determine the radial measure of output efficiency.
(c) Determine the cost efficiency when the prices of capital and labor are equal.
(d) Determine the allocative efficiency when the prices of capital and labor are equal.
(e) Determine the linear measure of input efficiency (see Definition 9.4) when $w_K = w_L = 0.5$.

Table 9.1. Input-output data for Exercises.

Firm	Capital	Labor	Output
1	2	6	50
2	4	2	50
3	5	9	150
4	8	4	200
5	8	8	50

(f) Determine the weighted measure of input efficiency (see Definition 9.8) when $f_K(\theta_K) = \theta_K^2$, $f_L(\theta_L) = \theta_L^2$, $w_K = 0.80$ and $w_L = 0.20$.
(g) Determine the hyperbolic measure of efficiency.
(h) Write down the linear program to compute the radial measure of input efficiency.

9.2. Assume the CRS model of technology. For Firm 5:

(a) Determine the radial measure of input efficiency.
(b) Determine the radial measure of output efficiency.
(c) Determine the cost efficiency when the prices of capital and labor are equal.
(d) Determine the allocative efficiency when the prices of capital and labor are equal.
(e) Determine the linear measure of input efficiency (see Definition 9.4) when $w_K = w_L = 0.5$.
(f) Determine the weighted measure of input efficiency (see Definition 9.8) when $f_K(\theta_K) = \theta_K^2$, $f_L(\theta_L) = \theta_L^2$, $w_K = 0.80$ and $w_L = 0.20$.
(g) Determine the hyperbolic measure of efficiency.
(h) Write down the linear program to compute the radial measure of input efficiency.

9.3. Assume the VRS model of technology. For Firm 5:

(a) Determine the radial measure of input efficiency.
(b) Determine the radial measure of output efficiency.
(c) Determine the cost efficiency when the prices of capital and labor are equal.
(d) Determine the allocative efficiency when the prices of capital and labor are equal.
(e) Determine the linear measure of input efficiency (see Definition 9.4) when $w_K = w_L = 0.5$.
(f) Determine the weighted measure of input efficiency (see Definition 9.8) when $f_K(\theta_K) = \theta_K^2$, $f_L(\theta_L) = \theta_L^2$, $w_K = 0.80$ and $w_L = 0.20$.
(g) Determine the hyperbolic measure of efficiency.
(h) Write down the linear program to compute the radial measure of input efficiency.

9.10 Bibliographical Notes

The monographs of Fare et. al. [1985, 1994] provide numerous linear programming models to measure a vast array of efficiency measures. These books also contain a superb summary of the early pioneering work in this field, which includes Debreu [1951], Farrell [1957], Farrell and Fieldhouse [1962] and Afriat [1967, 1972]. See also Russell [1985] and Sengupta [1989]. The Russell measure of input efficiency was introduced in Fare and Lovell [1978]. Banker [1984] introduced the concept of the most productive scale size. Varian [1984] also discusses the nonparametric approach and connects it to Afriat's earlier works.

9.11 Solutions to Exercises

9.1 (a) The input possibility set $L^{HR}(50)$ is the convex, input free disposable hull of the input vectors $x_1 = (2, 6)$ and $x_2 = (4, 2)$. The line passing through these two points is $L = -2K + 10$. The labor-capital ratio of x_5 is $L = K$. The line $L = K$ intersects the line passing through points x_1 and x_2 at $(10/3, 10/3)$. Thus, the input efficiency is $(10/3)/8 = 5/12 = 0.416\overline{6}$.

(b) To determine output efficiency, we must identify the largest value of y_i for which $x_5 \in L^{HR}(y_i)$. Since $x_5 \geq x_4$ it immediately follows that $x_5 \in L^{HR}(200)$, and so output efficiency is $50/200 = 0.25$.

(c) The slope of the isocost line is minus the ratio of the factor prices, which is $-(p_K/p_L) = -1$. The minimum isocost line is a line with this slope that is tangent to $L^{HR}(50)$. The tangent point is $x_2 = (4, 2)$ and thus the isocost line is $L = -K + 6$. This line intersects the line $L = K$ at the point $(3, 3)$. Hence, the cost efficiency is $3/8 = 0.375$. (Alternatively, one can normalize prices so that $p_K = p_L = 1$, and so the cost efficiency is the ratio of minimum cost, which is 6, to the cost of x_5, which is 16.)

(d) By definition, cost efficiency equals the product of allocative efficiency and technical efficiency. Since cost efficiency = 0.375 and technical efficiency = 5/12, allocative efficiency is 0.9.

(e) All points $(K, L) \in L^{HR}(50)$ such that $(K, L) \leq (8, 8)$ are candidates to determine this linear measure of input efficiency. Here, the only points that matter belong to the Efficient Frontier, which is defined by the line segment joining points x_1 and x_2. Each point on this line segment is of the form $(K, -2K + 10)$ (as long as $2 \leq K \leq 4$). Thus, we seek to

$$\min \left\{ 0.5 \left(\frac{K}{8} + \frac{10 - 2K}{8} \right) : \ 2 \leq K \leq 4 \right\}.$$

Since the objective function is linear, at least one of the endpoints defining the line segment will be an optimal solution. The solution is easily seen to be $x_2 = (4, 2)$ with linear efficiency score of $0.5(4/8 + 2/8) = 0.375$. Note that this point does not equal $(10/3, 10/3)$, which is used to define the radial measure of input efficiency.

(f) Again, all points $(K, L) \in L^{HR}(50)$ such that $(K, L) \leq (8, 8)$ are candidates to determine this weighted measure of input efficiency. Here, the only points that matter belong to the Efficient Frontier, which is defined by the line segment joining points x_1 and x_2. Each point on this line segment is of the form $(K, -2K + 10)$ (as long as $2 \leq K \leq 4$). Thus, we seek to

$$\min \left\{ 0.8 \left(\frac{K}{8} \right)^2 + 0.2 \left(\frac{10 - 2K}{8} \right)^2 : \ 2 \leq K \leq 4 \right\}.$$

This problem is equivalent to minimizing the convex quadratic function $1.6K^2 - 8K + 20$ on $[2, 4]$. Setting the derivative equal to zero, the solution is

$K = 2.5$ and thus $L = -2K + 10 = 5$. The objective function value is 0.15625, but this does not have a practical interpretation as the radial measure of input efficiency does.

(g) We shall use (9.24). We have $\rho_i = 1$ for $i = 1, 2, 5$, since $y_5/y_i = 1$ in each of these cases. Consider ρ_4. The input possibility set $L^{HR}(200)$ is the input free disposable hull of x_4. The point x_5 lies on the boundary of this set (contained in the line $K = 8$). Thus, $\mathcal{D}(x_5, 200) = 1$ and so $\rho_4 = 1$, too. (Notice that $y_5/y_4 = 0.25$ does not matter here.) Finally, we consider ρ_3. To compute this, we need to determine the input possibility set $L^{HR}(150)$. It is easily seen to be the convex, input free disposable hull of the input vectors x_3 and x_4. The line passing through these two points is $L = -(5/3)K + 52/3$. The labor-capital ratio of x_5 is $L = K$. The line $L = K$ intersects the line passing through points x_3 and x_4 at $(6.5, 6.5)$. Thus, $\mathcal{D}^{-1}(x_5, 150) = 6.5/8 = 0.8125$. Since $y_5/y_3 = 1/3$, we have that $\rho_3 = 0.8125$. Hence, $\mathcal{H}^{HR}(x_5, y_5) = 0.8125$.

(h) The linear program to compute the radial measure of input efficiency is

$$\min \Big\{ \theta : \begin{aligned} 2\lambda_1 + 4\lambda_2 + 5\lambda_3 + 8\lambda_4 + 8\lambda_5 &\leq 8\theta, \\ 6\lambda_1 + 2\lambda_2 + 9\lambda_3 + 4\lambda_4 + 8\lambda_5 &\leq 8\theta, \\ \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5 &= 1, \\ \lambda_i \geq 0, \;\; 1 \leq i \leq 5 \Big\}. \end{aligned}$$

9.2 (a) The input possibility set $L^{CRS}(50)$ is the convex, input free disposable hull of the vectors $(y_5/y_i)x_i$, $i = 1, 2, \ldots, 5$. Here, firms 3 and 4 determine this set, i.e., $L^{CRS}(50)$ is the convex, input free disposable hull of the vectors $\hat{x}_3(50) := (50/150)(5, 9) = (5/3, 3)$ and $\hat{x}_4(50) := (50/200)(8, 4) = (2, 1)$. The line passing through these two points is $L = -6K + 13$. The labor-capital ratio of x_5 is $L = K$. The line $L = K$ intersects the line passing through points $\hat{x}_3(50)$ and $\hat{x}_4(50)$ at $(13/7, 13/7)$. Thus, the input efficiency is $(13/7)/8 = 13/56 = 0.2321$.

(b) Since the CRS model of technology satisfies constant returns-to-scale, the output and input efficiencies are the same. Thus, output efficiency is 0.2321.

(c) The slope of the isocost line is minus the ratio of the factor prices, which is $-(p_K/p_L) = -1$. The minimum isocost line is a line with this slope that is tangent to $L^{CRS}(50)$. The tangent point is $\hat{x}_4(50) = (2, 1)$ and thus the isocost line is $L = -K + 3$. This line intersects the line $L = K$ at the point $(1.5, 1.5)$. Hence, the cost efficiency is $1.5/8 = 3/16 = 0.1875$. (Alternatively, one can normalize prices so that $p_K = p_L = 1$, and so the cost efficiency is the ratio of minimum cost, which is 3, to the cost of x_5, which is 16.)

(d) By definition, cost efficiency equals the product of allocative efficiency and technical efficiency. Since cost efficiency = 3/16 and technical efficiency = 13/56, allocative efficiency is $21/26 = 0.8077$.

(e) All points $(K, L) \in L^{CRS}(50)$ such that $(K, L) \leq (8, 8)$ are candidates to determine this linear measure of input efficiency. Here, the only points that

matter belong to the Efficient Frontier, which is defined by the line segment joining points $\hat{x}_3(50)$ and $\hat{x}_4(50)$. Each point on this line segment is of the form $(K, -6K + 13)$ (as long as $5/3 \leq K \leq 2$). Thus, we seek to

$$\min\left\{0.5\left(\frac{K}{8} + \frac{13 - 6K}{8}\right) :\ 5/3 \leq K \leq 2\right\}.$$

Since the objective function is linear, at least one of the endpoints defining the line segment will be an optimal solution. The solution is easily seen to be $\hat{x}_4(50) = (2, 1)$ with linear efficiency score of $0.5(2/8 + 1/8) = 3/16$. Note that this point does not equal $(13/7, 13/7)$, which is used to define the radial measure of input efficiency.

(f) Again, all points $(K, L) \in L^{CRS}(50)$ such that $(K, L) \leq (8, 8)$ are candidates to determine this weighted measure of input efficiency. Here, the only points that matter belong to the Efficient Frontier, which is defined by the line segment joining points $\hat{x}_3(50)$ and $\hat{x}_4(50)$. Each point on this line segment is of the form $(K, -6K + 13)$ (as long as $5/3 \leq K \leq 2$). Thus, we seek to

$$\min\left\{0.8\left(\frac{K}{8}\right)^2 + 0.2\left(\frac{13 - 6K}{8}\right)^2 :\ 2 \leq K \leq 4\right\}.$$

This problem is equivalent to minimizing the convex quadratic function $8K^2 - 31.2K + 33.8$ on $[5/3, 2]$. Setting the derivative equal to zero, the solution is $K = 1.95$ and thus $L = -6K + 13 = 1.30$. The objective function value is 0.0528, but this does not have a practical interpretation as the radial measure of input efficiency does.

(g) Following Remark 9.21, $\mathcal{H}^{CRS}(x_5, y_5) = \sqrt{0.2321} = 0.4818$.

(h) The linear program to compute the radial measure of input efficiency is

$$\begin{aligned}\min\Big\{\theta :\ & 2\lambda_1 + 4\lambda_2 + 5\lambda_3 + 8\lambda_4 + 8\lambda_5 \leq 8\theta,\\ & 6\lambda_1 + 2\lambda_2 + 9\lambda_3 + 4\lambda_4 + 8\lambda_5 \leq 8\theta,\\ & 50\lambda_1 + 50\lambda_2 + 150\lambda_3 + 200\lambda_4 + 50\lambda_5 \geq 50,\\ & \lambda_i \geq 0,\ \ 1 \leq i \leq 5\Big\}.\end{aligned}$$

9.3 (a) The input possibility set $L^{VRS}(50)$ is identical to $L^{HR}(50)$. Hence, the radial measure of input efficiency is the same and equals $0.41\overline{6} = 5/12$.

(b) The input possibility set $L^{VRS}(200)$ is the input free disposable hull of the input vector x_4. As $x_5 \geq x_4$ it follows that $x_5 \in L^{VRS}(200)$. Since 200 is the highest output, we conclude that the output efficiency is $50/200 = 0.25$. (Not surprisingly, since $L^{VRS}(200)$ is identical to $L^{HR}(200)$.)

(c)-(f) The cost, allocative, linear, and weighted input efficiencies are identical to those assuming the HR model of technology, since $L^{VRS}(50)$ is identical to $L^{HR}(50)$.

(g) The nonlinear program to compute the hyperbolic measure of efficiency is

$$\begin{aligned}\min\Big\{\theta:\ & 2\lambda_1+4\lambda_2+5\lambda_3+8\lambda_4+8\lambda_5\le 8\theta,\\ & 6\lambda_1+2\lambda_2+9\lambda_3+4\lambda_4+8\lambda_5\le 8\theta,\\ & 50\lambda_1+50\lambda_2+150\lambda_3+200\lambda_4+50\lambda_5\ge 50/\theta,\\ & \lambda_1+\lambda_2+\lambda_3+\lambda_4+\lambda_5=1,\\ & \lambda_i\ge 0,\ \ 1\le i\le 5\Big\}.\end{aligned}$$

The optimal solution for λ is $\lambda_1^* = 0.429285$, $\lambda_2^* = 0.282860$, and $\lambda_4^* = 0.287855$. This identifies the point $((4.29285, 4.29285), 93.17821) \in \mathcal{T}^{VRS}$ to which (x_5, y_5) is being compared. The implied (optimal) θ is $4.29285/8 = 50/93.17821 = 0.53661$.

Remark 9.23. With the construction of the *two-dimensional projection* (see Chapter 10), this calculation will be almost trivial!

(h) The linear program to compute the radial measure of input efficiency is

$$\begin{aligned}\min\Big\{\theta:\ & 2\lambda_1+4\lambda_2+5\lambda_3+8\lambda_4+8\lambda_5\le 8\theta,\\ & 6\lambda_1+2\lambda_2+9\lambda_3+4\lambda_4+8\lambda_5\le 8\theta,\\ & 50\lambda_1+50\lambda_2+150\lambda_3+200\lambda_4+50\lambda_5\ge 50,\\ & \lambda_1+\lambda_2+\lambda_3+\lambda_4+\lambda_5=1,\\ & \lambda_i\ge 0,\ \ 1\le i\le 5\Big\}.\end{aligned}$$

10

The Two-Dimensional Projection

Much of the conceptual understanding of DEA models of technology and their corresponding efficiency measures can be understood by examining a two-dimensional projection embedded in the technology. A two-dimensional projection is an example of a well-behaved single-input, single-output technology. We shall show how to *compute* this projection, thereby permitting a graphical determination of the input, output, and joint input-output efficiency measures from a *single* graph in the plane.

10.1 Definition

Definition 10.1. *The* **two-dimensional projection associated with** $(x_0, y_0) \in \mathcal{T}$ *is the set*

$$\mathcal{T}(x_0, y_0) := \{(\alpha, \beta) \in I\!R^2_+ : (\alpha x_0, \beta y_0) \in \mathcal{T}\}.$$

The two-dimensional projection is defined *relative* to a particular point in the technology. It is never empty as it always contains the point $(1, 1)$.

The variable α defines an input scale on x_0 and the variable β defines an output scale on y_0. As such, it will be useful in what follows to think of the two-dimensional projection as defining a single-input, single-output technology with α denoting the level of input and β denoting the level of output. To formalize this interpretation, define

$$\mathcal{P}(x_0, y_0) := \{(x, y) \in \mathbb{R}^n_+ \times \mathbb{R}^m_+ : \\ x = \alpha x_0,\ y = \beta y_0 \text{ for some } (\alpha, \beta) \in \mathbb{R}^2_+\}. \quad (10.1)$$

Each point in $P(x_0, y_0)$ can be identified with a point in the (α, β)-plane. With this identification in mind,

$$\mathcal{T}(x_0, y_0) = \mathcal{P}(x_0, y_0) \cap \mathcal{T}. \quad (10.2)$$

Technically, $\mathcal{P}(x_0, y_0) \cap \mathcal{T} \subset \mathbb{R}^n_+ \times \mathbb{R}^m_+$, but since every point in this set is uniquely characterized by two scalars, α and β, we may think of it as a subset of the two-dimensional (α, β) space. With this understanding there is no need to formally write down this identification.

10.2 Characterizations

The basic properties of the two-dimensional section are inherited from the technology itself.

Proposition 10.2. *A two-dimensional projection inherits the properties of closure, convexity and free disposability from the technology set* $\mathcal{T}$.[1]

Proof. The properties of closure and convexity are a direct consequence of the fact that the set $\mathcal{T} \cap \mathcal{P}(x_0, y_0)$ is both closed and convex. The property of free disposability is obvious. □

Definition 10.3. *The* **Efficient Frontier of the two-dimensional projection** $\mathcal{T}(x_0, y_0)$ *is the collection of points* $(\alpha, \beta) \in \mathcal{T}(x_0, y_0)$ *such that there does not exist an* $(\alpha', \beta') \in \mathcal{T}(x_0, y_0)$ *for which* $\alpha' \leq \alpha$ *and* $\beta' \geq \beta$ *and* $(\alpha', \beta') \neq (\alpha, \beta)$.

Proposition 10.4. *If the technology set* $\mathcal{T}$ *is closed and exhibits free disposability, then each two-dimensional projection* $\mathcal{T}(x_0, y_0)$ *is the free disposable hull of its efficient frontier.*

We proceed to characterize the two-dimensional projections derived from the VRS, CRS and HR technologies. Define

$$\alpha_m := \min\{\alpha : \ (\alpha, \beta) \in \mathcal{T}^{VRS}(x_0, y_0)\}, \tag{10.3}$$

$$\beta_m := \max\{\beta : \ (\alpha_m, \beta) \in \mathcal{T}^{VRS}(x_0, y_0)\}, \tag{10.4}$$

$$\beta_M := \max\{\beta : \ (\alpha, \beta) \in \mathcal{T}^{VRS}(x_0, y_0)\}, \tag{10.5}$$

$$\alpha_M := \min\{\alpha : \ (\alpha, \beta_M) \in \mathcal{T}^{VRS}(x_0, y_0)\}. \tag{10.6}$$

The maxima and minima are achieved since (i) the output constraints, $y(\lambda) \geq \beta y_0$, guarantee there is a finite bound on β, and (ii) $\mathcal{T}^{VRS}(x_0, y_0)$ is closed (Proposition 10.2).

Definition 10.5. *The* **production function** $\Phi^{VRS}_{(x_0,y_0)} : [\alpha_m, \alpha_M] \to I\!R_+$ **derived from the two-dimensional projection associated with the VRS technology** *is*

$$\Phi^{VRS}_{(x_0,y_0)}(\alpha) := \max\{\beta : \ (\alpha, \beta) \in \mathcal{T}^{VRS}(x_0, y_0)\}.$$

[1] A two-dimensional projection exhibits free disposability whenever $(\alpha, \beta) \in \mathcal{T}(x_0, y_0)$ and $\alpha' \geq \alpha$ and $\beta' \leq \beta$, then $(\alpha', \beta') \in \mathcal{T}(x_0, y_0)$, too.

Let

$$Gr(\Phi^{VRS}_{(x_0,y_0)}) := \{(\alpha, \Phi^{VRS}_{(x_0,y_0)}(\alpha))\}$$

denote the graph of the function $\Phi^{VRS}_{(x_0,y_0)}(\cdot)$. The following proposition follows directly from the definitions.

Proposition 10.6. *$Gr(\Phi^{VRS}_{(x_0,y_0)})$ is the Efficient Frontier of $\mathcal{T}^{VRS}(x_0, y_0)$ and $\mathcal{T}^{VRS}(x_0, y_0)$ is the free disposable hull of $Gr(\Phi^{VRS}_{(x_0,y_0)})$.*

Define the convex polytope[2]

$$\begin{aligned} C := \{(\lambda, \alpha, \beta) \geq 0 :\ & x(\lambda) \leq \alpha x_0,\ y(\lambda) \geq \beta y_0, \\ & e(\lambda) = 1,\ \alpha \leq \alpha_M,\ \beta \leq \beta_M\}. \end{aligned} \tag{10.7}$$

Let $P(\cdot)$ denote the *projection* of $\mathbb{R}^N_+ \times \mathbb{R}_+ \times \mathbb{R}_+$ onto $\mathbb{R}_+ \times \mathbb{R}_+$; that is, if $c = (\lambda, \alpha, \beta)$, then $P(c) = (\alpha, \beta)$. It follows from the definitions that

$$\mathcal{T}^{VRS}(x_0, y_0) = P(C). \tag{10.8}$$

Proposition 10.7. *The production function $\Phi^{VRS}_{(x_0,y_0)}(\cdot)$ is nondecreasing, concave and piecewise linear. If $\Phi^{VRS}_{(x_0,y_0)}(\cdot)$ is not constant, then it is an increasing function.*

Proof. Strong disposability of input immediately implies that $\Phi^{VRS}_{(x_0,y_0)}(\cdot)$ is nondecreasing. As for concavity, pick distinct points $\alpha_i \in [\alpha_m, \alpha_M]$, $i = 1, 2$, and a $\lambda \in [0, 1]$. By definition, for each $i = 1, 2$,

$$(\alpha_i, \Phi^{VRS}_{(x_0,y_0)}(\alpha_i)) \in \mathcal{T}^{VRS}(x_0, y_0).$$

Since $\mathcal{T}^{VRS}(x_0, y_0)$ is convex,

$$\Big(\lambda\alpha_1 + (1-\lambda)\alpha_2,\ \lambda\Phi^{VRS}_{(x_0,y_0)}(\alpha_1) + (1-\lambda)\Phi^{VRS}_{(x_0,y_0)}(\alpha_2)\Big) \in \mathcal{T}^{VRS}(x_0, y_0),$$

which immediately implies that

$$\Phi^{VRS}_{(x_0,y_0)}(\lambda\alpha_1 + (1-\lambda)\alpha_2)) \geq \lambda\Phi^{VRS}_{(x_0,y_0)}(\alpha_1) + (1-\lambda)\Phi^{VRS}_{(x_0,y_0)}(\alpha_2).$$

This establishes concavity of $\Phi^{VRS}_{(x_0,y_0)}(\cdot)$.

It remains to establish that $\Phi^{VRS}_{(x_0,y_0)}(\cdot)$ is piecewise linear and an increasing function if it is not constant. Let V denote the subset of extreme points of $\mathcal{T}^{VRS}(x_0, y_0)$ that belong to $Gr(\Phi^{VRS}_{(x_0,y_0)})$. The set V is nonempty since it contains the point (α_M, β_M). It follows from (10.8) that each $v \in V$ must be the projection of some extreme point of C. Since C is a polytope, it has only a finite number of extreme points. Consequently, the set V is finite. Since

[2] The constraint $e(\lambda) = 1$ can be expressed via two inequalities, namely, $e(\lambda) \geq 1$ and $e(\lambda) \leq 1$.

$Gr(\Phi^{VRS}_{(x_0,y_0)})$ contains only a finite number of extreme points, it consists of a sequence of line segments, the endpoints of which belong to V. This shows that $\Phi^{VRS}_{(x_0,y_0)}(\cdot)$ is piecewise linear.

Finally, suppose $\Phi^{VRS}_{(x_0,y_0)}(\cdot)$ is not constant and is not increasing. Given that $\Phi^{VRS}_{(x_0,y_0)}(\cdot)$ is nondecreasing, there must exist $\alpha_1 < \alpha_2$ such that

$$\Phi^{VRS}_{(x_0,y_0)}(\alpha_1) = \Phi^{VRS}_{(x_0,y_0)}(\alpha_2) := \beta.$$

By definition of α_M, β_M, it must be the case that $\beta < \beta_M$. The line segment joining (α_1, β) to (α_M, β_M) must belong to $\mathcal{T}^{VRS}(x_0, y_0)$. Since the line segment also lies above the point (α_2, β), it follows that $\beta < \Phi(\alpha_2)$, a clear contradiction. □

Remark 10.8. $\mathcal{T}^{VRS}(x_0, y_0)$ is a convex polyhedron, since it is the intersection of the halfspaces defined by the lines containing adjacent points in V and the nonnegativity constraints on α and β. The points in V define the vertices of this convex polyhedron.

We now provide a geometric characterization of $\mathcal{T}^{CRS}(x_0, y_0)$.

Proposition 10.9. *The two-dimensional projection associated with the CRS technology is the constant returns-to-scale hull of the two-dimensional projection associated with the VRS technology, i.e.,*

$$\mathcal{T}^{CRS}(x_0, y_0) = \mathcal{CRS}(\mathcal{T}^{VRS}(x_0, y_0)).$$

Moreover, there exists an $m > 0$ such that

$$\mathcal{T}^{CRS}(x_0, y_0) = \{(\alpha, \beta) \in I\!R^2 : \ 0 \le \beta \le m\alpha\},$$

and the line $y = mx$ is tangent to the boundary of $\mathcal{T}^{VRS}(x_0, y_0)$.

Proof. It follows from the definitions that

$$\mathcal{T}^{VRS}(x_0, y_0) = \{(\alpha, \beta) : \ x(\lambda) \le \alpha x_0, \ y(\lambda) \ge \beta y_0, \ e(\lambda) = 1, \ \lambda_i \ge 0\},$$
$$\mathcal{T}^{CRS}(x_0, y_0) = \{(\alpha, \beta) : \ x(\mu) \le \alpha x_0, \ y(\mu) \ge \beta y_0, \ \mu_i \ge 0\}.$$

Consequently, if $(\alpha, \beta) \in \mathcal{T}^{VRS}(x_0, y_0)$ and $s > 0$, then $s(\alpha, \beta) \in \mathcal{T}^{CRS}$ (set $\mu_i = s\lambda_i$). This shows that

$$\mathcal{CRS}(\mathcal{T}^{VRS}(x_0, y_0)) \subset \mathcal{T}^{CRS}(x_0, y_0).$$

Conversely, if $(\alpha, \beta) \in \mathcal{T}^{CRS}(x_0, y_0)$ and $\mu \neq 0$, then $(1/\sum_i \mu_i)(\alpha, \beta) \in \mathcal{T}^{VRS}$. This shows that

$$\mathcal{T}^{CRS}(x_0, y_0) \subset \mathcal{CRS}(\mathcal{T}^{VRS}(x_0, y_0)),$$

which establishes the first claim.

Since $\mathcal{T}^{CRS}(x_0, y_0)$ is a constant returns-to-scale, free disposable technology in $\mathbb{R}^2_+$,

$$\mathcal{T}^{CRS}(x_0, y_0) = \{(\alpha, \beta) : \ 0 \leq \beta \leq \hat{m}\alpha\}$$

for some finite positive value $\hat{m}$. Since the constant returns-to-scale technology $\mathcal{T}^{CRS}(x_0, y_0)$ contains $\mathcal{T}^{VRS}(x_0, y_0)$, it must contain C, and so $\hat{m} \geq m$. It remains to show the reverse inequality $\hat{m} \leq m$. Pick a point (α, β) for which $\beta = \hat{m}\alpha$, $\alpha \neq 0$. Since $(\alpha, \beta) \in \mathcal{T}^{CRS}(x_0, y_0)$ there exists a nonnegative λ for which $x(\lambda) \leq \alpha x_0$ and $y(\lambda) \geq \beta y_0$. Obviously, $\lambda \neq 0$. Now define $s := (\sum_i \lambda_i)^{-1}$ and $\mu_i := s\lambda_i$ for each i. Note that $\sum_i \mu_i = 1$. It follows that $(s\alpha, s\beta) \in \mathcal{T}^{VRS}(x_0, y_0)$ and that $\hat{m} \leq m$, as required. □

Proposition 10.10. *Fix $(x_0, y_0) \in \mathcal{T}^{HR}$. The two-dimensional projection $\mathcal{T}^{HR}(x_0, y_0)$ is the free disposable hull of the set $\{(\alpha_i, \beta_i),\ 1 \leq i \leq N\}$, where $(\alpha_i, \beta_i) := (\mathcal{D}(x_0, y_i)^{-1},\ y_i/y_0)$ for each i.*

Proof. It follows from the definition of the input distance function that each $(\alpha_i, \beta_i) \in \mathcal{T}^{HR}(x_0, y_0)$. Consequently, the free disposable hull of the (α_i, β_i) is contained in $\mathcal{T}^{HR}(x_0, y_0)$. It remains to show the reverse inclusion. Pick an $(\alpha, \beta) \in \mathcal{T}^{HR}(x_0, y_0)$. Since $\alpha x_0 \in L^{HR}(\beta y_0)$, obviously $\alpha \geq \mathcal{D}(x_0, y_i)^{-1}$. Moreover, the HR technology implies that $L^{HR}(\beta y_0) = L^{HR}(y_i)$ for some index i and so $\beta \leq y_i/y_0$. Thus, (α, β) belong to the free disposable hull of (α_i, β_i), and the result now follows. □

10.3 Computing Efficiency

Given the two-dimensional projection the computation of the radial measures of input and output efficiency are easily obtained via the following identities:

$$\mathcal{RI}(x_0, y_0) := \min\{\alpha : \ (\alpha, 1) \in \mathcal{T}(x_0, y_0)\}, \tag{10.9}$$

$$\mathcal{RO}(x_0, y_0) := [\max\{\beta : \ (1, \beta) \in \mathcal{T}(x_0, y_0)\}]^{-1}. \tag{10.10}$$

Remark 10.11. We have previously argued that the input and output efficiencies associated with $\mathcal{T}^{CRS}$ are equal. This fact follows directly from the two-dimensional projection, since $(\mathcal{RI}(x_0, y_0), 1)$ and $(1, \mathcal{RO}(x_0, y_0)^{-1})$ must both lie on the line $y = \hat{m}x$. In particular,

$$\hat{m} = \mathcal{RI}(x_0, y_0)^{-1} = \mathcal{RO}(x_0, y_0)^{-1}.$$

The hyperbolic measure of efficiency can also be *easily* computed from the two-dimensional projection. This is because

$$\mathcal{H}(x_0, y_0) = \max\{\alpha : \ (\alpha, 1/\alpha) \in \mathcal{T}(x_0, y_0)\}.$$

To graphically compute the hyperbolic measure, simply find the point of intersection of the curve $\beta = 1/\alpha$ with the boundary of the two-dimensional projection.

10.4 Scale Characterizations

Definition 10.12. *The* **most productive scale size (mpss)** *is the scalar value*

$$s^* := \max\{\beta/\alpha : (\alpha x_0, \beta y_0) \in \mathcal{T}^{VRS}\}.$$

Proposition 10.13. *The reciprocal of the mpss equals the input and output efficiencies for* $\mathcal{T}^{CRS}$.

Proof. Let $\hat{m}$ be the slope of the line $y = \hat{m}x$ that characterizes $\mathcal{T}^{CRS}(x_0, y_0)$. Clearly, the mpss equals $\hat{m}$, and $\mathcal{RI}(x_0, y_0) = \hat{m}^{-1}$, as noted in Remark 10.11.

For the following definition, let (α_i, β_i), $i = 1, 2$, denote the endpoints of the line segment contained in the intersection of $\mathcal{T}^{VRS}(X_0, Y_0)$ with $\mathcal{T}^{CRS}(X_0, Y_0)$. Without loss of generality, we assume that $(\alpha_1, \beta_1) \leq (\alpha_2, \beta_2)$. (Typically, the line segment collapses to a single point, in which case $\beta_1 = \beta_2$.)

Definition 10.14. *Assume that* $(1, 1)$ *lies on the boundary of* $\mathcal{T}^{VRS}(x_0, y_0)$. *If* $1 < \beta_1$, *then* (x_0, y_0) *exhibits increasing returns to scale. If* $1 > \beta_2$, *then* (x_0, y_0) *exhibits decreasing returns to scale. Finally, if* $\beta_1 \leq 1 \leq \beta_2$, *then* (x_0, y_0) *exhibits constant returns to scale.*

10.5 Example

The example data are displayed in Figure 4.4 on p. 61, which was used to generate the input possibility sets for the HR, CRS and VRS technologies displayed in Figures 4.5, 4.7 and 4.8, respectively. We shall explain how to generate the two-dimensional projection associated with the data point $(x_2, y_2) = ((5, 4), 10)$ for each technology.

We begin with the VRS technology. The proof of Proposition 10.6 established that $\mathcal{T}^{VRS}(x_2, y_2)$ is a convex polyhedron and is the free disposable hull of $Conv(V)$, where V denotes the finite set of extreme points or vertices of $\mathcal{T}^{VRS}(x_2, y_2)$. Obviously, if we can generate a finite set V' that contains V, then $\mathcal{T}^{VRS}(x_2, y_2)$ will be the free disposable hull of $Conv(V')$, too. We shall compute a set of points that contains *all* vertices of the C defined in (10.7), and take V' to be the projection of these points onto the (α, β) plane. Since each vertex of $\mathcal{T}^{VRS}(x_2, y_2)$ is the projection of some vertex of C, it directly follows that $V \subseteq V'$. One may think of the set V' as candidate vertices.

The set $\mathcal{T}^{VRS}(x_2, y_2)$ is the projection of the set C onto the (α, β) plane (see Definitions 10.7 and 10.8). Let $r \in \mathbb{R}^2_+$ denote the vector of slack variables associated with the input constraints defined so that $x(\lambda) + r = \alpha x_0$, and let $q \geq 0$ denote the slack variable associated with the output constraint defined so that $y(\lambda) - q = \beta y_0$. Define

$$\hat{C} := \{(\alpha, \beta, \lambda, r, q) \geq 0 : \; x(\lambda) + r = \alpha x_0, \; y(\lambda) - q = \beta y_0, \; e(\lambda) = 1\}. \quad (10.11)$$

Clearly, there is a one-to-one correspondence between the points $(\alpha, \beta, \lambda) \in C$ to the points $(\alpha, \beta, \lambda, r, q) \in \hat{C}$. Let p denote a vertex of C and let $\hat{p}$ denote the point in $\hat{C}$ that corresponds to p. Since there are four constraints that define $\hat{C}$ (two inputs, one output, and the constraint on λ), the vector $\hat{p}$ will have at most four positive coordinates.[3] Since the coordinates α and β of $\hat{p}$ will be positive, we conclude that *at most two* of the λ_i components of $\hat{p}$ can be positive.

First consider the case when two of the λ_i components of $\hat{p}$ are positive, say components k and l. In this case, *none* of the slack variables can be positive; in particular, this means that $x(\lambda) = \lambda_k x_k + \lambda_l x_l = \alpha x_2$ and $y(\lambda) = \beta y_2$. In order for $x(\lambda) = \alpha x_2$, exactly one of the vectors x_k or x_l must lie *above* the ray $\mathcal{R}(x_2) = \{sx : \ s \geq 0\}$ while the other vector must lie *below* the ray $\mathcal{R}(x_2)$. (Neither point can lie on the ray; otherwise, both points would lie on the ray, which in turn would imply that p is not a vertex.) In Figure 10.1, there are three points that lie above the ray (points 3, 5 and 6) and two points that lie below the ray (points 1 and 4). Consequently, there are *six* line segments that will intersect the ray as shown in the figure. The intersection of each line segment with the ray determines a unique λ vector (and thus α and β) that in turn generates a possible vertex of C. Table 10.1 shows the λ values as well as the corresponding α values and β values for the VRS and HR technologies.

Example 10.15. Consider the line segment joining points 1 and 6. It intersects the ray at point $x_{16} = 0.45(4, 1) + 0.55(4, 5) = (4, 3.2)$. Since $(4, 3.2) = 0.8(5, 4)$, the α value equals 0.80. Under the VRS technology the corresponding y value is $0.45(10) + 0.55(30) = 21$, which implies that $\beta_{VRS} = 21/10 = 2.10$. Under the more conservative HR technology, the corresponding y value is the *minimum* of the y-values of 10 and 30, and this is why $\beta_{HR} = 1.0$.

It is possible that a vertex p has only one positive λ_i value. To cover this possibility, for each index i we set $\lambda_i = 1$ and find the smallest value of α_i so that $x_i \leq \alpha_i x_2$. Obviously, $\alpha_i = \max\{x_i^k / x_2^k, \ k = 1, 2\}$. The corresponding y-value is simply y_i. We denote such points by x_i^+ and the relevant data are shown in Table 10.1.

Example 10.16. Consider $x_4 = (3, 2)$. Here, $\alpha_4 = \max\{3/5, 2/4\} = 0.60$. Of course, $(3, 2) \leq 0.60(5, 4) = (3, 2.4) := x_4^+$ with equality in the first input. The y-value here is $y_4 = 14$ and so $\beta_{VRS} = \beta_{HR} = 1.40$.

The (α, β) points in Table 10.1 corresponding to the VRS and HR technologies are plotted in Figure 10.2, which also depicts $\mathcal{T}^{VRS}(x_2, y_2)$, $\mathcal{T}^{HR}(x_2, y_2)$ and $\mathcal{T}^{CRS}(x_2, y_2)$, as well as the radial input and output efficiency measures and the hyperbolic measure of efficiency.

[3] A well-known fact of linear programming. See Theorem C.29, p. 468.

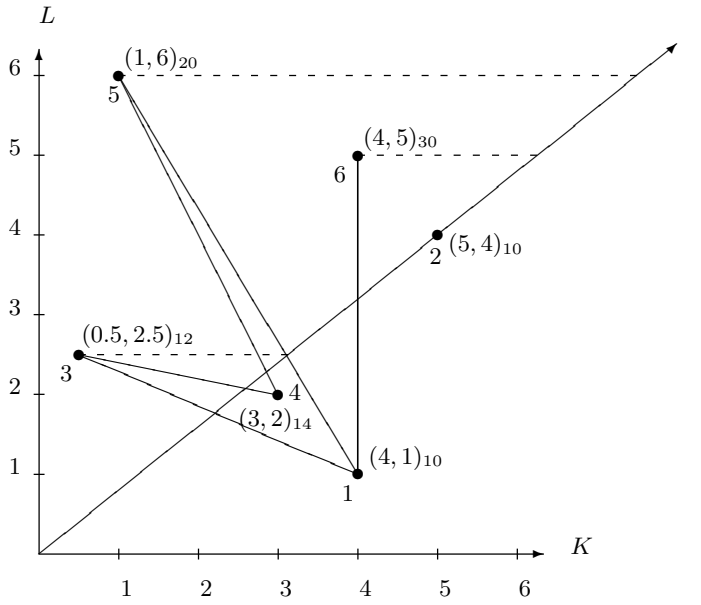


Fig. 10.1. Graphical determination of the two-dimensional projection for data point 2.

Table 10.1. Computing the two-dimensional projection.

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	α	β_{VRS}	β_{HR}
x_{13}	0.4884		0.5116				0.4419	1.102	1.000
x_{43}			0.1600	0.8400			0.5200	1.368	1.200
x_{45}				0.9286	0.0714		0.5714	1.443	1.400
x_{15}	0.7027				0.2973		0.6216	1.297	1.000
x_{46}				0.4737		0.5263	0.7053	2.242	1.400
x_{16}	0.4500					0.5500	0.8000	2.100	1.000
x_1^+	1.0000						0.8000	1.000	1.000
x_2		1.000					1.0000	1.000	1.000
x_3^+			1.0000				0.6250	1.200	1.200
x_4^+				1.000			0.6000	1.400	1.400
x_5^+					1.0000		1.5000	2.000	2.000
x_6^+						1.0000	1.2500	3.000	3.000

Remark 10.17. The boundary of $\mathcal{T}^{VRS}(x_2, y_2)$ (excluding the points where $\beta = 0$) is characterized by a finite number of adjacent line segments whose slopes are decreasing. The slope of the first line segment is infinity—this will always be the case when the $y_i > 0$—and the slope of the last line segment (technically, a ray) is zero.

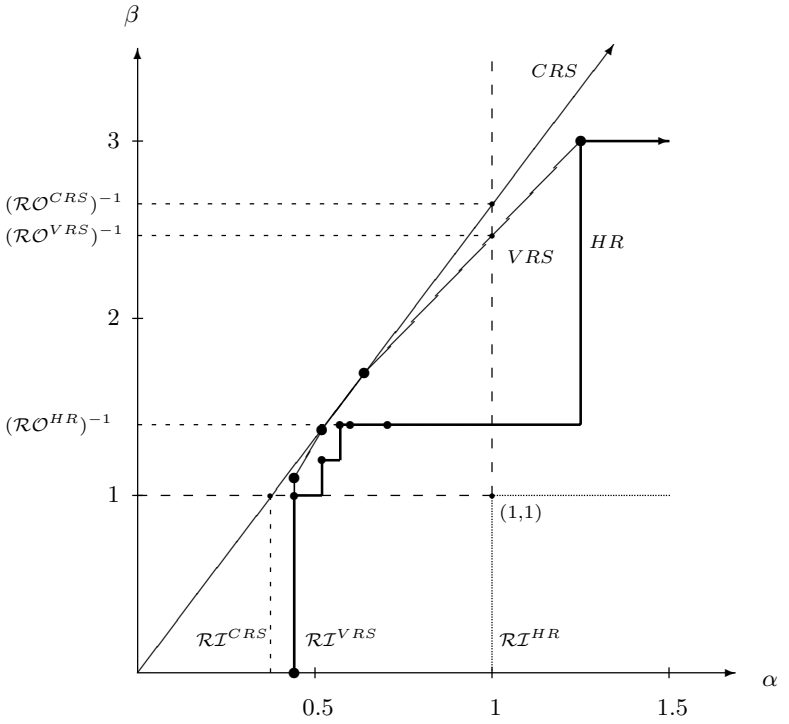


Fig. 10.2. The two-dimensional section for data point 2.

Remark 10.18. The two-dimensional projection $\mathcal{T}^{HR}(x_2, y_2)$ is *not* convex. The input possibility sets are each convex, however, as they should. Each two-dimensional projection of an HR technology will have the "staircase" shape as depicted in the figure.

As shown in the figure, the points $(0.44, 1)$ and $(1, 2.5)$ lie on the boundary of $\mathcal{T}^{VRS}(x_2, y_2)$, and so $\mathcal{RI}^{VRS}(x_2, y_2) = 0.44$ and $\mathcal{RO}^{VRS}(x_2, y_2) = 0.40$. The vertex of $\mathcal{T}^{VRS}(x_2, y_2)$ with the maximum β to α ratio corresponds to the point $(0.64, 1.69)$. Thus, the mpss is $1.69/0.64 = 2.64$ and $\mathcal{RI}^{CRS}(x_2, y_2) = \mathcal{RO}^{CRS}(x_2, y_2) = 0.64/1.69 = 0.38$. With respect to the HR technology, $\mathcal{RI}^{HR}(x_2, y_2) = 0.44$ and $\mathcal{RO}^{HR}(x_2, y_2) = 0.71$.

10.6 Extensions

It is possible to construct other useful two-dimensional projections.

1. *Partial productivity.* Fix an $(x, y) \in \mathcal{T}^{VRS}$. For each input i and output j, define the set

$$\{(\alpha,\beta):\ ((x_1,\ldots,\alpha x_i,\ldots,x_n),(y_1,\ldots,\beta y_j,\ldots,y_m))\in\mathcal{T}^{VRS}\}.$$

2. *Rate of technical substitution.* Fix an $(x,y)\in\mathcal{T}^{VRS}$. For each input i and input j, $i<j$, and define the set

$$\{(\alpha,\beta):\ ((x_1,\ldots,\alpha x_i,\ldots,\beta x_j,\ldots,x_n),\ y))\in\mathcal{T}^{VRS}\}.$$

3. *Rate of output substitution.* Fix an $(x,y)\in\mathcal{T}^{VRS}$. For each output i and output j, $i<j$, and define the set

$$\{(\alpha,\beta):\ (x,\ (y_1,\ldots,\alpha y_i,\ldots,\beta y_j,\ldots,y_m))\in\mathcal{T}^{VRS}\}.$$

10.7 Pivoting Algorithm

In this section, we describe a pivoting algorithm to compute a set of extreme points $E\subset\mathcal{T}^{VRS}(x_0,y_0)$ such that

$$\mathcal{T}^{VRS}(x_0,y_0)=\mathcal{FDH}(Conv(E)).$$

In lieu of using general notation, we shall describe the algorithm for the example data shown in Figure 10.1 and reference point $(x_0,y_0)=(x_2,y_2)$. It will be useful to be familiar with the terminology and results presented in Appendix C.5.

Here is an overview. We know from our previous calculations that the vertices for the two-dimensional section are shown in Figure 10.2. Presented in northeasterly order, they are $(0.44,0)$, $(0.44,1.10)$, $(0.52,1.37)$, $(0.64,1.69)$, $(0.80,2.10)$ and $(1.25,3.00)$. The algorithm proceeds in two phases, I and II, respectively.

Phase I of the algorithm determines α_m, the minimum value of α, equal to 0.44 in the example. As explained in Appendix C.5, it is sufficient to examine only the vertices of the convex polyhedron $\hat{C}$, that is, the α value associated with one of the vertices of C will be the minimum value of α we seek. There are too many vertices to examine one-by-one. The *simplex algorithm* of linear programming generates "adjacent" vertices whose corresponding α values continue to decrease until the optimum is found. Moving from one vertex to an adjacent vertex is known as *executing a pivot operation.*

At the end of Phase I, a vertex of $\hat{C}$ whose projection onto the two-dimensional section is $(0.44,0)$ is found. Phase II then executes a sequence of pivot operations that will generate a sequence of vertices of $\hat{C}$ whose respective projections onto the two-dimensional projection are $(0.44,1.10)$, $(0.52,1.37)$, $(0.64,1.69)$, $(0.80,2.10)$ and $(1.25,3.00)$.

Before we describe Phases I and II, we begin by describing the relationship between a vertex of $\hat{C}$ and the so-called *simplex tableau*, and then describe how one executes a pivot operation.

10.7.1 Vertices and the Simplex Tableau

The equations that define the unprojected two-dimensional section $\hat{C}$ in (10.11) for the example data may be expressed in the canonical polyhedral set form $Ax = b$ given by

$$\begin{bmatrix} 5.0 & 0.0 & -4.0 & -5.0 & -0.5 & -3.0 & -1.0 & -4.0 & -1.0 & 0.0 & 0.0 \\ 4.0 & 0.0 & -1.0 & -4.0 & -2.5 & -2.0 & -6.0 & -5.0 & 0.0 & -1.0 & 0.0 \\ 0.0 & 10.0 & -10.0 & -10.0 & -12.0 & -14.0 & -20.0 & -30.0 & 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 0.0 & 0.0 & 0.0 \end{bmatrix} \begin{pmatrix} \alpha \\ \beta \\ \lambda \\ r_1 \\ r_2 \\ q_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \tag{10.12}$$

where for notational convenience we let $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6)^T$. The initial basis is set to $\mathcal{B}_1 = \{\alpha, r_2, q_1, \lambda_2\}$. The basis matrix B_1 corresponding to this basis is[4]

$$B_1 \;=\; \begin{bmatrix} 5.0 & 0.0 & 0.0 & -5.0 \\ 4.0 & -1.0 & 0.0 & -4.0 \\ 0.0 & 0.0 & 1.0 & -10.0 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}, \tag{10.13}$$

whose inverse is

$$B_1^{-1} \;=\; \begin{bmatrix} 0.2 & 0.0 & 0.0 & 1.0 \\ 0.8 & -1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 10.0 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}. \tag{10.14}$$

By multiplying both sides of $Ax = b$ by B_1^{-1}, the new canonical polyhedral set form is

$$(B_1^{-1}A)x \;=\; B_1^{-1}b,$$

which is represented in the first four rows of Table 10.2. The symbol "RHS" stands for the right-hand side of the equality constraints and represents $B_1^{-1}b$. In lieu of writing the variable names next to the numbers (as is usually written in equation form), the variable names are given in the column headings. This representation is standard (for those familiar with the simplex algorithm in linear programming) and each set of four rows in Table 10.2 is called a **tableau**.

Remark 10.19. The four columns associated with the basis, namely, α, λ_2, r_2 and q_1, are marked in bold. These four columns, when *permuted* to match the *order* of the basis $\mathcal{B}_1$, namely, α, r_2, q_1 and λ_2, correspond to the 4×4 identity matrix. This is not surprising, as $B_1^{-1}B_1 = I$.

Each tableau identifies a vertex of the convex polyhedron $\hat{C}$. In the standard form (see Definition C.25, p. 467), all variables are constrained to be

[4] A basis matrix is an ordered subset of the columns of the A matrix corresponding to the ordered list of the variables in the basis.

Table 10.2. Tableaux associated with Phase I.

Basis	α	β	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	r_1	r_2	q_1	RHS
α	**1.00**	0.00	0.20	**0.00**	0.90	0.40	0.80	0.20	-0.20	**0.00**	**0.00**	1.00
r_2	**0.00**	0.00	-2.20	**0.00**	2.10	-0.40	5.20	1.80	-0.80	**1.00**	**0.00**	0.00
q_1	**0.00**	10.00	0.00	**0.00**	-2.00	-4.00	-10.00	-20.00	0.00	**0.00**	**1.00**	10.00
λ_2	**0.00**	0.00	1.00	**1.00**	1.00	1.00	1.00	1.00	0.00	**0.00**	**0.00**	1.00
α	1.00	0.00	1.14	0.00	0.00	0.57	-1.43	-0.57	0.14	-0.43	0.00	1.00
λ_3	0.00	0.00	-1.05	0.00	1.00	-0.19	2.48	0.86	-0.38	0.48	0.00	0.00
q_1	0.00	10.00	-2.09	0.00	0.00	-4.38	-5.05	-18.29	-0.76	0.95	1.00	10.00
λ_2	0.00	10.00	2.05	1.00	0.00	1.19	-1.48	0.14	0.38	-0.48	0.00	1.00
α	1.00	0.00	0.00	-0.56	0.00	-0.09	-0.60	-0.65	-0.07	-0.16	0.00	0.44
λ_3	0.00	0.00	0.00	0.51	1.00	0.42	1.72	0.93	-0.19	0.23	0.00	0.51
q_1	0.00	10.00	0.00	1.02	0.00	-3.16	-6.56	-18.14	-0.37	0.47	1.00	11.03
λ_1	0.00	0.00	1.00	0.49	0.00	0.58	-0.72	0.07	0.19	-0.23	0.00	0.49

non-negative. Only the basic variables may have positive values. *Non-basic* variables are all variables not in the basis. Each non-basic variable has value equal to zero. The values of the variables associated with the vertex in the first tableau in Table 10.2 are $(\alpha, r_2, q_1, \lambda_2) = (1.00, 0.00, 10.00, 1.00)$ and $(\beta, \lambda_1, \lambda_3, \lambda_4, \lambda_5, \lambda_6, r_1) = (0, 0, 0, 0, 0, 0, 0)$.

10.7.2 Pivot Operation

Basis $\mathcal{B}_1$ and $\mathcal{B}_2$ are *adjacent* if their respective list of variables differ in exactly one place. For example, the bases in the first and second tableaus in Table 10.2 are $\mathcal{B}_1 = (\alpha, r_2, q_1, \lambda_2)$ and $\mathcal{B}_2 = (\alpha, \lambda_3, q_1, \lambda_2)$, respectively. They are adjacent since their entries differ only in the second element. Moving from a vertex to an adjacent vertex is executing a pivot operation. In the parlance of the simplex algorithm, to obtain $\mathcal{B}_2$ from $\mathcal{B}_1$ one "pivots in" the variable λ_3 and "pivots out" the variable r_2.

Executing a pivot operation is best illustrated by showing how to generate the 2nd tableau in Table 10.2 from the first tableau. The basis matrix B_2 corresponding to this basis (constructed from the columns in the first tableau associated with the same order as the basis) is

$$B_2 = \begin{bmatrix} 1.00 & 0.90 & 0.00 & 0.00 \\ 0.00 & 2.10 & 0.00 & 0.00 \\ 0.00 & -2.00 & 1.00 & 0.00 \\ 0.00 & 1.00 & 0.00 & 1.00 \end{bmatrix}, \tag{10.15}$$

whose inverse is

$$B_2^{-1} = \begin{bmatrix} 1 & -\frac{0.90}{2.10} & 0 & 0 \\ 0 & \frac{1}{2.10} & 0 & 0 \\ 0 & -\frac{2.00}{2.10} & 0 & 0 \\ 0 & -\frac{1.00}{2.10} & 0 & 1 \end{bmatrix}. \tag{10.16}$$

The second tableau (viewed as a matrix) is obtained by multiplying the first tableau by B_2^{-1}.

Remark 10.20. Generating the second tableau from the first tableau can also be obtained by applying a sequence of familiar row operations to a system of linear equations. In effect, to obtain the second tableau we seek to transform the column associated with the variable λ_3, as follows:

$$\begin{pmatrix} 0.90 \\ 2.10 \\ -2.00 \\ 1.00 \end{pmatrix} \Longrightarrow \begin{pmatrix} 0.00 \\ 1.00 \\ 0.00 \\ 0.00 \end{pmatrix}. \tag{10.17}$$

To obtain a zero in the first element in the column multiply the second row by $-0.90/2.10$ and add it to the first row. Similarly, to obtain zeroes in the third and fourth elements in the column (i) multiply the second row by $2.00/2.10$ and add it to the third row and (ii) multiply the second row by $-1.00/2.10$ and add it to the fourth row. To obtain a one in the second row, divide the second row by 2.10.

The third basis $(\alpha, \lambda_3, q_1, \lambda_1)$ is obtained by pivoting in the variable λ_1 for the variable λ_2. Accordingly, the basis matrix B_3 corresponding to this basis (constructed from the columns in the second tableau associated with the same order as the basis) is

$$B_3 = \begin{bmatrix} 1.00 & 0.00 & 0.00 & 1.14 \\ 0.00 & 1.00 & 0.00 & -1.05 \\ 0.00 & 0.00 & 1.00 & -2.09 \\ 0.00 & 0.00 & 0.00 & 2.05 \end{bmatrix}, \tag{10.18}$$

whose inverse is

$$B_3^{-1} = \begin{bmatrix} 1 & 0 & 0 & -\frac{1.14}{2.05} \\ 0 & 1 & 0 & \frac{1.05}{2.05} \\ 0 & 0 & 1 & \frac{2.09}{2.05} \\ 0 & 0 & 0 & \frac{1}{2.05} \end{bmatrix}. \tag{10.19}$$

The third tableau (viewed as a matrix) is obtained by multiplying the second tableau by B_3^{-1}.

The *pivot element* is the number in each tableau inside the square. It is associated with the column of the variable entering the basis and the row of the variable that is leaving the basis. How does one select the pivot element?

Fix the variable that has been selected to enter the basis. We must decide which variable will leave the basis. In the standard form of each tableau, the RHS values must always be *non-negative*. Given this requirement, the interpretation (provided in Remark 10.20) of the pivot operation as a sequence of row operations shows that the pivot element must always be *positive*. In the column associated with λ_3 in the first tableau there are two possible pivot elements, namely, those associated with variables r_2 and λ_2, respectively. If one had chosen the pivot element associated with λ_2, the resulting value of the right-hand side for r_2 would be -1.00, which is not allowed. In order to ensure that all right-hand side values are non-negative, one performs the **ratio test** to select the pivot element: for each possible pivot element (in the column of the variable entering the basis) take the ratio of the right-hand side value to the pivot element and select the pivot element that has the *smallest* value. Here, one would compute the ratios 0.00/2.10 and 1.00/1.00 and therefore the ratio test would require that the first one be chosen as the pivot element. (Ties can be broken arbitrarily.)

10.7.3 Phase I

The purpose of Phase I of the algorithm is to compute α_m, the minimum value of α (see Definition 10.3). This problem is a linear programming problem. As explained in the overview at the beginning of this section, the algorithm executes a sequence of pivot operations.

We now describe how to select the variable that will enter the basis. (Once this has been determined the ratio test will determine which variable must leave the basis.) We know the pivot element must have a positive value. If we select a column whose entry in the α row is *positive*, the resulting value for the variable α *cannot increase*. If all of the entries in the RHS column are positive, the resulting α value will *decrease*. For example, as a result of executing the pivot to obtain the second tableau in Table 10.2, the new RHS value for α is $1.00 + (-0.90/2.10)(0.00) = 1.00$. The value for α did not decrease because the RHS value was zero. Now consider the pivot to obtain the third tableau: the new RHS value for α is

$$1.00 + (-1.14/2.05)(1.00) = 0.44,$$

which is lower than before because the RHS value was positive. We conclude that if the goal is to decrease the value of α, then *only those columns whose entry in the α row is positive should be considered*.[5] In principle, any valid column will work; in practice, one selects the column with the largest (absolute) value, which generates the steepest (local) change. Finally, if there are no more columns with positive entries, then a vertex has been found that achieves the lowest α value.

[5] On the other hand, if the goal is to decrease the value of α, then only those columns whose entry in the α row is negative should be considered.

To summarize, here are the steps associated with Phase I to find the minimum α value:

Step 1. Select the first basis and set up the first tableau as previously described in the first subsection.

Step 2. Select the next pivot element.

(a) Pick the column (variable to enter the basis) that has the largest value in the entry associated with the α row (here the first row). Break ties arbitrarily. If there is no such column, Phase I is complete.
(b) Perform the ratio test as described above. Consider only those ratios associated with *positive* entries. If there is no such entry then the problem would be unbounded, which may happen in general but not for the problems we consider here. Select the row (variable to leave the basis) that achieves the minimum ratio.

Step 3. Execute the pivot. Form the basis matrix associated with the basis variables as illustrated above. Make sure to maintain the order. Multiply the tableau (viewed as a matrix) by the inverse of the basis matrix. Alternatively, execute the row operations as described in Remark 10.20.

Step 4. Repeat Steps 2 and 3.

At the end of Phase I the first vertex of the two-dimensional section has been found. In our example, it is the point $(0.44, 0) = (\alpha_m, 0)$. The value for β will always be zero as this variable never enters the basis in Phase I.

10.7.4 Phase II

Phase II generates a sequence of vertices that can be used to generate the various two-dimensional sections. Please refer to Table 10.3, which shows the sequence of tableaus so generated. (The first tableau in this table is identical to the last tableau of Phase I, except that the second and third rows have been interchanged so that the α and β rows will appear consecutively.)

The first step in Phase II is to obtain the point (α_m, β_m), which equals $(0.44, 1.10)$ in the example (see Definition 10.4). This is achieved by automatically selecting the variable β to enter the basis. After executing the pivot the second tableau in Table 10.3 is generated.

The subsequent steps of Phase II are identical to Phase I with one critical exception, namely, the choice of which variable to enter the basis. This is because the objective now is *not* to minimize the value of α. Let (α_1, β_1) denote the current vertex and let (α_2, β_2) denote a candidate adjacent vertex. Intuitively, the goal should be to find the adjacent vertex (α_2, β_2) that will

Table 10.3. Tableaux associated with Phase II.

Basis	α	β	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	r_1	r_2	q_1	RHS
α	1.00	0.00	0.00	-0.56	0.00	-0.09	-0.60	-0.65	-0.07	-0.16	0.00	0.44
q_1	0.00	10.00	0.00	1.02	0.00	-3.16	-6.56	-18.14	-0.37	0.47	1.00	11.03
λ_3	0.00	0.00	0.00	0.51	1.00	0.42	1.72	0.93	-0.19	0.23	0.00	0.51
λ_1	0.00	0.00	1.00	0.49	0.00	0.58	-0.72	0.07	0.19	-0.23	0.00	0.49
α	1.00	0.00	0.00	-0.56	0.00	-0.09	-0.60	-0.65	-0.07	-0.16	0.00	**0.44**
β	0.00	1.00	0.00	0.10	0.00	-0.31	-0.66	-1.81	-0.04	0.05	0.10	**1.10**
λ_3	0.00	0.00	0.00	0.51	1.00	0.42	1.72	0.93	-0.19	0.23	0.00	0.51
λ_1	0.00	0.00	1.00	0.49	0.00	0.58	-0.72	0.07	0.19	-0.23	0.00	0.49
α	1.00	0.00	0.16	-0.48	0.00	0.00	-0.72	-0.64	-0.04	-0.02	0.00	**0.52**
β	0.00	1.00	0.54	0.37	0.00	0.00	-1.05	-1.78	0.06	-0.08	0.10	**1.37**
λ_3	0.00	0.00	-0.72	0.16	1.00	0.00	2.24	0.88	-0.32	0.40	0.00	0.16
λ_4	0.00	0.00	1.72	0.84	0.00	1.00	-1.24	0.12	0.32	-0.40	0.00	0.84
α	1.00	0.00	-0.36	-0.36	0.73	0.00	0.91	0.00	-0.27	0.09	0.00	**0.64**
β	0.00	1.00	-0.91	0.69	2.02	0.00	3.47	0.00	-0.58	0.73	0.10	**1.69**
λ_6	0.00	0.00	-0.82	0.18	1.14	0.00	2.55	1.00	-0.36	0.45	0.00	0.18
λ_4	0.00	0.00	1.82	0.82	-0.14	1.00	-1.55	0.00	0.36	-0.45	0.00	0.82
α	1.00	0.00	0.00	-0.20	0.70	0.20	0.60	0.00	-0.20	0.00	0.00	**0.80**
β	0.00	1.00	0.00	1.10	1.95	0.50	2.70	0.00	-0.40	0.50	0.10	**2.10**
λ_6	0.00	0.00	0.00	0.55	1.08	0.45	1.85	1.00	-0.20	0.25	0.00	0.55
λ_1	0.00	0.00	1.00	0.45	-0.08	0.55	-0.85	0.00	0.20	-0.25	0.00	0.45
α	1.00	0.00	1.00	0.25	0.62	0.75	-0.25	0.00	0.00	-0.25	0.00	**1.25**
β	0.00	1.00	2.00	2.00	1.80	1.60	1.00	0.00	0.00	0.00	0.20	**3.00**
λ_6	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	1.00
r_1	0.00	0.00	5.00	2.25	-0.38	2.75	-4.25	0.00	1.00	-1.25	0.00	2.25

maximize the *slope* $(\beta_2 - \beta_1)/(\alpha_2 - \alpha_1)$ of the line segment joining these adjacent vertices. It remains to figure out how to know which column (variable) to select to achieve this objective.

The coordinates of the vertex associated with the second tableau in Table 10.3 is $(0.44, 1.10)$. The new vertex in the third tableau is $(0.52, 1.37)$, which is obtained by executing the pivot associated with the pivot element 0.58 shown in the tableau. Notice that the new α value is

$$0.52 = 0.44 + \Big(\frac{0.09}{0.58}\Big)0.49, \tag{10.20}$$

and the new β value is

$$1.37 = 1.10 + \Big(\frac{0.31}{0.58}\Big)0.49. \tag{10.21}$$

The slope of the line segment joining these adjacent vertices is

$$\frac{\Delta\beta}{\Delta\alpha} := \frac{1.37 - 1.10}{0.52 - 0.44} = \frac{(\frac{0.31}{0.58})0.49}{(\frac{0.09}{0.58})0.49} = \frac{0.31}{0.09}\,. \tag{10.22}$$

We make two observations. First, in general, the slope of the line segment joining adjacent vertices is the ratio of the entries in the β row to the α row, i.e., $(-0.31)/(-0.09)$. Second, as Figure 10.2 clearly indicates, the sequence of vertices generated must move in the northeasterly direction, that is, the new α and β values *must increase*. For this to happen, the variables that will both enter and leave the basis must each have a *negative* value in the element corresponding to the α and β rows, respectively—examine equations (10.20) and (10.21) in conjunction with Remark 10.20. Since we seek to maximize the slope (10.22), the variable that will enter the basis is the one whose ratio of the (negative) β entry to its (negative) α entry is the largest. (Ties can be broken arbitrarily.) In the second tableau, only the variables λ_2, λ_4, λ_5, λ_6, r_1 and r_2 can be considered. The variables λ_2 and r_2 can be eliminated as their respective β entries are not negative. Of the remaining variables, the ratios 0.31/0.09, 0.66/0.60, 1.81/0.65, 0.04/0.07 are computed and the ratio 0.31/0.09 is, indeed, the highest. Consequently, the variable λ_4 will enter the basis. The next step is to perform the ratio test to determine which variable leaves the basis, which is λ_1 in the example. A pivot is executed and the process continues. It stops at the last tableau since there is no adjacent vertex of $(1.25, 3.00)$ in the northeasterly direction. (In fact, at this point an extreme direction has been found.)

Remark 10.21. Phase II may generate vertices of $\hat{C}$ that do *not* project into vertices of the two-dimensional section. It is possible that the slopes of the line segments joining a sequence of vertices are equal; thus, the vertices in the interior of the sequence will project onto the interior of the line segment joining the outermost vertices of this sequence. From the geometric perspective, when this happens all such vertices of $\hat{C}$ belong to a *face* of $\hat{C}$ whose projection maps onto a line segment in the two-dimensional section.

10.8 Exercises

10.1. The vertices of a two-dimensional section $\mathcal{T}^{VRS}(X_0, Y_0)$ are (0.25, 0), (0.25, 0.50), (0.50, 4) and (2, 10).

(a) Graphically depict $\mathcal{T}^{VRS}(X_0, Y_0)$ and $\mathcal{T}^{CRS}(X_0, Y_0)$ (on the same graph).
(b) Determine the VRS input efficiency.
(c) Determine the VRS output efficiency.
(d) Determine the hyperbolic efficiency $\mathcal{H}^{VRS}(X_0, Y_0)$.
(e) Determine the CRS input efficiency.

(f) Determine the CRS output efficiency.
(g) Determine the hyperbolic efficiency $\mathcal{H}^{CRS}(X_0, Y_0)$.

10.2. The vertices of a two-dimensional section $\mathcal{T}^{VRS}(X_0, Y_0)$ are (0.10, 0.20), (0.25, 0.75) and (1, 1).

(a) What is the most productive scale size?
(b) Characterize the returns-to-scale for the reference firm.

10.3. The input-output data for this exercise is given in Table 10.4. (It is the same data used for Exercise 9.1.)

Table 10.4. Input-output data for Exercise 10.3.

Firm	Capital	Labor	Output
1	2	6	50
2	4	2	50
3	5	9	150
4	8	4	200
5	8	8	50

(a) Determine the two-dimensional section $\mathcal{T}^{HR}(x_5, y_5)$ using the graphical approach described in Section 10.5.
(b) Use your answer to (a) to determine the input, output and hyperbolic efficiencies for Firm 5 assuming the HR model of technology.
(c) Determine the two-dimensional section $\mathcal{T}^{VRS}(x_5, y_5)$ using the graphical approach described in Section 10.5.
(d) Use your answer to (c) to determine the input, output and hyperbolic efficiencies for Firm 5 assuming the VRS model of technology.
(e) Use your answer to (c) to determine the input, output and hyperbolic efficiencies for Firm 5 assuming the CRS model of technology.
(f) Use your answer to (c) to determine the $mpss$ associated with Firm 5.

10.4. For the data given in Table 10.4 determine the two-dimensional section $\mathcal{T}^{VRS}(x_5, y_5)$ using the pivoting algorithm described in Section 10.7.

10.9 Bibliographical Notes

Banker [1984] discusses the two-dimensional projection as a device to communicate the concepts of efficiency. The algorithm for computing the two-dimensional projection is due to Hackman et. al. [1994]. Rosen et. al. [1998] show how to apply the two-dimensional projection to examine marginal rates of substitution among inputs. A pivoting algorithm that computes a general Efficient Frontier can be found in Hackman and Passy [2002].

10.10 Solutions to Exercises

10.1 (a) The set $\mathcal{T}^{VRS}(X_0, Y_0)$ is

$$\Big\{(\alpha, \beta) \geq 0 : \ \alpha \geq 0.25, \ \beta \leq 14\alpha - 3, \ \beta \leq 4\alpha + 2, \ \beta \leq 10\Big\}.$$

The set $\mathcal{T}^{CRS}(X_0, Y_0)$ is the smallest convex cone in $\mathbb{R}^2_+$ containing $\mathcal{T}^{VRS}(X_0, Y_0)$, and is

$$\Big\{(\alpha, \beta) \geq 0 : \ \beta \leq 8\alpha\Big\}.$$

(b) The line $\beta = 14\alpha - 3$ passing through points $(0.25, 0.50)$ and $(0.5, 4)$ intersects the line $\beta = 1$ at the point $(2/7, 1)$, and so the input efficiency for the VRS model of technology is $2/7$.

(c) The line $\alpha = 1$ intersects the line $\beta = 4\alpha + 2$ passing through points $(0.50, 4)$ and $(2, 10)$ at the point $(1, 6)$, and so the output efficiency for the VRS model of technology is $1/6$.

(d) The curve $\beta = 1/\alpha$ intersects $\mathcal{T}^{VRS}(X_0, Y_0)$ at a point lying on the line segment joining points $(0.25, 0.50)$ and $(0.5, 2)$. The point of intersection must therefore satisfy $\beta = 1/\alpha$ and $\beta = 14\alpha - 3$. The α value of the point of intersection is the positive root of the quadratic equation $14\alpha^2 - 3\alpha - 1$. The positive root is $\alpha = (3 + \sqrt{9 + 4(14)})/2(14) = 0.3951$ and so $\beta = 2.5311$. Consequently, $\mathcal{H}^{VRS}(X_0, Y_0)$ is 0.3951.

(e) The line $\beta = 8\alpha$ passing through origin and $(0.5, 4)$ intersects the line $\beta = 1$ at the point $(1/8, 1)$, and so the input efficiency for the CRS model of technology is $1/8$.

(f) The line $\alpha = 1$ intersects the line $\beta = 8\alpha$ passing through the origin and $(0.50, 4)$ at the point $(1, 8)$, and so the output efficiency for the CRS model of technology is $1/8$. (We also directly know that the input and output efficiencies for the CRS model are identical.)

(g) The curve $\beta = 1/\alpha$ intersects $\mathcal{T}^{CRS}(X_0, Y_0)$ at a point lying on the line segment joining the origin and $(0.50, 4)$. The point of intersection must therefore satisfy $\beta = 1/\alpha$ and $\beta = 8\alpha$. The α value of the point of intersection is the positive root of the quadratic equation $8\alpha^2 - 1$. The positive root is $\alpha = \sqrt{1/8} = 0.3536$ and so $\beta = 2.8281$. Consequently, $\mathcal{H}^{CRS}(X_0, Y_0)$ is 0.3536. (We also directly know that the hyperbolic efficiency in this setting is the square root of the input/output efficiency.)

10.2 (a) The *mpss* is the maximum of the following ratios $0.20/0.10, 0.75/0.25$ and $1/1$, which is 3.

(b) The vertex associated with the *mpss* is $(0.25, 0.75)$. Since $0.25 < 1$, the returns-to-scale for the reference firm is decreasing.

10.3 Table 10.5 shows the final calculations used to answer these questions. Here is how they are obtained. The point $x_1^+ = (6, 6)$ has output 50 in both the HR and VRS models. The point $x_2^+ = (4, 4)$ has output 50 in both

the HR and VRS models. The point $x_3^+ = (9, 9)$ has output 150 in both the HR and VRS models. The point $x_4^+ = (8, 8)$ has output 200 in both the HR and VRS models. The line $L = -2K + 10$ joining points x_1 and x_2 intersects the line $L = K$ (that passes through the origin and x_5) at $x_{12} = (1/3)x_1 + (2/3)x_2 = (10/3, 10/3)$, and this input vector has output of 50 in both the HR and VRS models. The line $L = -(1/3)K + 20/3$ joining points x_1 and x_4 intersects the line $L = K$ (that passes through the origin and x_5) at $x_{14} = (1/2)x_1 + (1/2)x_2 = (5, 5)$, and this input vector has output of 50 in the HR model but $(1/2)(50) + (1/2)(200) = 125$ in the VRS model. The line $L = 7K - 26$ joining points x_2 and x_3 intersects the line $L = K$ (that passes through the origin and x_5) at $x_{23} = (2/3)x_2 + (1/3)x_3 = (13/3, 13/3)$, and this input vector has output of 50 in the HR model but $(2/3)(50) + (1/3)(150) = 250/3$ in the VRS model. The line $L = -(5/3)K + 52/3$ joining points x_3 and x_4 intersects the line $L = K$ (that passes through the origin and x_5) at $x_{34} = (1/2)x_3 + (1/2)x_4 = (6.5, 6.5)$, and this input vector has output of 150 in the HR model but $(1/2)(150) + (1/2)(200) = 175$ in the VRS model.

Table 10.5. Computing the two-dimensional projection for Exercise 10.3.

	λ_1	λ_2	λ_3	λ_4	λ_5	α	β_{VRS}	β_{HR}
x_{12}	2/3	1/3				$0.41\bar{6}$	1.0	1.0
x_{14}	1/2			1/2		0.6250	2.5	1.0
x_{23}		2/3	1/3			$0.54\bar{1}$	$1.\bar{6}$	1.0
x_{34}			1/2	1/2		0.8125	3.5	3.0
x_1^+	1.0					0.7500	1.0	1.0
x_2^+		1.0				0.5000	1.0	1.0
x_3^+			1.0			1.1250	3.0	3.0
x_4^+				1.0		1.0000	4.0	4.0
x_5					1.0	1.0000	1.0	1.0

(a) The vertices that define the staircase shape of $\mathcal{T}^{HR}(x_5, y_5)$ are $(0.41\bar{6}, 0)$, $(0.41\bar{6}, 1)$, $(0.8125, 1)$, $(0.8125, 3)$, $(1, 3)$ and $(1, 4)$.

(b) It is obvious that input efficiency is $0.41\bar{6}$ and output efficiency is 0.25. As for the hyperbolic efficiency the curve $\beta = 1/\alpha$ intersects $\mathcal{T}^{HR}(x_5, y_5)$ along the line segment joining points $(0.8125, 1)$ and $(0.8125, 3)$. The point of intersection is obviously $(0.8125, 1.2308)$ and the hyperbolic efficiency for the HR model is therefore 0.8125.

(c) The vertices that characterize $\mathcal{T}^{VRS}(x_5, y_5)$ are $(0.41\bar{6}, 0)$, $x_{12} = (0.416\bar{6}, 1)$, $x_{14} = (0.625, 4)$, $x_{34} = (0.8125, 3.5)$, and $x_4 = (8, 4)$. Therefore,

$$\mathcal{T}^{VRS}(x_5, y_5) = \{(\alpha, \beta) \geq 0 : \ \alpha \geq 0.416\bar{6},\ \beta \leq 7.2\alpha - 2,$$
$$\beta \leq 5.\bar{3}K - 0.8\bar{3},\ \beta \leq 2.\bar{6} + 1.\bar{3},\ \beta \leq 4\Big\}.$$

(d) The line $\beta = 1$ intersects $\mathcal{T}^{VRS}(x_5, y_5)$ at the point $(0.41\bar{6}, 1)$ and so the input efficiency for the VRS model is $0.41\bar{6}$. The line $\alpha = 1$ intersects $\mathcal{T}^{VRS}(x_5, y_5)$ at the point $(1, 4)$ and so the output efficiency for the VRS model is 0.25. As for the hyperbolic efficiency, the curve $\beta = 1/\alpha$ intersects the line $\beta = 7.2\alpha - 2$. The α value of the point of intersection is the positive root of the quadratic equation $7.2\alpha^2 - 2\alpha - 1$. The positive root is $\alpha = (2 + \sqrt{4 + 4(7.2)})/2(7.2) = 0.5366$, and so $\beta = 1.8636$. The hyperbolic efficiency for the VRS model is therefore 0.5366.

Remark 10.22. The point of intersection is

$$(\hat{\alpha}, \hat{\beta}) := (0.5366, 1.8636) = 0.42429x_{12} + 0.57571x_{14}.$$

Since $x_{12} = (1/3)x_1 + (2/3)x_2$ and $x_{14} = (1/2)x_1 + (1/2)x_4$, it follows that the actual point $(\hat{x}, \hat{y}) \in \mathcal{T}^{VRS}$ used to determine the hyperbolic efficiency measure for the VRS model is

$$0.42929(x_1, y_1) + 0.28286(x_2, y_2) + 0.28786(x_4, y_4),$$

which equals $((4.29285, 4.29285), 93.1782)$. By construction, we have $4.29285/8 = 50/93.1782 = 0.5366$.

(e) The set $\mathcal{T}^{CRS}(x_5, y_5)$ is the smallest convex cone in $\mathbb{R}^2_+$ containing $\mathcal{T}^{VRS}(x_5, y_5)$. It is given by

$$\Big\{(\alpha, \beta) \geq 0 : \ \beta \leq (3.5/0.8125)\alpha = (56/13)\alpha = 4.3077\alpha\Big\}.$$

The line $\beta = 1$ intersects $\mathcal{T}^{CRS}(x_5, y_5)$ at the point $(13/56, 1)$ and so the input efficiency for the VRS model is 0.2321. The line $\alpha = 1$ intersects $\mathcal{T}^{CRS}(x_5, y_5)$ at the point $(1, 56/13)$ and so the output efficiency for the VRS model is 0.2321, too—we knew this directly. As for the hyperbolic efficiency, the curve $\beta = 1/\alpha$ intersects the line $\beta = (56/13)\alpha$. The α value of the point of intersection is the positive root of the quadratic equation $(56/13)\alpha^2 - 1$. The positive root is $\alpha = \sqrt{13/56} = 0.4818$ and so $\beta = 2.0755$. The hyperbolic efficiency for the VRS model is therefore 0.4818.

Remark 10.23. The point of intersection is

$$(\hat{\alpha}, \hat{\beta}) := (0.4818, 2.0755) = 0.5930x_{34}.$$

Since $x_{34} = (1/2)x_3 + (1/2)x_4$, it follows that the actual point $(\hat{x}, \hat{y}) \in \mathcal{T}^{VRS}$ used to determine the hyperbolic efficiency measure for the $CVRS$ model is

$$0.2965(x_3, y_3) + 0.2965(x_4, y_4) = ((3.8545, 3.8545), 103.7750).$$

By construction, we have $3.8545/8 = 50/103.7750 = 0.4818$.

(f) The *mpss* is the slope 56/13.

Remark 10.24. Notice how almost trivial it is to determine the various efficiency measures once the vertices of the two-dimensional section are determined.

10.4 The equations that define the unprojected two-dimensional section $\hat{C}$ in (10.11) for the problem data can be expressed in the canonical polyhedral set form $Ax = b$ given by

$$\begin{bmatrix} 8.0 & 0.0 & -2.0 & -4.0 & -5.0 & -8.0 & -8.0 & -1.0 & 0.0 & 0.0 \\ 8.0 & 0.0 & -6.0 & -2.0 & -9.0 & -4.0 & -8.0 & 0.0 & -1.0 & 0.0 \\ 0.0 & 50.0 & -50.0 & -50.0 & -150.0 & -200.0 & -50.0 & 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 0.0 & 0.0 & 0.0 \end{bmatrix} \begin{pmatrix} \alpha \\ \beta \\ \lambda \\ r_1 \\ r_2 \\ q_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \tag{10.23}$$

where for notational convenience we let $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)^T$. The initial basis is set to $\mathcal{B}_1 = \{\alpha, r_2, q_1, \lambda_2\}$. The sequence of pivots associated with Phase I are shown in Table 10.6.

Table 10.6. Tableaux associated with Phase I.

Basis	α	β	λ_1	λ_2	λ_3	λ_4	λ_5	r_1	r_2	q_1	RHS
α	1.00	0.00	0.75	0.50	0.38	0.00	0.00	-0.13	0.00	0.00	1.00
r_2	0.00	0.00	4.00	-2.00	4.00	-4.00	0.00	-1.00	1.00	0.00	0.00
q_1	0.00	50.00	0.00	0.00	-100.00	-150.00	0.00	0.00	0.00	1.00	50.00
λ_5	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	1.00
α	1.00	0.00	0.00	0.88	-0.38	0.75	0.00	0.06	-0.19	0.00	1.00
λ_1	0.00	0.00	1.00	-0.50	1.00	-1.00	0.00	-0.25	0.25	0.00	0.00
q_1	0.00	50.00	0.00	0.00	-100.00	-150.00	0.00	0.00	0.00	1.00	50.00
λ_5	0.00	0.00	0.00	1.50	0.00	2.00	1.00	0.25	-0.25	0.00	1.00
α	1.00	0.00	0.00	0.00	-0.38	-0.42	-0.58	-0.08	-0.04	0.00	0.42
λ_1	0.00	0.00	1.00	0.00	1.00	-0.33	0.33	-0.17	0.17	0.00	0.33
q_1	0.00	50.00	0.00	0.00	-100.00	-150.00	0.00	0.00	0.00	1.00	50.00
λ_2	0.00	0.00	0.00	1.00	0.00	1.33	0.67	0.17	-0.17	0.00	0.67

The sequence of pivots associated with Phase II are shown in Table 10.7. The first tableau in this table is identical to the last tableau of Phase I, except that the second and third rows have been interchanged so that the α and β rows will appear consecutively. All of the information obtained graphically to determine the two-dimensional projection is contained in this table. The bases identify which input vectors are used to generate the vertices. For example, the vertex $(0.625, 2.5)$ is a 50-50 mixture of the points x_1 and x_4. The ratio of the β

to α rows identify the slopes of the line segments joining adjacent vertices. For example, the pivot element marked in the second tableau corresponds to the β/α ratio of $(-3.00)/(-0.41\bar{6}) = 7.2$, which is the slope of the line segment joining vertices $(0.41\bar{6}, 1)$ and $(0.625, 2.5)$. Also, the pivot element marked in the third tableau corresponds to the β/α ratio of $(-2.00)/(-0.0375) = 5.\bar{3}$, which is the slope of the line segment joining vertices $(0.625, 2.5)$ and $(0.8125, 3.5)$.

Table 10.7. Tableaux associated with Phase II.

Basis	α	β	λ_1	λ_2	λ_3	λ_4	λ_5	r_1	r_2	q_1	RHS
α	1.00	0.00	0.00	0.00	-0.38	-0.42	-0.58	-0.08	-0.04	0.00	0.42
q_1	0.00	50.00	0.00	0.00	-100.00	-150.00	0.00	0.00	0.00	1.00	50.00
λ_1	0.00	0.00	1.00	0.00	1.00	-0.33	0.33	-0.17	0.17	0.00	0.33
λ_2	0.00	0.00	0.00	1.00	0.00	1.33	0.67	0.17	-0.17	0.00	0.67
α	1.00	0.00	0.00	0.00	-0.38	-0.42	-0.58	-0.08	-0.04	0.00	**0.42**
β	0.00	1.00	0.00	0.00	-2.00	-3.00	0.00	0.00	0.00	0.02	**1.00**
λ_1	0.00	0.00	1.00	0.00	1.00	-0.33	0.33	-0.17	0.17	0.00	0.33
λ_2	0.00	0.00	0.00	1.00	0.00	1.33	0.67	0.17	-0.17	0.00	0.67
α	1.00	0.00	0.00	0.31	-0.38	0.00	-0.38	-0.03	-0.09	0.00	**0.63**
β	0.00	1.00	0.00	2.25	-2.00	0.00	1.50	0.38	-0.38	0.02	**2.50**
λ_1	0.00	0.00	1.00	0.25	1.00	0.00	0.50	-0.13	0.13	0.00	0.50
λ_4	0.00	0.00	0.00	0.75	0.00	1.00	0.50	0.13	-0.13	0.00	0.50
α	1.00	0.00	0.38	0.41	0.00	0.00	-0.19	-0.08	-0.05	0.00	**0.81**
β	0.00	1.00	2.00	2.75	0.00	0.00	2.50	0.13	-0.13	0.02	**3.50**
λ_3	0.00	0.00	1.00	0.25	1.00	0.00	0.50	-0.13	0.13	0.00	0.50
λ_4	0.00	0.00	0.00	0.75	0.00	1.00	0.50	0.13	-0.13	0.00	0.50
α	1.00	0.00	0.75	0.50	0.38	0.00	0.00	-0.13	0.00	0.00	**1.00**
β	0.00	1.00	3.00	3.00	1.00	0.00	3.00	0.00	0.00	0.02	**4.00**
r_2	0.00	0.00	8.00	2.00	8.00	0.00	4.00	-1.00	1.00	0.00	4.00
λ_4	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	1.00

11

Multi-Stage Efficiency Analysis

In this chapter, we explore systems that consist of several stages arranged in series. Succeeding stages (or subsystems) use a mixture of exogenous inputs and intermediate outputs of preceding stages. We explore how to (i) assess the efficiency of each stage within the aggregate system, and (ii) analyze possible tradeoffs of the subsystem efficiencies.

As a starting point, one can treat each subsystem as a system in its own right. In this manner, the technology set for each subsystem is constructed using the relevant input-output data from its own peers, and the technology set of the aggregate system is constructed using aggregated inputs and outputs without regard to intermediate input-output factors that link the various stages. As we subsequently demonstrate, adopting this approach makes it possible for the aggregate system to be rated very inefficient while each subsystem is rated efficient, and for the aggregate system to be rated near efficient while each subsystem is rated highly inefficient.

We describe an expansion of the ordinary technology sets to develop a corresponding efficiency measurement framework that *simultaneously* computes the efficiency of each subsystem and the aggregate system. The measurement framework has the following properties:

- If each subsystem is rated efficient, then so must the aggregate system.
- Each subsystem's efficiency and the aggregate efficiency cannot exceed the efficiency obtained using the classical DEA approach, and can be expected to be far lower.
- The methodology described herein provides the recipe on how to obtain the operational improvements at *all* levels of the hierarchy, as it explicitly integrates the computation of the various efficiencies.

The approach described in this chapter expands the technology sets of each subsystem by allowing each to acquire resources from the other in exchange for delivery of the appropriate (intermediate or final) product, and to form composites from both subsystems. Managers of each subsystem will not agree to vertical integration initiatives unless each subsystem will be more

efficient than what each can achieve by separately applying conventional efficiency analysis. A Pareto efficient frontier characterizes the acceptable set of efficiencies of each subsystem from which the managers will negotiate to select the final outcome.

Three proposals for the choice for the Pareto efficient point are discussed: the one that achieves the largest equiproportionate reduction in the classical efficiencies, the one that achieves the largest equal reduction in efficiency, and the one that maximizes the radial contraction in the aggregate consumption of resources originally employed before integration. We show how each choice for the Pareto efficient point determines a derived measure of aggregate efficiency. We introduce a "consistent pricing principle" and show it characterizes the proposed models. An extensive numerical example is used to illustrate exactly how the subsystems can significantly improve their operational efficiencies via integration beyond what would be predicted by conventional analysis.

11.1 A Representative Multi-Stage System

We limit our application to two stages in series, and we make the following simplifying assumptions:

- *Technology.* Each subsystem 1 uses capital and labor to produce an intermediate product used by subsystem 2 to produce final product. Subsystem 2 also requires capital and labor, which are assumed to be completely transferable resources between stages. All technologies described herein exhibit constant returns-to-scale.

- *Market.* Each subsystem 2 has a unique supplier given by its subsystem 1. There is a market for the intermediate product, and the transfer price subsystem 1 charges subsystem 2 is the prevailing market price. Competitive markets exist so that either subsystem is a *price-taker* in the input and output markets. That is, it may expand or contract its output without affecting its price or cost of inputs.

- *Organization.* Each subsystem is viewed as a profit center, and each manager is given decision-making authority. Although we do not explicitly model the incentive scheme for the managers, we assume each manager is highly motivated to improve his own system's efficiency. Efficiency can be thought of as the proxy for performance, which is why each manager will not consent to an acquisition of resources unless he directly benefits from it. From an organizational perspective, the modeling approach described herein is viewed as a natural starting point for efficiency improvement.

To ease notational burdens and to make concrete the conceptual discussions to follow, we shall analyze multi-stage systems such as the one depicted in Figure 11.1. Each DMU_j (j = 1,...,N) consists of two subsystems in series.

Subsystem $1j$ (hereafter abbreviated S_{1j}) uses capital K_{1j} and labor L_{1j} to produce intermediate product I_j. Subsystem $2j$ (hereafter abbreviated by S_{2j}) uses capital K_{2j} and labor L_{2j} together with I_j to produce final output F_j. Constant returns-to-scale (CRS) will be assumed throughout. The models we develop will be illustrated with a 10 DMU numerical example. The data on inputs and outputs for S_{1j}, (K_{1j}, L_{1j}, I_j), and S_{2j}, $(K_{2j}, L_{2j}, I_j, F_j)$ for the example are given in Table 1.

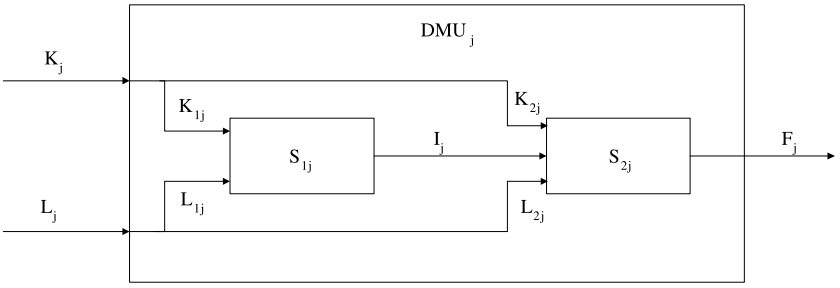


Fig. 11.1. Aggregate DMU with two stages in tandem.

Table 11.1. Data for the numerical example.

DMU	K_1	K_2	K	L_1	L_2	L	I	F
1	32	81	113	67	83	150	46.928	64.941
2	96	28	124	40	81	121	47.431	46.492
3	79	51	130	89	79	168	79.694	67.388
4	41	80	121	35	26	61	32.978	31.124
5	99	8	107	33	74	107	45.921	35.018
6	72	29	101	15	36	51	24.861	29.146
7	21	88	109	64	23	87	32.250	34.049
8	60	39	99	71	49	120	64.659	45.176
9	7	86	93	80	16	96	21.531	21.062
10	10	40	50	33	11	44	12.519	10.189

11.2 Description of Multi-Stage Technology

11.2.1 Classical Models of Technology

For a given data set of input-output pairs $(X_j, Y_j) \in \mathbb{R}^n_+ \times \mathbb{R}^m_+$, $j = 1, 2, \ldots, N$, the classical CRS technology $\mathcal{T}$ is

$$\mathcal{T} := \{(X,Y) : \sum_j \lambda_j X_j \leq X,\ \sum_j \lambda_j Y_j \geq Y\},$$

where we now and hereafter suppress the nonnegativity constraints imposed on the intensity variables (the λ_j's). Given the technology set $\mathcal{T}$ and $(X_0, Y_0) \in \mathcal{T}$, the classical (CL) radial measure of input efficiency is

$$\theta_0^{CL} := \text{Min}\ \{\theta_0 :\ (\theta_0 X_0, Y_0) \in \mathcal{T}\}.$$

For the multi-stage systems depicted in Figure 11.1, the classical descriptions of the technology sets for each subsystem are

$$\mathcal{T}_1 := \left\{((K,L),\ I) :\ \textstyle\sum_j \lambda_{1j} K_{1j} \leq K,\ \sum_j \lambda_{1j} L_{1j} \leq L,\ \sum_j \lambda_{1j} I_j \geq I\right\},$$
$$\mathcal{T}_2 := \left\{((K,L,I),\ F) :\ \textstyle\sum_j \lambda_{2j} K_{2j} \leq K,\ \sum_j \lambda_{2j} L_{2j} \leq L,\ \sum_j \lambda_{2j} I_j \leq I,\ \sum_j \lambda_{2j} F_j \geq F\right\}.$$

For each Decision-Making Unit (DMU_0) in the data set, the classical measures of input efficiency for each stage are

$$\theta_{10}^{CL} := \text{Min}\{\theta_{10} :\ ((\theta_{10}K_{10}, \theta_{10}L_{10}),\ I_0) \in \mathcal{T}_1\},$$
$$\theta_{20}^{CL} := \text{Min}\{\theta_{20} :\ ((\theta_{20}K_{20}, \theta_{20}L_{20}, \theta_2 I_0),\ F_0) \in \mathcal{T}_2\}.$$

Computational results are reported in Table 2.

Table 11.2. Classical efficiency evaluation for S_{1j} and S_{j2}.

	S_{1j}		S_{2j}	
DMU_j	θ_{1j}^{CL}	Benchmarks	θ_{2j}^{CL}	Benchmarks
1	1.00	1	1.00	1
2	0.943	5,8	1.00	2
3	0.973	5,8	1.00	3
4	0.958	5,8	0.904	3,6
5	1.00	5	1.00	5
6	1.00	6	1.00	6
7	0.941	1,9	1.00	7
8	1.00	8	1.00	8
9	1.00	9	0.918	1,7
10	0.748	1,9	0.735	1,7

11.2.2 Expanded Model of Technology

A numerical example using one of the DMUs in the data set will be used to explain how to expand the technology to provide better opportunities for each subsystem to improve its efficiency. (The model used to generate this example is formally described in the next section.) In what follows, we make the following assumptions:

1. Each stage is managed as a profit center.
2. S_{1j} may sell its intermediate product on the open market for the same price it charges S_{2j}.
3. Both S_{1j} and S_{2j} may sell any amount of their respective outputs on the open market without affecting input cost or price.

The observed S_{25} of DMU_5 uses 8 units of capital, 74 units of labor and 45.92 units of intermediate product to produce 35.02 units of final product. The manager of S_{25} (hereafter named 'M_{25}'), while always looking to improve efficiency, is content for now as his system is rated efficient by classical efficiency analysis. Now suppose the manager of S_{15} (hereafter named 'M_{15}') comes to M_{25} with the following proposal: "I can show you how to increase your output by 6.3%, while *simultaneously* reducing your cost of inputs by 11.75%. Interested?" That is, M_{15} is proposing a way for M_{25} to use $(K, L, I) = (7.06, 65.29, 40.52)$ to produce $F = 37.23$ instead of M_{25}'s current production plan that uses $(K, L, I) = (8, 74, 45.92)$ to produce $F = 35.02$. To expand his output by 6.3%, M_{25} would normally expect (under CRS) to have to increase his inputs by 6.3%, and so, in effect, M_{15} is offering M_{25} to consume only 100(1 - 0.1175)/1.063 = 83% of his input to achieve the same output level. While M_{25} is obviously intrigued by M_{15}'s proposal, M_{25} demands an explanation as to how M_{15} proposes to accomplish this seemingly impossible task, as M_{25} knows that *both* S_{15} and S_{25} were rated input efficient by classical analysis. M_{15} obliges with the following explanation.

Using classical descriptions of technology for each subsystem, M_{15} found a composite subsystem 1 process that uses $(34.48, 40.80)$ units of capital and labor to produce 37.16 units of intermediate product, and a composite subsystem 2 process that uses $(37.04, 45.98, 31.76)$ units of capital, labor and intermediate product to produce the 37.23 units of final product, which M_{15} promised to deliver to M_{25}. The total amounts of capital and labor required by these two composite processes are 71.52 and 86.78, respectively. With the 7.06 units of capital and 65.29 units of labor acquired from M_{25}, M_{15} still needs 64.46 units of capital and 21.49 units of labor, which he possesses as these totals represent only 65.1% of his current capacity of $(99, 33)$ units of capital and labor. With respect to the intermediate product, M_{25} notes that while he is now purchasing 45.92-40.52 = 5.40 *less* units of intermediate product from M_{15}, the difference between what the subsystem 2 composite requires and what the subsystem 1 composite currently produces of intermediate product is also 5.40 = 37.16 - 31.76 units, which M_{15} will sell on the open market to compensate him for the loss in revenue from M_{25}. M_{25} is satisfied that M_{15}'s proposal is conceptually sound.

M_{25} now understands why M_{15} is so eager to offer this proposal to M_{25}: under the proposal, M_{15} will be able to free up 34.9% of *his* inputs, a considerable savings, while still producing his same level of output. Since M_{15} cannot achieve this savings without M_{25}'s consent, M_{25} realizes he must understand

exactly how M_{15} was able to devise this seemingly ingenious plan, so that he will be in position to negotiate with M_{15} a better deal for himself.

11.2.3 Expanded Subsystem Technology Sets

For every DMU$_j$, the manager of S_{1j} now realizes that the classical efficiency analysis constructed the efficient frontier using *only* subsystem 1 processes. It does not consider the possibility that M_{1j} may have the options of adopting an alternative subsystem 2 production process *and* acquiring resources from M_{2j} (as long as M_{2j} would agree). With these options, the technology set for S_{1j}, which defines the collection of input pairs (K, L) that can produce at least I_j, has been expanded.

Under the CRS assumption, M_{10} knows that $\omega_{20}((K_{20}, L_{20}, I_0),\ F_0) \in \mathcal{T}_2$ for all $\omega_{20} \geq 0$. In order to entice M_{20} to agree, M_{10} selects a value $\theta_{20} < 1$, and offers M_{20} the opportunity to achieve the input-output point of $\omega_{20}((\theta_{20}K_{20}, \theta_{20}L_{20}, \theta_{20}I_0),\ F_0)$. In order for M_{10} to meet his obligation to M_{20} and his objective, namely to produce I_0 with resources (K, L), he must find two composite processes $((\hat{K}_1, \hat{L}_1),\ \hat{I}_1) \in \mathcal{T}_1$ and $((\hat{K}_2, \hat{L}_2, \hat{I}_2),\ \hat{F}) \in \mathcal{T}_2$ for which the following four *inventory balance equations* must hold:

[E1] *Capital.* The supply of capital from M_{10} and M_{20}, $K + \omega_{20}(\theta_{20}K_{20})$, must be no smaller than the demand for capital by both composite subsystems, $\hat{K}_1 + \hat{K}_2$.

[E2] *Labor.* The supply of labor from M_{10} and M_{20}, $L + \omega_{20}(\theta_{20}L_{20})$, must be no smaller than the demand for labor by both composite subsystems, $\hat{L}_1 + \hat{L}_2$.

[E3] *Intermediate Product.* The supply of intermediate product from M_{20} and the composite Stage 1 process, $\hat{I}_1 + \omega_{20}(\theta_{20}I_0)$, must be no smaller than the demand for intermediate product by M_{10} and the composite Stage 2 process, $I_0 + \hat{I}_2$.

[E4] *Final Product.* The supply of final product from the composite Stage 2 process, $\hat{F}$, must be no smaller than the demand for final product by M_{20}, $\omega_{20}F_0$.

Let $\mathcal{T}_1^E(\theta_{20})$ denote the collection of input-output pairs $((K, L),\ I_0)$ that satisfy the inventory balance equations [E1]–[E4] listed above for DMU$_0$. Given θ_{20}, it would make sense for M_{10} to find the least amount of capital and labor to satisfy his own output requirement of I_0. Accordingly, he should solve the following linear programming model, which we shall denote as the *Acquisition* (AQ) model:

$$\theta_{10}^{AQ}(\theta_{20}) := \min \theta_{10} \tag{11.1}$$

$$\sum_j \lambda_{1j}K_{1j} \ + \ \sum_j \lambda_{2j}K_{2j} \leq \theta_{10}K_{10} \ + \ \omega_{20}[\theta_{20}K_{20}],$$

$$\sum_j \lambda_{1j} L_{1j} \; + \; \sum_j \lambda_{2j} L_{2j} \leq \theta_{10} L_{10} \; + \; \omega_{20}[\theta_{20} L_{20}],$$
$$\sum \lambda_{1j} I_j \; + \; \omega_{20}[\theta_{20} I_0] \geq I_0 \; + \; \sum \lambda_{2j} I_j,$$
$$\sum \lambda_{2j} F_j \geq \omega_{20} F_0.$$

In the proposal of M_{15} to M_{25} that was described in Section 2.3, M_{15} selected $\theta_{25} = 0.83$, and solved the AQ model, whose solution was $\omega_{25} = 1.063$ with $\theta_{15}^{AQ}(\theta_{25}) = 0.651$.

Now that M_{20} understands how M_{10} was able to achieve *his* objective, M_{20} realizes he can play the same game. Let $\mathcal{T}_2^E(\theta_{10})$ denote the collection of input-output pairs $((K, L, I), \; F_0)$ that satisfy analogous four inventory balance requirements as described above. Given θ_{10} it would make sense for M_{20} to find the least amount of capital and labor to satisfy his own output requirement of F_0. Accordingly, he would solve his own *Acquisition* (AQ) model, namely, the following linear programming model:

$$\theta_{20}^{AQ}(\theta_{10}) := \min \theta_{20} \tag{11.2}$$

$$\sum_j \lambda_{1j} K_{1j} \; + \; \sum_j \lambda_{2j} K_{2j} \leq \theta_{20} K_{20} \; + \; \omega_{10}[\theta_{10} K_{10}],$$
$$\sum_j \lambda_{1j} L_{1j} \; + \; \sum_j \lambda_{2j} L_{2j} \leq \theta_{20} L_{20} \; + \; \omega_{10}[\theta_{10} L_{10}],$$
$$\sum \lambda_{1j} I_j \; + \; \theta_{20} I_0 \geq \sum \lambda_{2j} I_j \; + \; \omega_{10} I_0,$$
$$\sum \lambda_{2j} F_j \geq F_0.$$

For example, suppose M_{25} selects $\theta_{15} = 0.9$. Solution of his AQ model gives $\omega_{15} = 0.686$ and $\theta_{25}^{AQ}(\theta_{15}) = 0.697$. Note how much better off M_{25} is and worse off M_{15} is as compared to M_{15}'s original proposal. Both managers will agree that either proposal will outperform the classical analysis.

We emphasize the following point about describing the subsystem technologies. Since we allow the possibility of one subsystem manager to acquire resources from the other, as long as they can agree, the potential acquisition of resources consistent with the "θ_{10}—θ_{20} agreement" must now be embedded in the respective descriptions of technology given by $\mathcal{T}_1^E(\theta_{20})$ and $\mathcal{T}_2^E(\theta_{10})$ to reflect the set of all production possibilities.

11.3 Pareto efficient Frontiers

For the serial system we discuss here, a gain by one manager is a loss by the other manager. Regardless of the final choice for how the two subsystems shall vertically integrate, the agreed-upon choice for θ_{10} and θ_{20} should minimally

result in a Pareto efficient outcome; that is, $(\theta_{10}, \theta_{20}) = (\theta_{10}^{AQ}(\theta_{20}), \theta_{20}^{AQ}(\theta_{10}))$. Otherwise, neither manager, M_{10} nor M_{20}, would agree to the vertical integration.

The efficient frontier corresponding to DMU_5 in our example is depicted in Figure 11.2.[1] When θ_{15} is set to 1.00, θ_{25} is at its lowest value 0.66. On the other hand, when θ_{25} is set to 1.00, θ_{15} is assigned its lowest value 0.43.

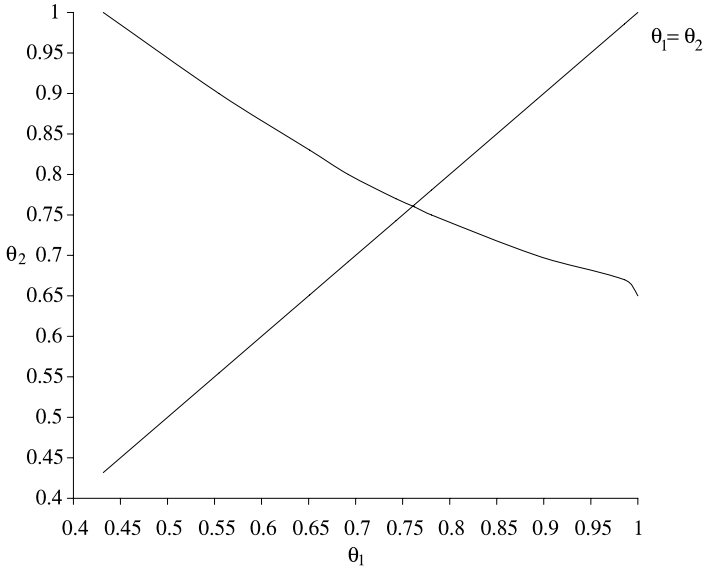


Fig. 11.2. Efficient frontier for DMU_5, θ_{15} *vs.* θ_{25}.

Two remarks concerning the Acquisition Models (11.1) and (11.2) that determine the Pareto efficient frontier are in order. First, it is not necessary to solve both Acquisition Models, as there is a one-to-one correspondence between the solutions for each Acquisition Model—the solution to Model (11.2) can be obtained from the solution to Model (11.1) by dividing λ_{1j}^* and λ_{2j}^* by ω_{20}^*, and setting $\omega_{10}^* = (\omega_{20}^*)^{-1}$. Second, the solutions to either Model (11.1) or Model (11.2) must necessarily lie below their respective classical efficiency counterparts, since the linear program to compute θ_{10}^{CL} is a special case of Model (11.1) in which $\omega_{20} = 0$ and $\lambda_{2j} = 0$ and the linear program to compute θ_{20}^{CL} is a special case of Model (11.2) in which $\omega_{10} = 0$ and $\lambda_{1j} = 0$. From an economic perspective, M_{20} would never agree to a proposal from M_{10} if the proposed θ_{20} exceeds what he could achieve on his own, and M_{10} would never offer a proposal to M_{20} in which he receives an efficiency θ_{10} that exceeds what he could achieve on his own, too.

[1] This frontier was constructed using the pivoting algorithm described in Hackman and Passy (2002).

In principle, any point on the Pareto efficient frontier is a candidate. There are three natural choices:

1. Select the point that achieves the largest *equiproportionate* reduction in the respective classical single-stage efficiencies θ_{10}^{CL} and θ_{20}^{CL}. For example, the 45^o line in Figure 11.2 intersects the frontier at the equiproportional point $\theta_{15} = \theta_{25} = 0.762$, a point which might be considered as fair for both subsystems.
2. Select the point that achieves the largest equal reduction in efficiency. (When $\theta_{10}^{CL} = \theta_{20}^{CL} = 1$, as is the case for DMU_5, these two choices will obviously coincide.)
3. Select the point that achieves for the vertically integrated unit the largest radial contraction in the aggregate amounts of capital and labor originally employed, which we more fully discuss in the next section.

11.4 Aggregate Efficiency

11.4.1 Measures of Aggregate Input Efficiency

From the perspective of an aggregate DMU, the classical model of technology is

$$\mathcal{T}_A := \left\{((K, L),\ F) :\ \sum_j \lambda_j(K_{1j} + K_{2j}) \le K,\ \sum_j \lambda_j(L_{1j} + L_{2j}) \le L,\ \sum_j \lambda_j F_j \ge F\right\}. \tag{11.3}$$

For the aggregate DMU_0 (denoted hereafter as $A0$), the classical model ignores the intermediate product I_0 as it represents internal production. The corresponding classical input efficiency measure is

$$\theta_{A0}^{CL} := \text{Min}\{\theta_{A0} :\ ((\theta_{A0}(K_{10} + K_{20}), \theta_{A0}(L_{10} + L_{20})),\ F_0) \in \mathcal{T}_A\}.$$

Färe and Grosskopf [1996] provide an in-depth development of models of technology for general multi-stage systems. One of their basic models (Färe and Grosskopf [1996, pp. 20-23]) allows *complete transferability* (CT) of capital and labor flows between the stages. Applied to the two-stage systems we analyze in this chapter, their model is formulated as

$$\begin{aligned}\mathcal{T}_A^{CT} := \{((K, L),\ F) :\ &\sum_j \lambda_{1j}K_{1j} + \sum_j \lambda_{2j}K_{2j} \le K,\\ &\sum_j \lambda_{1j}L_{1j} + \sum_j \lambda_{2j}L_{2j} \le L,\\ &\sum_j \lambda_{1j}I_j - \sum_j \lambda_{2j}I_j \ge 0,\\ &\sum_j \lambda_{2j}F_j \ge F\}.\end{aligned}$$

The third constraint above represents *inventory balance* of intermediate product to ensure that the supply of I produced by the composite S_1 will be sufficient to satisfy the demand for I by the composite S_2. Assuming complete transferability of resources, the Färe-Grosskopf measure of input efficiency would be computed as

$$\theta_{A0}^{CT} := \text{Min}\{\theta_{A0} : \ ((\theta_{A0}(K_{10}+K_{20}), \theta_{A0}(L_{10}+L_{20})), \ F_0) \in \mathcal{T}_A^{CT}\}.$$

11.4.2 Derived Measure of Aggregate Efficiency

For each Pareto efficient point $(\theta_{10}, \theta_{20})$, let $K(\theta_{10}, \theta_{20})$ and $L(\theta_{10}, \theta_{20})$ denote, respectively, the aggregate amounts of capital and labor which the vertically integrated unit would use to produce *both* F_0 and I_0. A natural choice for a derived measure of aggregate input efficiency is

$$\theta_{A0}^{D}(\theta_{10}, \theta_{20}) \ := \ \text{Max}\left\{\frac{K(\theta_{10}, \theta_{20})}{K_{10}+K_{20}}, \ \frac{L(\theta_{10}, \theta_{20})}{L_{10}+L_{20}}\right\}.$$

We now show how to compute $K(\theta_{10}, \theta_{20})$ using both Acquisition Models (11.1) and (11.2). (The derivation for $L(\theta_{10}, \theta_{20})$ is analogous.) First suppose that $\omega_{20} \leq 1$. Examine the right-hand side of the first constraint in (11.1). In return for delivering $\omega_{20}F_0$ units of final product to S_{20} and meeting its own requirements of producing I_0, S_{10} uses $\omega_{20}[\theta_{20}K_{20}]$ units of capital it acquires from S_{20} and $\theta_{10}K_{10}$ for its own production needs. For the vertically integrated unit to produce a total of F_0, S_{20} will have to produce the remaining amount $(1-\omega_{20})F_0$ by its own production process and thereby consume $(1-\omega_{20})K_{20}$ units of capital. In this case,

$$K(\theta_{10}, \theta_{20}) = \theta_{10}K_{10} + \omega_{20}[\theta_{20}K_{20}] + (1-\omega_{20})K_{20}.$$

Now suppose $\omega_{20} \geq 1$. Since $\omega_{10} = \omega_{20}^{-1} \leq 1$, we shall examine the right-hand side of the first constraint in (11.2). Here, in return for delivering $\omega_{10}I_0$ units of intermediate output to S_{10}, S_{20} uses $\omega_{10}[\theta_{10}K_{10}]$ of capital it acquires from S_{10}, and $\theta_{20}K_{20}$ units for its own production needs. For S_{10} to produce a total of I_0, it will need $(1-\omega_{10})K_{10}$ units of capital to produce the remaining amount $(1-\omega_{10})I_0$ using its current production process. In this case,

$$K(\theta_{10}, \theta_{20}) = \omega_{10}[\theta_{10}K_{10}] + \theta_{20}K_{20} + (1-\omega_{10})K_{10}.$$

To summarize, we have

$$K(\theta_{10}, \theta_{20}) = \begin{cases} \theta_{10}K_{10} \ + \ \omega_{20}[\theta_{20}K_{20}] + (1-\omega_{20})K_{20}, & \omega_{20} \leq 1, \\ \omega_{10}[\theta_{10}K_{10}] \ + \ \theta_{20}K_{20} + (1-\omega_{10})K_{10}, & \omega_{10} \leq 1, \end{cases}$$

$$L(\theta_{10}, \theta_{20}) = \begin{cases} \theta_{10}L_{10} \ + \ \omega_{20}[\theta_{20}L_{20}] + (1-\omega_{20})L_{20}, & \omega_{20} \leq 1, \\ \omega_{10}[\theta_{10}L_{10}] \ + \ \theta_{20}L_{20} + (1-\omega_{10})L_{10}, & \omega_{10} \leq 1. \end{cases}$$

A numerical example will help explain our proposed derived measure of aggregate efficiency. We solved Model (11.1) for DMU_3 with $\theta_{23} = 0.9$. The result is $\theta_{13} = 0.9667$ and $\omega_{23} = 0.1439$. The two composite subsystems constructed by the linear program are, respectively,

$$((\hat{K}_1, \hat{L}_1),\ \hat{I}_1) = ((70.88, 83.87),\ 76.38)$$
$$((\hat{K}_2, \hat{L}_2), \hat{I}_2),\ \hat{F}_2) = ((12.09, 12.39, 7.01),\ 9.696).$$

Now consider the four inventory balance equations [E1]–[E4] associated with this Pareto efficient point ($\theta_{13} = 0.9667, \theta_{23} = 0.9$):

$$\begin{aligned} 70.88 + 12.09 &\leq (0.9667)[79] + 0.1439[0.90 \cdot 51] = 82.97, \\ 83.87 + 12.39 &\leq (0.9667)[89] + 0.1439[0.90 \cdot 79] = 96.26, \\ 76.38 + 0.1439[0.90 \cdot 79.694] &\geq [79.694] + 7.01, \\ 9.696 &\geq 0.1439(67.39). \end{aligned}$$

Note how S_{13} is only promising to deliver 14.39% of final output; the remaining 85.61% must be produced by S_{23} using its own production process. The derived capital in this case is

$$(0.9667)[79] + 0.1439[0.90 \cdot 51] + (1 - 0.1439) \cdot 51 = 126.63,$$

and the derived labor is

$$(0.9667)[89] + 0.1439[0.90 \cdot 79] + (1 - 0.1439) \cdot 79 = 163.89.$$

When $(126.63, 163.89)$ is compared to the original values of $(130, 168)$, we obtain $\theta^D_{A3} = 0.9755$.

The derived aggregate measure of efficiency is measured along the Pareto efficient frontier corresponding to Models (11.1) and (11.2). It can never be larger than 1.0. Conceptually, any point on the Pareto efficient frontier could be used to define the aggregate efficiency. As discussed at the end of the last section there are two obvious choices: the equiproportional solution, $(\theta_{10} = \rho\theta^{CL}_{10}, \theta_{20} = \rho\theta^{CL}_{20})$, where $\rho \leq 1$, and the equal contraction solution in which $\theta_{10} = \theta_{20}$. We propose a third alternative: Minimize θ^D_{A0} on the Pareto efficient frontier, which we shall denote by θ^P_{A0}. To compute θ^P_{A0}, a bi-level programming problem, we iteratively solve Model (11.2) (resp. Model (11.1)) for different θ_{10} (resp. θ_{20}) values.

11.4.3 Computational Results

Table 3 reports the computational results for each measure of aggregate efficiency. First, we compare θ^{CT}_{Aj} to θ^{CL}_{Aj} for $j = 1, ..., 10$. In stark contrast to the relative efficiency nature of DEA, when additional flexibility of transferring resources between stages is available, the CT model is able to use this flexibility to identify potential improvement opportunities for *all* DMUs. Indeed, the

relevant benchmarks, which report the reference sets for each of the evaluated DMUs, contain *both* S_{1j} and S_{2j} stages (first and second rectangular brackets, respectively, in column 5 of Table 3). From a measurement perspective, it will always be the case that $\theta_{Aj}^{CT} \leq \theta_{Aj}^{CL}$.

Table 11.3. Aggregate efficiency measures.

	Classical Efficiency		Complete Transferability		Expanded Technology	
DMU_j	θ_{Aj}^{CL}	Benchmarks	θ_{Aj}^{CT}	Benchmarks	θ_{Aj}^{P}	Benchmarks
1	1.00	1	0.983	[1,8], [6]	0.988	[8], [6]
2	0.829	1,6	0.777	[5,8] [1]	0.842	[8] [1]
3	0.922	1,6	0.899	[1,8], [6]	0.953	[1], [6]
4	0.893	6	0.853	[5], [1,7]	0.922	[5], [1]
$\boxed{5}$	0.710	1,6	0.666	[5,8], [1]	0.781	[8], [6]
6	1.00	6	0.956	[5], [1,7]	0.981	[5], [7]
7	0.799	1,6	0.751	[5,8], [1]	0.867	[8], [6]
8	0.854	1,6	0.816	[8], [1,6]	0.868	[-], [6]
9	0.480	1,6	0.449	[5,8], [1]	0.621	[8], [6]
10	0.486	1,6	0.456	[5,8], [1]	0.712	[8], [6]

When comparing θ_{Aj}^{CT} to θ_{Aj}^{P} in Table 3, we see that $\theta_{Aj}^{P} > \theta_{Aj}^{CT}$ always holds. (We have been unable to establish any definitive relationship between θ_{Aj}^{CL} and θ_{Aj}^{P}.) Thus, for our numerical example, the CT model indeed finds the maximal possible contraction from the point of view of the aggregate DMU. From an organizational perspective, it may not be possible to implement this solution. We know it is impossible to achieve a better result than θ_{Aj}^{P} *along the Pareto efficient frontier*. Consequently, to implement the solution proposed by θ_{Aj}^{CT} when it is smaller than θ_{Aj}^{P} will require either M_{1j} or M_{2j} to consent to a restructuring that would make him *worse* off than he can achieve by negotiating directly with the manager of the other subsystem. When consent is required, it would make more sense for an aggregate manager to select a Pareto efficient point that both managers will accept. Table 4 records the Pareto efficient points for the two stages and their corresponding θ_{Aj}^{D} values. For every DMU_j, the minimal value for the derived aggregate efficiency (θ_{Aj}^{P}) is given in a box and the equiproportional choice is highlighted in boldface. Observe the wide disparity in efficiencies for each stage corresponding to the Pareto aggregate efficiency. Since the equiproportionate choice seems to sacrifice little in the way of aggregate efficiency, it is a practical alternative that is easier for the managers to agree on.

Table 11.4. Derived aggregate efficiency along the Pareto frontier of the two stages.

DMU	Measures	Efficiency Values								
1	θ_{11}	1.000	0.9999	**0.9865**	0.9703	0.9272				
	θ_{21}	0.9837	0.9837	**0.9865**	0.9900	1.000				
	θ^D_{A1}	0.991	0.990	0.989	0.988	0.990				
2	θ_{12}	0.9428	0.9409	0.9344	**0.8358**	0.8037	0.7638	0.7427	0.7014	0.5196
	θ_{22}	0.7970	0.7980	0.800	**0.8358**	0.8500	0.8600	0.8800	0.9000	1.000
	θ^D_{A2}	0.924	0.923	0.919	0.861	0.845	0.842	0.844	0.867	0.892
3	θ_{13}	0.9731	0.9718	0.9667	**0.9288**	0.8408	0.766	0.5266		
	θ_{23}	0.8753	0.8800	0.9000	**0.9288**	0.9500	0.9600	1.000		
	θ^D_{A3}	0.986	0.981	0.976	0.953	0.958	0.961	0.965		
4	θ_{14}	0.9580	0.9436	0.9292	**0.8936**	0.5912				
	θ_{24}	0.8600	0.8700	0.8800	**0.8936**	0.900				
	θ^D_{A4}	0.986	0.947	0.922	0.952	0.952				
5	θ_{15}	1.000	0.8924	**0.7608**	0.6922	0.6511	0.5547	0.4318		
	θ_{25}	0.6600	0.7000	**0.7608**	0.8000	0.8300	0.9000	1.000		
	θ^D_{A5}	1.000	0.908	0.819	0.781	0.781	0.843	0.892		
6	θ_{16}	1.000	0.9827	**0.9612**	0.9523	0.7469	0.6419			
	θ_{26}	0.9000	0.9300	**0.9612**	0.9750	0.9900	1.000			
	θ^D_{A6}	1.000	0.981	0.981	0.981	0.982	0.988			
7	θ_{17}	0.9416	0.8728	**0.7666**	0.7367	0.6962	0.5793			
	θ_{27}	0.4917	0.6250	**0.7666**	0.800	0.85	1.000			
	θ^D_{A7}	0.988	0.930	0.867	0.879	0.888	0.985			
8	θ_{18}	0.9976	0.9055	**0.8401**	0.6560	0.4822	0.1625			
	θ_{28}	0.8170	0.8300	**0.8401**	0.8740	0.9155	1.000			
	θ^D_{A8}	0.927	0.919	0.913	0.892	0.868	0.907			
9	θ_{19}	1.000	0.8086	0.6005	**0.5424**	0.4719	0.3966	0.1618		
	θ_{29}	0.3402	0.400	0.500	**0.5424**	0.600	0.700	0.9065		
	θ^D_{A9}	0.825	0.771	0.687	0.621	0.723	0.813	0.913		
10	$\theta_{1,10}$	0.7476	0.7000	**0.5589**	0.5226	0.4896	0.4386			
	$\theta_{2,10}$	0.3351	0.4000	**0.5589**	0.65	0.6670	0.6967			
	$\theta^D_{A,10}$	0.949	0.821	0.712	0.726	0.763	0.809			

11.5 A Consistent Pricing Principle

The dual linear fractional program to each manager's Acquisition Model, also known as the *multiplier* formulation, provides an alternative means to understand the tradeoff inherent in the Pareto efficient frontier for managers M_{10} and M_{20}. For M_{10}, the multiplier formulation is

$$\theta^{AQ}_{10} := \max \frac{\pi_I I_0}{\pi_K K_{10} \ + \ \pi_L L_{10}} \tag{11.4}$$

$$\frac{\pi_I I_j}{\pi_K K_{1j} + \pi_L L_{1j}} \leq 1, \quad j = 1, \ldots, n,$$

$$\frac{\pi_F F_j}{\pi_K K_{2j} + \pi_L L_{2j} + \pi_I I_j} \leq 1, \quad j = 1, \ldots, n,$$

$$\frac{\pi_F F_0}{\pi_K K_{20} + \pi_L L_{20} + \pi_I I_0} \geq \theta_{20},$$

and for M_{20}, the multiplier formulation is

$$\theta_{20}^{AQ} := \max \frac{\pi_F F_0}{\pi_K K_{20} \ + \ \pi_L L_{20} \ + \ \pi_I I_0} \tag{11.5}$$

$$\begin{aligned} \tfrac{\pi_I I_j}{\pi_K K_{1j} + \pi_L L_{1j}} &\leq 1, \quad j = 1, \ldots, n, \\ \tfrac{\pi_F F_j}{\pi_K K_{2j} + \pi_L L_{2j} + \pi_I I_j} &\leq 1, \quad j = 1, \ldots, n, \\ \frac{\pi_I I_0}{\pi_K K_{10} + \pi_L L_{10}} &\geq \theta_{10}. \end{aligned}$$

The last constraint in each model ensures a lower bound on the efficiency of the counterpart subsystem. As the lower bound parameter varies it changes the efficiency in the obvious way. For example, raising θ_{20} lowers S_{10}'s efficiency in (11.4), and raising θ_{10} lowers S_{20}'s efficiency in (11.5).

Observe that there is a single multiplier π_K for both K_1 and K_2, a single multiplier π_L for both L_1 and L_2, and a single multiplier π_I that is used to weigh the intermediate factor both when it is an output (of the first stage) and when it is an input (to the second stage). Since the capital, labor and intermediate product are freely transferable between stages, their respective weights in the multiplier formulation should be the same. We shall call this the *Consistent Pricing Principle*. Consistent pricing holds for the Färe-Grosskopf model as well. There, the linear fractional programming dual is

$$\theta_{20}^{CT} := \max \frac{\pi_F F_0}{\pi_K (K_{10} + K_{20}) \ + \ \pi_L (L_{10} + L_{20}) \ + \ \pi_I I_0}$$

$$\begin{aligned} \tfrac{\pi_I I_j}{\pi_K K_{1j} + \pi_L L_{1j}} &\leq 1, \quad j = 1, \ldots, n, \\ \tfrac{\pi_F F_j}{\pi_K K_{2j} + \pi_L L_{2j} + \pi_I I_j} &\leq 1, \quad j = 1, \ldots, n. \end{aligned}$$

In an ordinary application of DEA, M_{10} would prefer a larger value of the multiplier π_I, whereas M_{20} would prefer a smaller value of π_I. When both output-input ratios appear in the same optimization, necessarily there will be a tradeoff between the measurement of efficiency of both stages. The consistent pricing principle leads to a natural conflict between the efficiency measures of the two stages. A weighting scheme that might make M_{10} efficient might very well make M_{20} look inefficient, and vice-versa. Thus, there will be a need to coordinate the choice for this multiplier. Regardless of the weighting scheme ultimately agreed upon, it should not be possible to select an alternative set of weights that would make both stages at least as efficient while making one

of them more efficient. That is, it should minimally result in a Pareto efficient outcome; otherwise, neither manager M_{10} nor M_{20} would agree to the vertical integration.

11.6 Extensions

Technology: The basic model assumed constant returns-to-scale and computed standard radial measures of efficiency. It can be easily extended to the variable returns-to-scale models or, more generally, to other technologies used to generate measures of efficiency that eliminate the slacks in resource use.

Structure: The basic model contains just two stages. An obvious extension is to increase the number of serial stages in the model. Such an extension is possible as the notion of Pareto efficient frontier and the definition of Pareto aggregate efficiency easily generalize. The main (and significant) difficulty here is the time it will take to compute the various measures.

Choice of variables: The model can also be extended to allow more flexibility in the definition of the inputs and outputs. First, it is straightforward to incorporate additional input and output factors. Second, the model can be extended to allow the presence of inputs and outputs that are not completely transferable in some subsystems. For example, some inputs are specific to a particular subsystem and can not be shared. In this case, for the purpose of modeling and computing a subsystem's efficiency, it will be easier to work with the multiplier formulation using the consistent pricing principle.

Transaction costs: The basic model assumes no transaction costs when resources are moved between the two stages. These costs can be formulated by either adding some terms to the balance equation of the capital or by applying a certain "depreciation" term on each amount that is transferred.

11.7 Bibliographical Notes

This chapter is adapted from Golany et. al. [2006]. See references cited therein to recent attempts to model some form of multi-stage efficiency.

Färe and Grosskopf [1996] and [2000] develop a general multi-stage model with intermediate inputs-outputs, coined *network DEA*. In their framework, each internal stage's technology is modeled using a single-stage DEA model. The two-stage model of the *flow of material* described in this chapter is a special case of Färe and Grosskopf's multi-stage framework. However, the proposed aggregate efficiency measure is fundamentally different. In particular, in cases when the proposed aggregate efficiency is higher, it necessarily

follows that it will not be possible to disaggregate the Fare-Grosskopf aggregate efficiency measure into separate efficiency measures for which each submanager will consent. If there is an aggregate manager who may unilaterally reallocate resources without consent of the submanagers, then assessing aggregate efficiency using the Färe-Grosskopf framework may lead to superior results for the whole system. However, as we have noted, due to the linkage of inputs and outputs between the stages, in such a context one subsystem's efficiency can be vastly improved at the expense of potential improvement in the other subsystem, which may render meaningless the assessment of subsystem efficiency.

17

Index-Based Dynamic Production Functions

An economist typically works with aggregate data that record the cumulative amounts of inputs and outputs in some predetermined period of time (e.g., quarterly, yearly). With today's information systems, detailed shop-floor data are becoming increasingly available, which opens the door to a refined description of technology.

At a micro-level, the exact shape of the input curve must be known to project realized output rates over time. Within an activity or stage of production, this dynamic input-output process is conveniently encapsulated by a **dynamic production function**

$$x = (x_1(\cdot), x_2(\cdot), \ldots, x_n(\cdot)) \quad \xrightarrow{f} \quad y = (y_1(\cdot), y_2(\cdot), \ldots, y_m(\cdot)).$$

Each $x_i(t)$ represents the quantity of input i at time t, and each $y_j(t)$ represents the quantity of output j realized at time t. A dynamic production function $f(\cdot)$ is a *functional*, since both its domain and range are vectors of functions, not vectors of numbers. A dynamic production function defines a recipe with more flexible elements than a steady-state production function.

We begin by describing a motivating example of non-instantaneous behavior. Next, we define the class of functions used to model all flows of goods and services. The simplest description of dynamic production assumes instantaneous transformations. This assumption can be relaxed to incorporate constant lead times. Index-based dynamic production functions will be used to model these processes, and three practical ways of indexing to incorporate constant lead times will be described.

17.1 A Motivating Example

Consider a production system that takes two time periods to transform a unit of input into a unit of output. In the first three time periods, system input was observed to be 24, 48 and 96 units, respectively. What would be your

answer to the following question: "How much total output has emerged by the end of the first, second, and third time periods?"

The production system description is (purposely) ambiguous, but we begin with an "obvious" answer, namely, no units of output will be realized in each of the first two periods and 24 units of output will be realized by the end of the third period. Even this simple answer makes the implicit assumption that there was *no* input in the two periods *prior* to the first period; otherwise, these two input numbers should be included as output in the first two periods. For simplicity, we shall assume that there was no input prior to the first period.

Further investigation reveals that the system operates one 8-hour shift per day, say from 9:00am to 5:00pm, and the first three time periods correspond to Monday through Wednesday. The process involves a heating (light manufacturing) operation that requires a cooling period of exactly 16 hours before the semi-finished part is completed as a finished product. (Another example is a painting operation that requires parts to dry.) All semi-finished parts are stored in a room while cooling, and this room is available 24 hours a day. Since the production process occurs round-the-clock, we define a time period to correspond to a single 24-hour day, say 9:00am to 9:00am.

The length of each time period is 24 hours. Let $x_i(\tau)$, $\tau \in [0, 24]$, denote the input curve in each period (day) $i = 1, 2, 3$. The total or cumulative input in each period is

$$x_i := \int_0^{24} x_i(\tau) d\tau.$$

The **shape** of the input in each period i is

$$s_i(\tau) := x_i(\tau)/x_i, \quad \tau \in [0, 24).$$

It represents normalized input in that $\int_0^{24} s_i(\tau) d\tau = 1$ for each i. We know that $x_1 = 24$, $x_2 = 48$ and $x_3 = 96$ and that $s_i(\tau) = 0$ if $8 \leq t \leq 24$ for each i, but, as yet, no further information about the shapes of the input curves are known. Let $Y(t)$ denote the cumulative output obtained due to input in the time interval $[0, t]$. (Here, $t \in [0, \infty)$ represents a point in time.) From what we have learned so far,

$$Y(8) = 0,\ Y(24) = Y(32) = 24,\ Y(48) = Y(56) = 72,\ Y(72) = 168.$$

Suppose competitive pressures and transportation lead times dictate that shipping occurs 24 hours a day. In particular, suppose a sizeable percentage of the shipping activity occurs during the third shift, 1:00am - 9:00am, each day. (A one-day time period consisting of three 8-hr shifts is consistent with the original description.) It is useful now to view the overall production process as consisting of three *stages* in series:

$$\text{light manufacturing} \longrightarrow \text{cooling} \longrightarrow \text{shipping}.$$

Shipping cannot ship a semi-finished part and so it will be necessary to determine $Y(t)$ for $t \in [16, 24)$, i.e., between 1am–9am. However, it will not be

possible to obtain these values *unless* the actual shapes, $s_i(\tau)$, of each input curve are known. If each $s_i(\tau)$ is "front-loaded," so that most input occurred early in each shift, then

$$Y(16) \approx Y(24) = 24,\ Y(40) \approx Y(48) = 72,\ Y(56) \approx Y(72) = 168.$$

If, on the other hand, each $s_i(\tau)$ is "back-loaded," so that most input occurred later in each shift, then

$$Y(0) \approx Y(16) = 0,\ Y(24) \approx Y(40) = 24,\ Y(48) \approx Y(56) = 72.$$

The shape (or distribution) of input over time can have an enormous effect on how output emerges over time.

At an atomic level, the production process, although relatively short, is not instantaneous. The follow-on cooling operation takes 16 hours, which is too significant to ignore. To simplify matters, we have conveniently assumed it took *exactly* 16 hours for each part to cool. Suppose all parts belong to a product family, each part is identical from the light manufacturing perspective, but parts require different times to cool with the maximum time to cool being 16 hours. To keep track of exact inventories of completed parts, the input curves associated with each part within the family must be known. If the number of parts in the product family is large and if customer demands exhibit a high degree of substitution, then one may wish to model only the aggregate input curve associated with all parts in the family and keep track of only the aggregate inventory of completed parts. For example, it may be computationally necessary to reduce the number of variables in a formulation used for planning purposes (when considering all of the product families and the many other aspects to the process). If this is indeed the case, then it will be appropriate to view aggregate output as emerging *continuously* over a 16-hour period, the distribution of which depends on how the aggregate input function disaggregates into components associated with each part in the family.

Output emerging continuously over time can also arise when modeling processing times that are random. For example, one possible outcome of a testing or inspection process is a failed part that requires rework. The testing or inspection time typically depends on the type of diagnosis. Estimating the output of failed parts over time is required for planning resource requirements for a rework activity.

17.2 Input-Output Domain

We describe the class of functions we use to model the flows of goods and services. Points in time are modeled by the interval $(-\infty, \infty)$. Unless otherwise stated, each function of time is (i) finite-valued and nonnegative, and (ii) has compact support, i.e., the points in time where the function is positive is contained in a closed and bounded interval of time. There are two fundamental types of functions of time, event-based and rate-based flows. We describe each below.

17.2.1 Event-Based Flows

For discrete-parts manufacturing systems, the flows of inputs and outputs are event-based at the microscopic level. An **event-based flow** $z(\cdot)$ associates a nonnegative real number $z(\tau)$ to an event that occurs at time τ. For example, $z(\tau)$ might be the quantity of parts initiated into a production process at time τ or a value associated with a job completed at time τ. Event-based flows only take on positive values at those times when events occur. We assume the number of event occurrences in a bounded interval of time is finite.

We shall integrate event-based functions, but cannot use the ordinary (Riemann) integral, since the integral of an event-based function is always zero. For an event-based function $z(\cdot)$ a summation takes the place of an integral. For instance, the integral of $z(\cdot)$ on the interval $(-\infty, t]$ simply adds up all the values associated with the events that occur on or before time t, namely,

$$Z(t) := \sum_{\tau_i \leq t} z(\tau_i). \tag{17.1}$$

The function $Z(\cdot)$ is a step-function whose "jumps" occur at the times τ_i. The integral of $z(\cdot)$ on the time interval $(s, t]$ is the difference $Z(t) - Z(s)$.

17.2.2 Rate-Based Flows

For a **rate-based flow** $z(\cdot)$, the nonnegative real number $z(\tau)$ represents the rate (quantity per unit time) of flow at time τ. Rate-based flows sometimes represent a fluid approximation to event-based flows, and also arise quite naturally when modeling physical processes. We shall insist a rate-based flow is *piecewise continuous*, and the number of its discontinuities in any bounded time interval is finite. The intervals between adjacent discontinuity points define the pieces on which the rate-based flow is continuous. We shall often refer to the *cumulative* flow associated with a rate-based flow $z(\cdot)$, which is defined as the (Riemann) integral

$$Z(t) := \int_{-\infty}^{t} z(\tau) d\tau.$$

While it is certainly possible to imagine a "mixed" flow, in the developments to follow, a flow is either rate-based or event-based. Let D denote the set of all functions of time that are either event- or rate-based. A dynamic production function is a map from $\mathcal{X} \subset \mathsf{D}^n$ into D^m.

17.3 Instantaneous Processes

A dynamic production function is *instantaneous* if the outputs at time t are solely a function of the inputs at time t, and possibly other exogenous information that is t-dependent. It has the form

$$y(t) = [f(x)](t) = \Phi(x(t), t) := (\Phi_1(x(t), t), \ldots, \Phi_m(x(t), t)).$$

Example 17.1. Consider a single-input, single-output process characterized by

$$y(t) = ax(t),$$

where a unit of input instantaneously results in a units of output. The constant a could be less than one to model yield loss common to industries such as semiconductor manufacture.

Example 17.2. Another example from productivity measurement is a single-input, single-output process characterized by

$$y(t) = A(t)\phi(x(t)) = A(0)e^{gt}x(t);$$

here, output at time t is proportional to the input rate at time t, and the proportionality constant changes over time to reflect productivity improvements.

17.4 Index-Based Processes

17.4.1 Definition

Typically, when there are several inputs (e.g., multiple materials, subassemblies, machine and labor services), there is definite linkage in their use, especially in discrete-parts manufacturing. For example, in an assembly process, there is a well-defined recipe for the number of parts or subassemblies needed to make a finished product. The following definition captures this notion.

Definition 17.3. *A dynamic production function is* **index-based** *if each input vector $x(\cdot)$ in its domain $\mathcal{X}$ has the form*

$$x(t) = (x_1(t), x_2(t), \ldots, x_n(t)) = ([\xi_1(z)](t), [\xi_2(z)](t), \ldots, [\xi_n(z)](t)),$$

and the corresponding output vector has the form

$$\begin{aligned}[f(x)](t) = y(t) &= (y_1(t), y_2(t), \ldots, y_n(t)) \\ &= ([\psi_1(z)](t), [\psi_2(z)](t), \ldots, [\psi_n(z)](t)),\end{aligned}$$

where $\xi_i : \mathsf{D} \to \mathsf{D}$ and $\psi_i : \mathsf{D} \to \mathsf{D}$ are each one-to-one. That is, the components of each input vector and resulting output vector are uniquely determined by a single function $z(\cdot)$ called the **index***; the shape of any one input or output curve completely determines the shape of all remaining input and output curves.*

It is, of course, possible to define processes with several indexes, but we shall not explore this generalization.

The computational advantage of an index-based process can be considerable, since the structure of the input-output transformation reduces to specifying $n + m$ independent transformations that often possess relatively simple forms.

17.4.2 Fixed Proportions, Instantaneous Model

The simplest, classic example of an index-based process is the straightforward extension of a simple Leontief process to the dynamic setting. We call it a **fixed proportions dynamic model**. The production process is instantaneous and the inputs and outputs are in constant proportions. The input, output vectors are, respectively,

$$x(t) = (a_1, a_2, \ldots, a_n)z(t), \tag{17.2}$$
$$y(t) = (u_1, u_2, \ldots, u_m)z(t). \tag{17.3}$$

The vectors $a = (a_1, a_2, \ldots, a_n)$ and $u = (u_1, u_2, \ldots, u_m)$ are the **technical coefficients** that characterize this technology. The index $z(\cdot)$ is called the **intensity** of this process.

Example 17.4. An example from semiconductor manufacturing illustrates the use for vector-valued output. In semiconductor wafer manufacturing, each wafer consists of many die. In an ideal world, all die on the wafer would have identical characteristics. Due to (random) fluctuations, die on a single wafer are not identical and must be classified into different "bins" based on key operating characteristics. In this setting, $z(\cdot)$ indexes the amount of wafer starts and the u^k represent the (expected) proportion of die that will be classified into bin k, after accounting for yield loss. See Leachman et. al. [1996] and Leachman [2002] for a detailed description.

More generally, the technical coefficients could be functions of time, namely, each $a_i = a_i(\cdot)$ and $u_j = u_j(\cdot)$, in which case the input and output vectors are, respectively,

$$x(t) = (a_1(t), a_2(t), \ldots, a_n(t))z(t), \tag{17.4}$$
$$y(t) = (u_1(t), u_2(t), \ldots, u_m(t))z(t). \tag{17.5}$$

For this more general description, the inputs and output are in constant proportions at each point in time; however, these proportional constants may vary over time.

Example 17.5. Technical coefficients can change over time due to productivity improvements. Required inputs per unit of intensity can decline due to learning, operational improvements, etc., and the outputs per unit of intensity can increase due to better yields.

17.4.3 Fixed Proportions, Constant Lead Time Models

We describe three practical ways to use indexing to incorporate constant lead times into the fixed proportions dynamic model characterized by (17.2) and (17.3).

Indexing Non-storable Services

Here, $z(\cdot)$ indexes the non-storable labor and machine services, which are simultaneously used to produce a single output. The storable inputs, such as materials and subassemblies, are assumed to be withdrawn from inventory "just-in-time" for their use, but each may require a constant lead time for transportation, inspection, etc.

Let $\ell_k \geq 0$ denote the lead time for the k^{th} storable input. In this setting we have $x_i(t) = a_i z(t)$ for the i^{th} non-storable service input, and $x_k(t) = a_k z(t + \ell_k)$ for the k^{th} storable input, since its withdrawal from inventory occurs exactly ℓ_k time units before its use. As in the motivating example of Section 17.1, we assume output emerges a constant lead time $\rho \geq 0$ after use of the non-storable services; consequently, $y(t) = z(t - \rho)$.

Example 17.6. A production process uses two raw (storable) materials and one machine service to produce a single output. The relevant data are:

- 12 units of material 1 are required per unit of output, and it takes 2 hours to transport this material to the machine station. Here, $a_1 = 12$ and $\ell_1 = 2$.
- 18 units of material 2 are required per unit of output, and it takes 3 hours to transport this material to the machine station. Here, $a_2 = 18$ and $\ell_2 = 3$.
- 2 hours of machine service are required per unit of output. Here, $a_3 = 2$.
- After machining has taken place, it takes 5 hours to inspect the semi-finished output, after which the completed output is available to service demand. Here, $\rho = 5$.

Between hours 100 and 102, a total of 32 hours of machine services has been consumed, uniformly spread over this period. Here, $x_3(t) = 16$ for $t \in [100, 102]$, and $z(t) = 16$ for $t \in [100, 102]$ since $x_3(t) = z(t)$. In words, 8 units are being machined at a constant rate during this two-hour period of time. Given a lead time of 2 hours for material 1 (and the just-in-time assumption), the withdrawal rate of material 1 input is $x_1(t) = 12(8) = 96$ for $t \in [98, 100]$. Similarly, given a lead time of 3 hours for material 2 (and the just-in-time assumption), the withdrawal rate of material 2 input is $x_2(t) = 18(8) = 144$ for $t \in [97, 99]$. Given that it takes 5 hours to inspect the semi-finished output as it emerges from the machining process, the final output rate is $y(t) = 8$ for $t \in [105, 107]$. In terms of this index,

$$\begin{aligned} z(t) &= 16 \text{ for } t \in [100, 102], \\ x_1(t) &= 6z(t + \ell_1) = 6z(t + 2), \\ x_2(t) &= 9z(t + \ell_2) = 9z(t + 3), \\ x_3(t) &= z(t), \\ y(t) &= 0.5z(t - \rho) = 0.5z(t - 5). \end{aligned}$$

Indexing "Outs"

A second approach to indexing is to let $z(\cdot)$ index the output or "outs" of the process in which case $y(t) = z(t)$. On the input side, $x_i(t) = a_i z(t+\rho)$ for the i^{th} non-storable service input, since its usage occurs exactly ρ time units before the product emerges as output, and $x_k(t) = a_k z(t+\ell_k+\rho)$ for the k^{th} storable input, since its withdrawal from inventory occurs exactly $\rho+\ell_k$ time units before the product emerges as output.

Example 17.7. Consider the same data provided in Example 17.6. Between hours 105 and 107, a total of 16 units of completed output has emerged, uniformly spread over this period. Here, $z(t) = 8$ for $t \in [105, 107]$. The function $z(\cdot)$ here does *not* equal the $z(\cdot)$ function in the previous example since here it is being used to index output and not machine services. The consumption of resources has not changed, i.e., it is still the case that

$$\begin{aligned} x_1(t) &= 12(8) = 96,\ t \in [98, 100], \\ x_2(t) &= 18(8) = 144,\ t \in [97, 99], \\ x_3(t) &= 16,\ t \in [100, 102]. \end{aligned}$$

What has changed, however, is how these functions relate to the chosen index. In terms of this index,

$$\begin{aligned} z(t) &= 8 \text{ for } t \in [105, 107], \\ x_1(t) &= 12z(t+\ell_1+\rho) = 12z(t+7), \\ x_2(t) &= 18z(t+\ell_2+\rho) = 18z(t+8), \\ x_3(t) &= 2z(t+\rho) = 2z(t+5), \\ y(t) &= z(t). \end{aligned}$$

Example 17.8. Consider the same data provided in Example 17.6, except that now *two* simultaneous semi-finished outputs emerge after machining in a 3:1 ratio, i.e., a total of 12 units of semi-finished output 1 and a total of 4 units of semi-finished output 2 emerge uniformly over the interval [100, 102]. In this example, it takes 5 hours to inspect output 1 (as before), whereas now it takes 9 hours to inspect output 2.

The notion of outs in this example cannot apply simultaneously to both outputs, since there are two nonequal ρ's, i.e., $\rho^1 = 5$ and $\rho^2 = 9$. It has to apply to one of the outputs, from which the other output and inputs can be determined. If output 1 is chosen as the index, then

$$\begin{aligned} z(t) &= 6,\ t \in [105, 107], \\ x_1(t) &= 16z(t+7),\ t \in [98, 100], \\ x_2(t) &= 24z(t+8),\ t \in [97, 99], \\ x_3(t) &= (8/3)z(t+5),\ t \in [100, 102], \\ y^1(t) &= z(t),\ t \in [105, 107], \\ y^2(t) &= (1/3)z(t-7),\ t \in [109, 111]. \end{aligned}$$

On the other hand, if output 2 is chosen as the index, then

$$\begin{aligned} z(t) &= 2,\ t \in [97, 99], \\ x_1(t) &= 48z(t),\ t \in [97, 99], \\ x_2(t) &= 72z(t-1),\ t \in [98, 100], \\ x_3(t) &= 8z(t-3),\ t \in [100, 102], \\ y^1(t) &= 3z(t+4),\ t \in [105, 107], \\ y^2(t) &= z(t),\ t \in [109, 111]. \end{aligned}$$

Indexing "Starts"

For the special case when the storable lead times are all identical, say $\ell_i = \ell$, then a third approach to indexing is to let $z(\cdot)$ index the "starts" of the process in which case $x_k(t) = a_k z(t)$ for the k^{th} storable input and $x_i(t) = a_i z(t-\ell)$ for the i^{th} non-storable service input, since the usage of the non-storable service occurs exactly ℓ units after the withdrawal of the storable inputs. On the output side, $y(t) = z(t-\ell-\rho)$.

Example 17.9. Suppose the data in Example 17.6 is changed so that $\ell_1 = \ell_2 = 2.5$ (the average of 2 and 3). The machine services are still consumed uniformly over the interval [100, 102]. Since the function $z(\cdot)$ now indexes starts,

$$\begin{aligned} z(t) &= 8 \text{ for } t \in [97.5, 99.5], \\ x_1(t) &= 12z(t),\ t \in [97.5, 99.5], \\ x_2(t) &= 18z(t),\ t \in [97.5, 99.5], \\ x_3(t) &= 2z(t-2.5),\ t \in [100, 102], \\ y(t) &= z(t-7.5),\ t \in [105, 107]. \end{aligned}$$

When the storable lead times are not all identical, then the notion of a start cannot apply to all inputs simultaneously—it has to apply to one of the inputs, from which the other inputs and output can be determined. Suppose the lead times $\ell_1 = 2$ and $\ell_2 = 3$, as before. If input 1 is chosen as the index, then

$$\begin{aligned} z(t) &= 96,\ t \in [98, 100], \\ x_1(t) &= z(t),\ t \in [98, 100], \\ x_2(t) &= 1.5z(t+1),\ t \in [97, 99], \\ x_3(t) &= (1/6)z(t-2),\ t \in [100, 102], \\ y(t) &= (1/12)z(t-7),\ t \in [105, 107]. \end{aligned}$$

On the other hand, if input 2 is chosen as the index, then

$$\begin{aligned} z(t) &= 144,\ t \in [97, 99], \\ x_1(t) &= (2/3)z(t-1),\ t \in [98, 100], \end{aligned}$$

$$x_2(t) = z(t),\ t \in [97, 99],$$
$$x_3(t) = (1/9)z(t-3),\ t \in [100, 102],$$
$$y(t) = (1/18)z(t-8),\ t \in [105, 107].$$

Remark 17.10. Under restrictive assumptions, the notion of starts and outs for the whole process can apply and can be used to index the consumption of non-storable resources, the withdrawal of storable inputs, and subsequent final output. In general, one can still speak of starts and outs and use them to index the process: in the case of starts, one of the inputs is chosen as the index; in the case of outs, one of the outputs is chosen as the index. When the consumption of non-storable resources is chosen as the index, all inputs and outputs can be related to it (as long as lead times are constant).

17.5 Exercises

17.1. Consider the data of Example 17.6. Suppose between the hours of 205 and 209, a total of 192 hours of machine services has been consumed, uniformly spread over this period. Ignore the previous input described in this example.

(a) What are the specific values for $z(\cdot)$, $x_1(\cdot)$, $x_2(\cdot)$, $x_3(\cdot)$ and $y(\cdot)$ if $z(\cdot)$ is used to index the non-storable services?
(b) What are the specific values for $z(\cdot)$, $x_1(\cdot)$, $x_2(\cdot)$, $x_3(\cdot)$ and $y(\cdot)$ if $z(\cdot)$ is used to index the outs?
(c) Suppose, as in Example 17.8, $\rho_1 = 5$ and $\rho_2 = 9$.
  (i) What are the specific values for $z(\cdot)$, $x_1(\cdot)$, $x_2(\cdot)$, $x_3(\cdot)$ and $y(\cdot)$ if $z(\cdot)$ is used to index output 1?
  (ii) What are the specific values for $z(\cdot)$, $x_1(\cdot)$, $x_2(\cdot)$, $x_3(\cdot)$ and $y(\cdot)$ if $z(\cdot)$ is used to index output 2?
(d) Suppose, as in Example 17.9, $\ell_1 = \ell_2 = 2.5$. What are the specific values for $z(\cdot)$, $x_1(\cdot)$, $x_2(\cdot)$, $x_3(\cdot)$ and $y(\cdot)$ if $z(\cdot)$ is used to index the starts?
(e) Suppose, as in Example 17.9, $\ell_1 = 2$ and $\ell_2 = 3$.
  (i) What are the specific values for $z(\cdot)$, $x_1(\cdot)$, $x_2(\cdot)$, $x_3(\cdot)$ and $y(\cdot)$ if $z(\cdot)$ is used to index input 1?
  (ii) What are the specific values for $z(\cdot)$, $x_1(\cdot)$, $x_2(\cdot)$, $x_3(\cdot)$ and $y(\cdot)$ if $z(\cdot)$ is used to index input 2?

17.2. A production process uses two raw (storable) materials and one machine service to produce two outputs. The relevant data are:

- Two simultaneous semi-finished outputs emerge after machining in a 2:1 ratio.
- 9 units of material 1 are required per unit of aggregate output, and it takes 1 hour to transport this material to the machine station. Here, $a_1 = 9$ and $\ell_1 = 1$.

- 27 units of material 2 are required per unit of aggregate output, and it takes 4 hours to transport this material to the machine station. Here, $a_2 = 27$ and $\ell_2 = 4$.
- 3 hours of machine service are required per unit of aggregate. Here, $a_3 = 3$.
- After machining has taken place, it takes 8 hours to inspect semi-finished output 1 and 12 hours to inspect semi-finished output 2. Here, $\rho_1 = 8$ and $\rho_2 = 12$.
- Between hours 100 and 110, a total of 900 hours of machine services has been consumed, uniformly spread over this period.

(a) What are the specific values for $x_1(\cdot)$, $x_2(\cdot)$, $x_3(\cdot)$, $y_1(\cdot)$ and $y_2(\cdot)$?
(b) What is the general form for the functions $x_1(\cdot)$, $x_2(\cdot)$, $x_3(\cdot)$, $y_1(\cdot)$ and $y_2(\cdot)$ in terms of the index $z(\cdot)$ when the index is chosen to represent the
(i) non-storable machine service?
(ii) output 1?
(iii) output 2?
(iv) input 1?
(v) input 2?

17.6 Bibliographical Notes

See Hackman [1990] for an in-depth discussion of acceptable properties of dynamic production functions.

17.7 Solutions to Exercises

17.1 (a) We have

$$\begin{aligned} z(t) &= 48,\ t \in [205, 209], \\ x_1(t) &= 288,\ t \in [203, 207], \\ x_2(t) &= 432,\ t \in [202, 206], \\ x_3(t) &= 48,\ t \in [205, 209], \\ y(t) &= 24,\ t \in [210, 214]. \end{aligned}$$

(b) All functions remain unchanged, except that now $z(t) = 24$, $t \in [210, 214]$.
(c) For (i),

$$\begin{aligned} z(t) &= 18,\ t \in [210, 214], \\ y_1(t) &= 18,\ t \in [210, 214], \\ y_2(t) &= 6,\ t \in [214, 218]. \end{aligned}$$

All other functions remain unchanged. As for (ii), now $z(t) = 6$, $t \in [210, 214]$. All other functions remain unchanged from their values in (i).
(d) Here,

$$\begin{aligned} z(t) &= 24,\ t \in [202.5, 206.5], \\ x_1(t) &= 288,\ t \in [202.5, 206.5], \\ x_2(t) &= 432,\ t \in [202.5, 206.5]. \end{aligned}$$

All other functions remain the same.
(e) Only the $z(\cdot)$ function changes. For part (i), $z(t) = 288$, $t \in [203, 207]$, whereas for part (ii), $z(t) = 432$, $t \in [202, 206]$.

17.2 (a) The specific values are:

$$\begin{aligned} x_1(t) &= 270,\ t \in [99, 109], \\ x_2(t) &= 810,\ t \in [96, 106], \\ x_3(t) &= 90,\ t \in [100, 110], \\ y_1(t) &= 20,\ t \in [108, 118], \\ y_2(t) &= 10,\ t \in [112, 122]. \end{aligned}$$

(b) For part (i):

$$\begin{aligned} x_1(t) &= 3z(t+1), \\ x_2(t) &= 9z(t+4), \\ x_3(t) &= z(t), \\ y_1(t) &= (2/9)z(t-8), \\ y_2(t) &= (1/9)z(t-12). \end{aligned}$$

For part (ii):

$$
\begin{aligned}
x_1(t) &= (27/2)z(t+9),\\
x_2(t) &= (81/2)z(t+12),\\
x_3(t) &= (9/2)z(t+8),\\
y_1(t) &= z(t),\\
y_2(t) &= (1/2)z(t-4).
\end{aligned}
$$

For part (iii):

$$
\begin{aligned}
x_1(t) &= 27z(t+13),\\
x_2(t) &= 81z(t+16),\\
x_3(t) &= 9z(t+12),\\
y_1(t) &= 2z(t+4),\\
y_2(t) &= z(t).
\end{aligned}
$$

For part (iv):

$$
\begin{aligned}
x_1(t) &= z(t),\\
x_2(t) &= 3z(t+3),\\
x_3(t) &= (1/3)z(t-1),\\
y_1(t) &= (2/27)z(t-9),\\
y_2(t) &= (1/27)z(t-13).
\end{aligned}
$$

For part (v):

$$
\begin{aligned}
x_1(t) &= 3z(t-3),\\
x_2(t) &= z(t),\\
x_3(t) &= (1/9)z(t-4),\\
y_1(t) &= (2/81)z(t-12),\\
y_2(t) &= (1/81)z(t-16).
\end{aligned}
$$

Part III

Productivity and Performance Measurement

13

Index Numbers

A *price index* is a measure that summarizes the *change* in the prices of a basket of goods (or a group of inputs) from one time period to the next. A *Consumer Price Index* (CPI) is used by policy makers to estimate the inflation rate, which is then used, for example, to make cost-of-living adjustments. A price index is also constructed to measure the difference in prices in two different locations, in which case it could be used to compensate an employee for relocation. A *Producer Price Index* (PPI) is used by policy makers and businesses to estimate the price changes that affect the production side of the economy. In the United States, the Bureau of Labor Statistics (BLS) and the Bureau of Economic Analysis (BEA) compute a host of indexes for many categories (e.g., materials, energy, labor productivity).

While price indexes are most relevant for consumers, quantity indexes can be used to measure a firm's performance. A *quantity index* is a measure that summarizes the *change* in the quantities of outputs produced (or products consumed) or inputs used. The ratio of the growth in aggregate output to the growth in aggregate input can be used to assess a firm's productivity from one year to the next. Such a computation presupposes a rational method for aggregating the outputs and inputs, which is equivalent to constructing separate quantity indexes for output and input.

In this chapter, we develop several price and quantity indexes and explain the rationale for each. In the next two chapters, we show how to apply these indexes to assess a firm's productivity and overall performance from one year to the next.

13.1 Motivating Example

Consider the following price-quantity data collected for a consumer or producer for two time periods, t = 0 ("last year") and t = 1 ("current year"), displayed in Table 13.1. On the producer side, the goods may either be outputs, in which case the prices correspond to the per-unit revenues the firm

receives, or inputs, in which case the prices correspond to the per-unit costs the firm pays. The outputs or inputs can also belong to a category, e.g., product group, materials, labor, etc. The goal is to obtain a reasonable estimate of the price and quantity changes over the two periods. In the discussion to follow, the price-quantity data corresponds to a representative consumer.

Table 13.1. Price-quantity data for motivating example.

	Period 0		Period 1	
	Price	Quantity	Price	Quantity
Good 1	1.00	2	0.70	10
Good 2	2.00	3	2.25	4
Good 3	1.50	8	1.80	5

The consumer spent $E^0 = 20$ in period 0, $E^1 = 25$ in period 1 for an expenditure ratio of 1.25. The number 1.25, or equivalently $100(1.25 - 1) = 25\%$, cannot be used to assess increases in the cost-of-living, aggregate price, or aggregate quantity for three fundamental reasons. First, the expenditure ratio suffers from *substitution bias*, since it ignores the "substitution effect"—the consumer will adjust his choices of goods to purchase when prices change. Second, the expenditure ratio suffers from *utility bias*, since it ignores the "utility effect"—the consumer has preferences for the goods purchased. For example, the consumer may strongly prefer the consumption bundle in period 1 to the one purchased in period 0, in which case there is an "apples-oranges" problem. The third problem, *quality bias*, is always lurking in the shadows. Suppose a unit of good i in period 1 is "better" than a unit of good i in period 0. It is tacitly assumed that quantities are measured in *quality-adjusted* units, which, in practice, is a difficult task. For certain goods (e.g. computers), it is possible to form a **hedonistic index** to measure the quality changes (e.g., some weighted combination of factors such as CPU speed, memory, etc.). For government services, this task is so difficult the BEA often assumes no quality change has occurred. To the extent a price index understates the quality effect, it will overstate the price effect.

We now present several possible ways to construct a price index:

- *Average level of prices.* The average price levels are

$$\frac{(1.00 + 2.00 + 1.50)}{3} = 1.5000, \ \frac{(0.70 + 2.25 + 1.80)}{3} = 1.5833,$$

  in periods 0 and 1, respectively. The ratio of the average price levels is $1.5833/1.500 = 1.0556$, suggesting a 5.56% price increase. This index suffers from a fatal flaw: it is not invariant to a change in the unit of measurement. Consider how the index changes if the prices of good 1 are multiplied

by 10 due to a change in the units used to measure the quantities: the new index equals $(11.05/3)/(13.50/3) = 0.8185$, suggesting an 18.15% *decline* in prices.

- *Average or geometric average of price ratios.* The fatal flaw associated with the use of the average level of prices can be overcome by instead taking the ratio of (i) the average of the price ratios, or (ii) the geometric average of the price ratios. In case (i), the index is calculated as

$$\frac{(0.70/1.00) + (2.25/2.00) + (1.80/1.50)}{3} = 1.0083,$$

and in case (ii), the index is calculated as

$$\left(\frac{0.70}{1.00}\right)^{1/3}\left(\frac{2.25}{2.00}\right)^{1/3}\left(\frac{1.80}{1.50}\right)^{1/3} = 0.9813.$$

There is a statistical justification for either index. If the price ratios are independently and symmetrically distributed about a common mean, and this distribution is normal, the maximum likelihood estimator for the common mean corresponds to the first index. If the natural log of the price ratios are independently and symmetrically distributed about a common mean, and this distribution is normal, the maximum likelihood estimator for the common mean corresponds to the second index. It has long been empirically established that price ratios are not so distributed; price movements are co-mingled through the general equilibrium of the economy.

- *Weighted average of the price ratios.* The indexes mentioned so far do not use any quantity information, and thus have no hope of avoiding the utility bias. The next index weights the price ratios by the *expenditure shares* in period 0, which were

$$\frac{(1.00)(2)}{20} = 0.10, \ \frac{(2.00)(3)}{20} = 0.30, \ \frac{(1.50)(8)}{20} = 0.60,$$

for goods 1, 2 and 3, respectively. The index is calculated as

$$(0.10)\left(\frac{0.70}{1.00}\right) + (0.30)\left(\frac{2.25}{2.00}\right) + (0.60)\left(\frac{1.80}{1.50}\right) = 1.1275, \tag{13.1}$$

suggesting a price or cost-of-living increase of 12.75%.[1] This index is known as the *Laspeyres price index*, and is more commonly calculated as the ratio of what it would cost in year 1 to purchase the same goods in year 0, namely,

$$(0.70)(2) + (2.25)(3) + (1.80)(8) = 22.55,$$

to the expenditure in year 0, namely, 20. The index (13.1) is constructed "forward in time." A similar index is constructed by going "backwards in

[1] Of course, the average price ratio is a weighted average of the price ratios, albeit with equal weights.

time," namely, by interchanging the roles of period 0 with period 1 and *taking the reciprocal*[2], as follows:

$$\left\{(0.28)\left(\frac{1.00}{0.70}\right)+(0.36)\left(\frac{2.00}{2.25}\right)+(0.36)\left(\frac{1.50}{1.80}\right)\right\}^{-1}=0.9804, \quad (13.2)$$

suggesting a cost-of-living or price *decrease* of 1.96%. This index is known as the *Paasche price index*, and is more commonly calculated as the ratio of the expenditure in year 1, namely, 25, to what it would cost in year 0 to purchase the same goods in year 1, namely,

$$[1.00(10)+2.00(4)+1.50(5)]=25.50.$$

Both the Laspeyres and Paasche indexes have merit, but they lead to dramatically different estimates, 1.1275 and 0.9804, respectively.[3] A compromise is to take the *geometric mean* of these two indexes,

$$\sqrt{(1.1275)(0.9804)}=1.0514,$$

now known as the *Fisher ideal index*, which suggests a cost-of-living or price increase of 5.14%.

- *Weighted geometric mean of price ratios.* The most commonly used index of this type uses the average expenditure shares as the weights.[4] The average expenditure shares are

$$\frac{(0.10+0.28)}{2}=0.19,\ \frac{(0.30+0.36)}{2}=0.33,\ \frac{(0.60+0.36)}{2}=0.48,$$

for the goods 1, 2 and 3, respectively. Known as the *Tornqvist price index* it is calculated as

$$\left(\frac{0.70}{1.00}\right)^{0.19}\left(\frac{2.25}{2.00}\right)^{0.33}\left(\frac{1.80}{1.50}\right)^{0.48}=1.0803,$$

suggesting a cost-of-living or price increase of 8.03%.

So far, we have discussed price indexes. Interchanging the roles of prices and quantities leads to analogous quantity indexes. There is also an *implicit* method to obtaining a quantity index: Associate with each price index P the quantity index Q defined by the identity $P \cdot Q = E^1/E^0$. For the given price-quantity data, the

[2] The reciprocal is necessary. Consider what would happen if the prices in year 1 were twice the prices in year 0 and the quantities purchased were identical. Without taking the reciprocal, the index would suggest a price *decline* of 50%.

[3] This is largely due to the dramatic swings in the price components from one period to the next in the example data. These two indexes differ in practice by at most 1%. These indexes can lead to quite different results if the price-quantity data correspond to different locations within the same year.

[4] Of course, the geometric average of the price ratios is a weighted geometric mean albeit with equal weights.

- *implicit Laspeyres quantity index* is 1.1086, since $(1.1275)(1.1086) = 1.25$;
- *implicit Paache quantity index* is 1.2750, since $(0.9804)(1.2750) = 1.25$;
- *implicit Fisher ideal price index* is 1.1889, since $(1.0514)(1.1889) = 1.25$;
- *implicit Tornqvist price index* is 1.1571, since $(1.0803)(1.1571) = 1.25$.

Similarly, one may associate with each quantity index Q the *implicit* price index P so that $Q \cdot P = E^1/E^0$. Implicit indexes are convenient since, by construction, the product of the price and quantity indexes fully accounts for the growth (or decline) in the expenditure ratio. They are also useful when a portion of the detailed price-quantity data is not available.

To be concrete in the developments to follow, we shall interpret the price-quantity data in the producer context. We let p^t, x^t and $\Phi(x^t)$, denote the price vectors, input vectors, and maximal outputs for periods $t = 0, 1$. All economic variables (e.g., prices, quantities) are assumed positive.

13.2 Price Indexes

13.2.1 Konus Price Index

Definition 13.1. *The* **Konus price index** *is*

$$P_K(p^0, p^1, x) := \frac{Q(\Phi(x), p^1)}{Q(\Phi(x), p^0)} ,$$

where $Q(u, p)$ denotes the minimal cost (expenditure) function.

The Konus price index does not suffer from substitution bias or utility bias: it tackles the substitution bias via the expenditure function, which incorporates changes in quantities purchased due to price changes, and it tackles the utility bias by insisting that expenditures are compared for the *same* level of output.The output levels u^0 and u^1 will *not*, in general, coincide, and they certainly may not equal the output level $u := \Phi(x)$ corresponding to the arbitrarily chosen quantity vector x.

The Konus index depends on the particular choice of x. There is one useful setting in which the Konus index is *independent* of x, namely, when the technology satisfies homotheticity. A function $\Phi(\cdot)$ is homothetic if and only if it can be represented as $F(\phi(x))$ where $\phi(\cdot)$ exhibits constant returns-to-scale and $F(\cdot)$ is a suitable transform—see Definition 2.28.

Theorem 13.2. *P_K is independent of x if and only if $\Phi(\cdot)$ is homothetic.*

Proof. A production function $\Phi(\cdot)$ is homothetic if and only if the cost function factors as $Q(u, p) = \Lambda(u)P(p)$, where $\Lambda(\cdot)$ is strictly increasing and $P(\cdot)$ is homogeneous of degree one. (We established this result as an application of the duality between the cost and distance function—see Section 7.4.) If the expenditure function factors, then obviously the Konus index is independent

of x. Conversely, suppose the Konus index is independent of x. Fix p_0 and note that $Q(u, p_0) > 0$. We may express the cost function as

$$Q(u,p) = Q(u,p_0)\left(\frac{Q(u,p)}{Q(u,p_0)}\right) := f(u)P(p).$$

Clearly, $f(\cdot)$ and $P(\cdot)$ have the requisite properties, which shows that the expenditure function factors, as required. □

The following Proposition establishes easy-to-compute, natural bounds on the Konus price index.

Proposition 13.3. *For any choice of x,*

$$\min_i p_i^1/p_i^0 \le P_K(p^0,p^1,x) \le \max_i p_i^1/p_i^0.$$

Proof. Fix x or u. For each $i = 0, 1$, let y^i denote an optimum solution to the $Q(u, p^i)$ cost minimization problem. By definition of the y^i and the Konus price index,

$$P_K(p^0,p^1,x) = \frac{p^1 \cdot y^1}{p^0 \cdot y^0}, \tag{13.3}$$

$$p^1 \cdot y^1 \le p^1 \cdot y^0, \tag{13.4}$$

$$p^0 \cdot y^0 \le p^0 \cdot y^1. \tag{13.5}$$

Thus,

$$P_K(p^0,p^1,x) \ge \sum_i \left(\frac{p_i^1}{p_i^0}\right)\left[\frac{p_i^0\, y_i^1}{p^0 \cdot y^1}\right] \ge \min_i p_i^1/p_i^0, \tag{13.6}$$

$$P_K(p^0,p^1,x) \le \sum_i \left(\frac{p_i^1}{p_i^0}\right)\left[\frac{p_i^0\, y_i^0}{p^0 \cdot y^0}\right] \le \max_i p_i^1/p_i^0, \tag{13.7}$$

since a convex combination of positive numbers is bounded below (above) by the minimum (maximum) of those positive numbers. □

13.2.2 Laspeyres and Paasche Price Indexes

Considerably sharper bounds can be achieved if instead of selecting an arbitrary x in Proposition (13.3), somewhat more hospitable choices, namely, x^0 or x^1, are chosen.

Definition 13.4. *The* **Laspeyres-Konus price index** *is $P_K(p^0, p^1, x^0)$ and the* **Paasche-Konus price index** *is $P_K(p^0, p^1, x^1)$.*

Definition 13.5. *The* **Laspeyres price index** *is*

$$P_L(p^0, p^1, x^0, x^1) := \frac{p^1 \cdot x^0}{p^0 \cdot x^0} .$$

The **Paasche price index** *is*

$$P_P(p^0, p^1, x^0, x^1) := \frac{p^1 \cdot x^1}{p^0 \cdot x^1} .$$

We suppress the functional dependence of the prices and quantity bundles, and simply denote the Laspeyres and Paasche price indexes as P_L and P_P, respectively.

The following Corollary to Proposition 13.3 shows how the use the Laspeyres and Paasche price indexes to respectively bound the Laspeyres-Konus and Paasche-Konus indexes.

Assumption 4 *Each x^t is a cost (expenditure) minimizer, namely, each x^t satisfies*

$$p^t \cdot x^t = Q(\Phi(x^t), p^t). \tag{13.8}$$

Corollary 13.6. *Under Assumption 4,*

$$\min_i p_i^1/p_i^0 \leq P_K(p^0, p^1, x^0) \leq P_L,$$
$$P_P \leq P_K(p^0, p^1, x^1) \leq \max_i p_i^1/p_i^0.$$

If $\Phi(\cdot)$ is homothetic, then P_K is independent of x by Theorem 13.2, in which case

$$P_K(p^0, p^1, x^0) = P_K(p^0, p^1, x) = P_K(p^0, p^1, x^1)$$

for each choice of x. As an immediate consequence of Corollary 13.6, we have:

Corollary 13.7. *If $\Phi(\cdot)$ is homothetic, then*

$$P_P \leq P_K(p^0, p^1, x) \leq P_L$$

holds for each choice of x.

Remark 13.8. Under the homothetic setting, the Paasche price index lies below the Laspeyres price index, which typically occurs in practice. However, if $\Phi(\cdot)$ is not homothetic, it is possible that $P_P > P_L$.

Remark 13.9. The Paasche and Laspeyres price indexes will bound the Konus price index as in Corollary 13.7 if the homothetic condition is replaced by the condition $u^0 = u^1 = u$. However, this last condition is a bit much to expect for observed data.

Given the available data the indexes, P_P and P_L, are easy to construct. Can any statement be made about the relationship between P_P, P_L, and P_K when $\Phi(\cdot)$ is *not* homothetic? Under reasonable regularity conditions, the following relationships exist:

(i) If $P_L \leq P_P$, then for each $\mu \in [0,1]$, the value $\mu P_L + (1-\mu)P_P$ *must coincide with a Konus price index* for some choice of x, which, moreover, can be taken to be a convex combination $x = \lambda x^0 + (1-\lambda)x^1$ of x^0 and x^1 for some $\lambda \in [0,1]$.

(ii) If $P_P \leq P_L$, then there is at least one value of $\mu \in [0,1]$ for which the value $\mu P_L + (1-\mu)P_P$ *will coincide with a Konus price index* for some choice of x, which, moreover, can be taken to be a convex combination $x = \lambda x^0 + (1-\lambda)x^1$ of x^0 and x^1 for some $\lambda \in [0,1]$.

Assumption 5 *$\Phi(\cdot)$ is continuous, strictly quasiconcave and increasing.*

Theorem 13.10. *Under Assumption 5,*

a) if $P_P \geq P_L$, then for each value of $\mu \in [0,1]$ there exists a $\lambda_\mu \in [0,1]$ such that

$$P_K(p^0, p^1, \lambda_\mu x^0 + (1-\lambda_\mu)x^1) = \mu P_P + (1-\mu)P_L.$$

b) if $P_P \leq P_L$, then there exists a $\lambda \in [0,1]$ such that

$$P_P \leq P_K(p^0, p^1, \lambda x^0 + (1-\lambda)x^1) \leq P_L.$$

Proof. Let

$$\psi(\lambda) := P_K(p^0, p^1, \lambda x^0 + (1-\lambda)x^1).$$

The Theorem of the Maximum, Appendix H, guarantees that $Q(\Phi(x), p)$, and hence P_K, will be continuous in x as long as $\Phi(\cdot)$ is continuous and is well-behaved. Continuity of P_K in x implies continuity of $\psi(\cdot)$ on $[0,1]$. Since $\psi(\cdot)$ is continuous, its image $\psi[0,1]$ must be an interval.[5] Corollary 13.6 implies this interval contains the interval $[P_L, P_P]$ in the case of part (a) and intersects the interval $[P_P, P_L]$ in the case of part (b). The result follows. □

13.3 Fisher and Tornqvist Price Indexes

13.3.1 Fisher Ideal Price Index

In this section we make the following assumption:

Assumption 6 *$\Phi(x) = \sqrt{x^T A x}$, and $\Phi(\cdot)$ is restricted to an open domain $S \subset R^n_+$ on which $x^T A x > 0$ and $\Phi(\cdot)$ is concave. Without loss of generality, A is taken to be symmetric so that $A = A^T$.*[6]

[5] A continuous image of a connected set is connected, and a subset of the real line is connected if and only if it is an interval.

[6] If A is not symmetric, replace it with $1/2(A + A^T)$.

Since $\Phi(\cdot)$ exhibits constant returns-to-scale, and thus is homothetic, the Konus index equals the ratio $Q(1,p^1)/Q(1,p^0)$. In fact, we show below the Konus index equals the Fisher ideal index.

Definition 13.11. *The* **Fisher ideal price index** *is* $P_F := \sqrt{P_P\, P_L}$.

Lemma 13.12. *If* $p \cdot z = Q(1,p)$, *then* $p = (p \cdot z) z^T A$.

Proof. By definition,

$$Q(1,p) = \min\{p \cdot x : \sqrt{x^T A x} \geq 1\} = \min\{p \cdot x : x^T A x \geq 1\}.$$

Since z is a cost minimizer, first-order optimality conditions imply that $p = 2\lambda z^T A$. Thus, $p \cdot z = 2\lambda(z^T A z) = 2\lambda$, since the constraint must be tight at the optimum, and the result follows. □

Theorem 13.13. $P_F = P_K$ *under Assumption 6.*

Proof. Let $z_i = x^i/\Phi(x^i)$ and note that $p^i \cdot z_i = Q(1,p^i)$. Using Lemma 13.12,

$$p^1 = (p^1 \cdot z_1) z_1^T A \text{ and } p^0 = (p^0 \cdot z_0) z_0^T A. \tag{13.9}$$

Thus,

$$P_L = \frac{p^1 \cdot x^0}{p^0 \cdot x^0} = \frac{p^1 \cdot z_0}{p^0 \cdot z_0} = \frac{(p^1 \cdot z_1)(z_1^T A z_0)}{p^0 \cdot z_0}, \tag{13.10}$$

$$P_P = \frac{p^1 \cdot x^1}{p^0 \cdot x^1} = \frac{p^1 \cdot z_1}{p^0 \cdot z_1} = \frac{p^1 \cdot z_1}{(p^0 \cdot z_0)(z_0^T A z_1)}\,. \tag{13.11}$$

Since A is symmetric,

$$P_L \cdot P_P = \left(\frac{p^1 \cdot z_1}{p^0 \cdot z_0}\right)^2 = \left(\frac{Q(1,p^1)}{Q(1,p^0)}\right)^2, \tag{13.12}$$

and the result follows. □

13.3.2 Tornqvist Price Index

In this section, we assume $\Phi(\cdot)$ is homothetic so that $P_K = Q(1,p^1)/Q(1,p^0)$. We also make the following assumption:

Assumption 7 *The unit cost function* $Q(1,p)$ *has the translog form, i.e.,*

$$ln\, Q(1,p) = a_0 + \sum_i a_i\, ln\, p_i + 1/2 \sum_i \sum_j a_{ij}\, ln\, p_i\, ln\, p_j, \tag{13.13}$$

where $a_{ij} = a_{ji}$.[7]

[7] Since a well-behaved cost function is also homogeneous of degree one, it must also be true that $\sum_i a_i = 1$ and that $\sum_i a_{ij} = \sum_j a_{ij} = 0$.

Let $h(z) := a_0 + a^T z + 1/2\ z^T A z$ denote a quadratic form with A symmetric. An examination of (13.13) shows that $\ln Q(1,p) = h(z(p))$ where $z(p) := (\ln p_1, \ln p_2, \ldots, \ln p_n)$. By the chain rule

$$\frac{\partial h(z(p))}{\partial p_k} = \frac{\partial h(z(p))}{\partial z_k}\frac{1}{p_k}; \tag{13.14}$$

it therefore follows that

$$\frac{\partial h(z)}{\partial z_k} = p_k \frac{\partial \ln Q(1,p)}{\partial p_k} = \frac{p_k \frac{\partial Q(1,p)}{\partial p_k}}{Q(1,p)} = \epsilon_i^{Q(1,p)}\ , \tag{13.15}$$

the elasticity of cost $Q(1,p)$ with respect to input i. We use this fact below.

Since $\Phi(\cdot)$ is homothetic, the Konus index equals the ratio $Q(1,p^1)/Q(1,p^0)$. In fact, we show below the Konus price index equals the Tornqvist price index.

Definition 13.14. *The* **Tornqvist price index** *is*

$$P_T := \prod_i \left(\frac{p_i^1}{p_i^0}\right)^{(S_i^1+S_i^0)/2},$$

where S_i^t denotes the expenditure share of good i in period $t = 0, 1$.

Lemma 13.15. *Let $h(z) = a_0 + a^T z + 1/2\ z^T A z$ denote a quadratic form with A symmetric. For each choice of z^1 and z^0,*

$$h(z_1) - h(z_0) = 1/2\ [\nabla h(z_1) + \nabla h(z_0)] \cdot (z_1 - z_0)\ . \tag{13.16}$$

Proof. Since $\nabla h(z_i) = a^T + z_i A$, the right-hand side of (13.16)

$$\begin{aligned}
&= 1/2\ [(a^T + z_1^T A) + (a^T + z_0^T A)] \cdot (z_1 - z_0) \\
&= a^T(z_1 - z_0) + 1/2\ \{z_1^T A z_1 + z_1^T A z_0 - z_0^T A z_1 - z_0^T A z_0\} \\
&= (a^T z_1 + 1/2\ z_1^T A z_1) - (a^T z_0 + 1/2\ z_0^T A z_0) \\
&= h(z_1) - h(z_0). \square
\end{aligned}$$

Theorem 13.16. *If $\Phi(\cdot)$ is homothetic and Assumptions 4 and 7 hold, then $P_T = P_K$.*

Proof. Let $z_i = z(p^i)$, $i = 1, 2$. The following string of identities follow from Lemma 13.15, identity (13.15) and Shephard's Lemma (5.4.2), namely, $\nabla_p Q(1,p^i) = x^i$:

$$\begin{aligned}
\ln Q(1,p^1) - \ln Q(1,p^0) &= h(z_1) - h(z_0) \\
&= 1/2\ [\ \nabla h(z_1) + \nabla h(z_0)\ ] \cdot (z_1 - z_0) \\
&= 1/2\ \sum_k \left[\ \frac{\partial h(z_1)}{\partial z_k} + \frac{\partial h(z_0)}{\partial z_k}\ \right](z_1^k - z_0^k)
\end{aligned}$$

$$
\begin{aligned}
&= 1/2 \sum_k \left[\frac{p_k^1 \frac{\partial Q(1,p^1)}{\partial p_k}}{Q(1,p^1)} + \frac{p_k^0 \frac{\partial Q(1,p^0)}{\partial p_k}}{Q(1,p^0)} \right] \left(\ln \frac{p_k^1}{p_k^0} \right) \\
&= 1/2 \sum_k \left[\frac{p_k^1 x_k^1}{Q(1,p^1)} + \frac{p_k^0 x_k^0}{Q(1,p^0)} \right] \left(\ln \frac{p_k^1}{p_k^0} \right) \\
&= 1/2 \sum_k (S_k^1 + S_k^1) \left(\ln \frac{p_k^1}{p_k^0} \right) .
\end{aligned}
$$

Taking the exponential of each side proves the claim. □

13.4 Implicit Quantity Indexes

As mentioned in the motivating example, price indexes can be used to construct *implicit quantity indexes*, which we denote by $\hat{Q}$. The rule is this: *given* the price index P its associated implicit quantity index is defined via the identity $P \cdot \hat{Q} = E^1/E^0$. Examples include:

$$
\begin{aligned}
\textit{implicit Laspeyres quantity index} = \hat{Q}_L &:= \frac{E^1/E^0}{P_L}. \\
\textit{implicit Paasche quantity index} = \hat{Q}_P &:= \frac{E^1/E^0}{P_P}. \\
\textit{implicit Fisher ideal quantity index} = \hat{Q}_F &:= \frac{E^1/E^0}{P_F}. \\
\textit{implicit Tornqvist quantity index} = \hat{Q}_T &:= \frac{E^1/E^0}{P_T}.
\end{aligned}
$$

13.5 Quantity Indexes

Analogous quantity indexes can be constructed by simply interchanging the roles of p and x in the formulae presented. Examples include:

$$
\begin{aligned}
\textit{Laspeyres quantity index} = Q_L &:= \frac{x^1 \cdot p^0}{x^0 \cdot p^0}. \\
\textit{Paasche quantity index} = Q_P &:= \frac{x^1 \cdot p^1}{x^0 \cdot p^1}. \\
\textit{Fisher ideal quantity index} = Q_F &:= \sqrt{Q_L Q_P} = \sqrt{\frac{x^1 \cdot p^0}{x^0 \cdot p^0} \frac{x^1 \cdot p^1}{x^0 \cdot p^1}} . \\
\textit{Tornqvist quantity index} = Q_T &:= \prod_i \left(\frac{x_i^1}{x_i^0} \right)^{(S_i^1 + S_i^0)/2} .
\end{aligned}
$$

Example 13.17. For the data provided in the motivating example, we have

$$\begin{aligned}
Q_L &= \frac{1.00(10) + 2.00(4) + 1.50(5)}{1.00(2) + 2.00(3) + 1.50(8)} = \frac{25.50}{20.00} = 1.2750, \\
Q_P &= \frac{0.70(10) + 2.25(4) + 1.80(5)}{0.70(2) + 2.25(3) + 1.80(8)} = \frac{25.00}{22.55} = 1.1086, \\
Q_F &= \sqrt{Q_L Q_P} = \sqrt{(1.2750)(1.1086)} = 1.1889, \\
Q_T &= \left(\frac{10}{2}\right)^{0.19} \left(\frac{4}{3}\right)^{0.33} \left(\frac{5}{8}\right)^{0.48} = 1.1914.
\end{aligned}$$

Remark 13.18. For the example data, you will note a relationship between the implicit Laspeyres, Paasche, and Fisher ideal quantity indexes and the (direct) Laspeyres, Paasche, and Fisher ideal quantity indexes. You will be asked to show in an exercise that $\hat{Q}_L = Q_P$, $\hat{Q}_P = Q_L$ and $\hat{Q}_F = Q_F$.

13.6 Implicit Price Indexes

The quantity indexes can be used to construct *implicit price indexes*, which we denote by $\hat{P}$. The rule is this: *given* the quantity index Q its associated implicit quantity index is defined via the identity $\hat{P} \cdot Q = E^1/E^0$. Examples include:

$$\begin{aligned}
\textit{implicit Laspeyres price index} = \hat{P}_L &:= \frac{E^1/E^0}{Q_L}. \\
\textit{implicit Paasche price index} = \hat{P}_P &:= \frac{E^1/E^0}{Q_P}. \\
\textit{implicit Fisher ideal price index} = \hat{P}_F &:= \frac{E^1/E^0}{Q_F}. \\
\textit{implicit Tornqvist price index} = \hat{P}_T &:= \frac{E^1/E^0}{Q_T}.
\end{aligned}$$

Example 13.19. For the data in the motivating example, we have

$$\begin{aligned}
\hat{P}_L &= \frac{25/20}{1.2750} = 0.9804, \\
\hat{P}_P &= \frac{25/20}{1.1086} = 1.1275, \\
\hat{P}_F &= \frac{25/20}{1.1889} = 1.0514, \\
\hat{P}_T &= \frac{25/20}{1.1914} = 1.0492.
\end{aligned}$$

Remark 13.20. For the example data you will note a relationship between the implicit Laspeyres, Paasche, and Fisher ideal price indexes and the (direct) Laspeyres, Paasche, and Fisher ideal price indexes. You will be asked to show in an exercise that $\hat{P}_L = P_P$, $\hat{P}_P = P_L$ and $\hat{P}_F = P_F$.

13.7 Exercises

Exercises 13.1-13.7 use the price-quantity data given in the Table 13.2:

Table 13.2. Price-quantity data for Exercises 13.1–13.7.

	Period 0		Period 1	
	Price	Quantity	Price	Quantity
Good 1	2.50	40	2.25	50
Good 2	1.50	60	1.65	54
Good 3	0.20	150	0.21	140
Good 4	3.20	50	2.80	80

13.1. Determine the following price indexes:

(a) Ratio of the average price levels.
(b) Average of the price ratios.
(c) Geometric average of the price ratios.
(d) Laspeyres price index.
(e) Paasche price index.
(f) Fisher ideal price index.
(g) Tornqvist price index.

13.2. Determine the following implicit quantity indexes:

(a) Implicit Laspeyres quantity index.
(b) Implicit Paasche quantity index.
(c) Implicit Fisher ideal quantity index.
(d) Implicit Tornqvist quantity index.

13.3. Determine the appropriate bounds on the Konus price index:

(a) When there are no assumptions.
(b) $P_K(p^0, p^1, x^i)$ when each x^i, $i = 0, 1$, is an expenditure (or cost) minimizer.
(c) When the utility (or production) function is homothetic.

13.4. Determine the following quantity indexes:

(a) Ratio of the average quantity levels.

(b) Average of the quantity ratios.
(c) Geometric average of the quantity ratios.
(d) Laspeyres quantity index.
(e) Paasche quantity index.
(f) Fisher ideal quantity index.
(g) Tornqvist quantity index.

13.5. Determine the following implicit price indexes:

(a) Implicit Laspeyres price index.
(b) Implicit Paasche price index.
(c) Implicit Fisher ideal price index.
(d) Implicit Tornqvist price index.

13.6. Compare your answers to Exercises 13.1 and 13.5. Show that:

(a) $P_L = \hat{P}_P$.
(b) $P_P = \hat{P}_L$.
(c) $P_F = \hat{P}_F$.

13.7. Compare your answers to Exercises 13.2 and 13.4. Show that:

(a) $Q_L = \hat{Q}_P$.
(b) $Q_P = \hat{Q}_L$.
(c) $Q_F = \hat{Q}_F$.

13.8. Suppose consumer utility (producer output) is $\Phi(x_1, x_2) = \sqrt{x_1} + x_2$. Further suppose that base prices $p^0 = (1, 2)$ and period 1 prices $p^1 = (2, 1)$, and that the consumer/producer spent 10 to achieve maximum utility/output.

(a) Determine the Konus cost-of-living index $P_K(p^0, p^1, u)$ as a general function of u.
(b) Determine the value of the Konus cost-of-living index when $u = u_0$.
(c) Graphically depict the value of the Konus cost-of-living index as a function of the utility/output u. Discuss its properties and explain its shape.

13.9. Given appropriate data satisfying a number of reasonable assumptions it is possible to estimate the Konus cost-of-living index $P_K(p_0, p_1, \hat{x})$ via a nonparametric approach. This problem suggests how to undertake this task. Suppose data $\mathsf{D} = \{(p_t, x_t)\}_{t=1}^N$ for the observed price and input vectors has been collected over N periods. The data satisfies *producer maximization*, namely, in each period the producer maximized his output subject to his budget constraint. The expenditure in each period is $p_t \cdot x_t$. The production function $\Phi(\cdot)$ is assumed increasing and strictly quasiconcave. (Strict quasiconcavity ensures that the output maximizing input vector is unique.) In what follows, assume $\hat{x}$ does not coincide with any observed data.

(a) Let p and z denote generic observed price and input vectors chosen in a particular period (assumed to satisfy output maximization). The vector z is *revealed preferred* to vector $y \neq z$ if $p \cdot y \leq p \cdot z$. Explain why $\Phi(z) \geq \Phi(y)$.

(b) Use your answer to (a) to identify which observed consumption vectors x_t, $t = 1, 2, \ldots, N$, are revealed preferred to the vector $\hat{x}$.
(c) Use your answer to (b) to explicitly define an approximation $\mathcal{L}(\mathsf{D}, \hat{x})$ to $L^{\geq}_{\Phi}(\Phi(\hat{x}))$.
(d) Use your answer to (c) to estimate $P_K(p_0, p_1, \hat{x})$. Provide a specific formula in terms of the data.

13.8 Bibliographical Notes

The material in this chapter is drawn from several chapters in two monographs, Diewert and Nakamura [1993] and Diewert and Montmarquette [1983]. In particular, Diewert [1987] contains thorough references to the early literature on this subject. See also the original works of Konus [1939], Malmquist [1953]. For practical issues concerning the consumer price index, see the Boskin Commission Report [1996] and Gordon [2000].

13.9 Solutions to Exercises

13.1 (a) $\frac{6.91/4}{7.4/4} = 0.9338$.

(b) $(1/4)(2.25/2.50 + 1.65/1.50 + 0.21/0.20 + 2.8/3.2) = 0.9813$.

(c) $[(0.9)(1.10)(1.05)(0.875)]^{1/4} = 0.9766$.

(d) $P_L = [2.25(40) + 1.65(60) + 0.21(150) + 2.80(50)]/380 = 0.9487$.

(e) $P_P = 455/[2.50(50) + 1.50(54) + 0.20(140) + 3.20(80)] = 0.9286$.

(f) $P_F = \sqrt{(0.9487)(0.9286)} = 0.9386$.

(g) The average expenditure shares are $(1/2)(100/380 + 112.5/455) = 0.2552$, $(1/2)(90/380 + 89.1/455) = 0.2163$, $(1/2)(30/380 + 29.4/455) = 0.0718$, and $(1/2)(160/380 + 224/455) = 0.4567$, respectively. Thus, $P_T = (0.90)^{0.2552}(1.1)^{0.2163}(1.05)^{0.0718}(0.875)^{0.4567} = 0.9382$.

13.2 Here $E^1/E^0 = 455/380 = 1.1974$.

(a) $\hat{Q}_L = 1.1974/0.9487 = 1.2621$.

(b) $\hat{Q}_P = 1.1974/0.9286 = 1.2895$.

(c) $\hat{Q}_F = 1.1974/0.9386 = 1.2757$.

(d) $\hat{Q}_T = 1.1974/0.9382 = 1.2763$.

13.3 $\min\{0.9, 1.1, 1.05, 0.875\} = 0.875$ and $\max\{0.9, 1.1, 1.05, 0.875\} = 1.1$. Moreover, $P_L = 0.9487$ and $P_P = 0.9286$.

(a) $0.875 \leq P_K \leq 1.1$.

(b) $0.875 \leq P_K(p^0, p^1, x^0) \leq 0.9487$ and $0.9286 \leq P_K(p^0, p^1, x^1) \leq 1.1$.

(c) $0.9286 \leq P_K \leq 0.9487$.

13.4 (a) $\frac{(50+54+140+80)/4}{(40+60+150+50)/4} = 1.08$.

(b) $(1/4)\{50/40 + 54/60 + 140/150 + 80/50\} = 1.1708$.

(c) $[(50/40)(54/60)(140/150)(80/50)]^{1/4} = 1.1385$.

(d) $Q_L = [2.50(50) + 1.50(54) + 0.20(140) + 3.20(80)]/380 = 1.2895$.

(e) $Q_P = 455/[2.25(40) + 1.65(60) + 0.21(150) + 2.80(50)] = 1.2621$.

(f) $Q_F = \sqrt{(1.2895)(1.2621)} = 1.2762$.

(g) $Q_T = (50/40)^{0.2552}(54/60)^{0.2163}(140/150)^{0.0718}(80/50)^{0.4567} = 1.2762$.

13.5 (a) $\hat{P}_L = 1.1974/1.2895 = 0.9286$.

(b) $\hat{P}_P = 1.1974/1.2621 = 0.9487$.

(c) $\hat{P}_F = 1.1974/1.2757 = 0.9386$.

(d) $\hat{P}_T = 1.1974/1.2762 = 0.9383$.

13.6 (a)

$$\hat{P}_P = \frac{p^1 \cdot x^1/p^0 \cdot x^0}{p^1 \cdot x^1/p^1 \cdot x^0} = \frac{p^1 \cdot x^0}{p^0 \cdot x^0} = P_L.$$

(b)

$$\hat{P}_L = \frac{p^1 \cdot x^1 / p^0 \cdot x^0}{p^0 \cdot x^1 / p^0 \cdot x^0} = \frac{p^1 \cdot x^1}{p^0 \cdot x^1} = P_P.$$

(c)

$$\hat{P}_F = \frac{p^1 \cdot x^1}{p^0 \cdot x^0} \sqrt{\frac{p^0 \cdot x^0}{p^0 \cdot x^1} \frac{p^1 \cdot x^0}{p^1 \cdot x^1}} = P_F.$$

13.7 (a)

$$\hat{Q}_P = \frac{p^1 \cdot x^1 / p^0 \cdot x^0}{p^1 \cdot x^1 / p^0 \cdot x^1} = \frac{p^0 \cdot x^1}{p^0 \cdot x^0} = Q_L.$$

(b)

$$\hat{Q}_L = \frac{p^1 \cdot x^1 / p^0 \cdot x^0}{p^1 \cdot x^0 / p^0 \cdot x^0} = \frac{p^1 \cdot x^1}{p^1 \cdot x^0} = Q_P.$$

(c)

$$\hat{Q}_F = \frac{p^1 \cdot x^1}{p^0 \cdot x^0} \sqrt{\frac{p^0 \cdot x^0}{p^1 \cdot x^1} \frac{p^0 \cdot x^1}{p^1 \cdot x^1}} = Q_F.$$

13.8 (a) In general,

$$Q(u,p) = \min\{p_1 x_1 + p_2 x_2 : \sqrt{x_1} + x_2 \geq u\}.$$

First, we assume that both x_1 and x_2 will be positive in the optimal solution. If so, then first-order optimality conditions imply that

$$\begin{aligned} p_1 &= \lambda \frac{1}{2\sqrt{x_1}}, \\ p_2 &= \lambda, \end{aligned}$$

which yields $x_1 = p_2^2/4p_1^2$. Using the fact that $\Phi(x) = u$, this in turn implies that $x_2 = u - p_2/2p_1$. This solution only makes sense if $u \geq p_2/2p_1$. If $u \leq p_2/2p_1$, then $x_2 = 0$ and $x_1 = u^2$. (Due to the square root, x_1 will always be positive in the solution since $\sqrt{\epsilon} >> \epsilon$ so that $p_1/\sqrt{\epsilon} < p_2/\epsilon$ for sufficiently small ϵ.) We conclude that

$$\begin{aligned} Q(u,p) &= \begin{cases} \frac{p_2^2}{4p_1} + p_2(u - \frac{p_2}{2p_1}), & \text{if } u \geq \frac{p_2}{2p_1}, \\ p_1 u^2, & \text{otherwise,} \end{cases} \\ &= \begin{cases} p_2 u - \frac{p_2^2}{4p_1}, & \text{if } u \geq \frac{p_2}{2p_1}, \\ p_1 u^2, & \text{otherwise.} \end{cases} \end{aligned}$$

Now substituting in p^0 and p^1 for p we obtain that

$$Q(u,p^0) = \begin{cases} 2u-1, & \text{if } u \geq 1, \\ u^2, & \text{otherwise,} \end{cases}$$

$$Q(u,p^1) = \begin{cases} u-0.125, & \text{if } u \geq 0.25, \\ 2u^2, & \text{otherwise,} \end{cases}$$

from which the Konus cost-of-living index is $Q(u,p^1)/Q(u,p^0)$.

(b) Since the minimum cost in the base period is 10, it follows from the expression for $Q(u,p^0)$ that $u = 5.5$. Substituting this value for u in the expression for $Q(u,p^1)$, we obtain that $Q(u,p^1) = 5.375$. Hence the Konus cost-of-living index is $5.375/10 = 0.5375$.

(c) An examination of the functional form for the cost functions show that the Konus cost-of-living index is indeed a function of the choice of u. This is not too surprising as the cost function does not factor (the function $\Phi(\cdot)$ is not homothetic). Explicitly, $P_K(p^0,p^1,u) = 2$ on $(0,0.25]$, $(u-0.125)/u^2$ on $[0.25,1]$ and $(u-0.125)/(2u-1)$ on $[1,\infty)$. It may be verified that this function of u is strictly decreasing, differentiable and convex whose limit value is 0.5. In particular, we see that for very low values of u, the cost-of-living, as measured by this index, is high. Again, this is because to achieve a very low utility only good 1 will be purchased. The fact that its price has doubled is therefore bad.

13.9 (a) By assumption $\Phi(z) = \Gamma(p, p\cdot z)$. Moreover, since $p\cdot y \leq p\cdot z$ it follows that the input vector y is budget-feasible for the optimization problem defined by $\Gamma(p, p\cdot z)$. Consequently, it must be the case that $\Phi(z) \geq \Phi(y)$; otherwise, z would not be an optimal solution.

(b) By part (a) it is only necessary to check if $p_t \cdot \hat{x} \leq p_t \cdot x_t$; if so, then $\Phi(\hat{x}) \leq \Phi(x_t)$ and x_t would be revealed preferred to $\hat{x}$.

(c) The most obvious choice to approximate $L^{\geq}_{\Phi}(\Phi(\hat{x}))$ is to take $\mathcal{L}(\mathcal{D},\hat{x})$ to be the convex, input free disposable hull of those x_t for which $p_t \cdot \hat{x} \leq p_t \cdot x_t$.

(d) As discussed in Chapter 5, for any choice of price vector p,

$$Q(p) := \min\{p\cdot x :\ x \in \mathcal{L}(\mathcal{D},\hat{x})\} = \min\{p\cdot x_t :\ x_t \in \mathcal{L}(\mathcal{D},\hat{x})\}.$$

Accordingly, we estimate the Konus cost-of-living index as

$$P_K(p^0,p^1,\hat{x}) \approx \frac{Q(p^1)}{Q(p^0)} = \frac{\min\{p^1\cdot x_t :\ x_t \in \mathcal{L}(\mathcal{D},\hat{x})\}}{\min\{p^0\cdot x_t :\ x_t \in \mathcal{L}(\mathcal{D},\hat{x})\}}.$$

14

Productivity Measurement

To estimate productivity growth, the production function $\Phi(\cdot)$ now incorporates a time dimension, as follows. Given the input vector $x(t) = (x_1(t), \ldots, x_n(t))$ chosen at time t, the realized output $y(t)$ at time t equals $\Phi(x(t), t)$. In this chapter, we make the following assumption:

Assumption 8 *The production function is*

$$y(t) \;=\; \Phi(x(t), t) \;=\; A(t)\phi(x(t)), \tag{14.1}$$

where $A(\cdot)$ and $\phi(\cdot)$ are both continuously differentiable and $\phi(\cdot)$ is linearly homogeneous.[1]

Suppose $A(\cdot)$ is increasing with time. If $x(t) = x$ is constant, then output increases without increasing inputs. Alternatively, if a desired output level $y(t) = y$ is constant, then fewer inputs are required to achieve it. From this perspective, $A(t)$ represents the technical progress of the core technology $\phi(\cdot)$ through time. *Productivity growth* is defined as the (instantaneous) growth rate of $A(\cdot)$. It will be measured as the difference between the growth rate in output and a weighted average of the growth rates of the factor inputs. This difference is sometimes referred to as the *Solow residual*, so named after Robert M. Solow, who pioneered the theoretical framework used to measure productivity.

Before we proceed to show how to measure productivity growth, we establish a few basic properties about growth rates. All functions examined hereafter are assumed to be continuously differentiable.

14.1 Growth Rates

Definition 14.1. *The* **growth rate** *of a real-valued function of one variable $z(\cdot)$ from time t to time $t + \Delta t$ is the ratio*

[1] Analysis of productivity growth in a general case is developed in the exercises.

$$\frac{z(t+\Delta t)-z(t)}{\Delta t\ z(t)}\ .$$

The instantaneous growth rate of $z(\cdot)$ at time t is

$$\gamma_z(t)\ :=\ \lim_{\Delta t\to 0}\frac{z(t+\Delta t)-z(t)}{\Delta t\ z(t)}\ .$$

It can be positive or negative. The instantaneous growth rate equals

$$\gamma_z(t)=\frac{d}{dt}\ln z(t)=\frac{z'(t)}{z(t)}=\frac{\dot z}{z}\ , \tag{14.2}$$

where we use $\dot z$ to represent the time derivative of $z(\cdot)$ and suppress the functional dependence on t. For convenience, we shall drop the modifier instantaneous when referring to a function's growth rate.

Example 14.2. Suppose a dollar at time 0 is continuously compounded at the fixed, annual interest rate of r, and let $z(t)$ represent the amount of money at time t. Here, $z(t)=e^{rt}$ and $\gamma_z(t)=r$ for all t. More generally, if interest is accruing at a rate of $r(\tau)$ at time $\tau\in[0,t]$, then

$$z(t)\ =\ e^{\int_0^t r(\tau)d\tau},$$

and $\gamma_z(t)=r(t)$ for all t.

It follows from (14.2) that $\gamma_{f*g}=\gamma_f+\gamma_g$, $\gamma_{f\div g}=\gamma_f-\gamma_g$, and $\gamma_{f^\alpha}=\alpha\gamma_f$. More generally, consider the growth rate of the function

$$\theta(t)\ :=\ h(x_1(t),\ldots,x_n(t)).$$

By direct calculation,

$$\begin{aligned}\gamma_\theta(t)&=\frac{d}{dt}\ln h(x_1(t),\ldots,x_n(t)))\\&=\frac{\sum_i\frac{\partial h}{\partial x_i}\ \dot x_i}{h(x)}\\&=\sum_i\left[\frac{x_i\frac{\partial h}{\partial x_i}}{h(x)}\right]\ \frac{\dot x_i}{x_i}\\&=\sum_i\epsilon_i^h\gamma_i,\end{aligned}\tag{14.3}$$

where ϵ_i^h denotes the elasticity of $h(\cdot)$ with respect to x_i. Identity (14.3) shows that the growth rate of $\theta(\cdot)$ is the weighted sum of the growth rates of each $x_i(\cdot)$ with the weights given by the elasticities.

Example 14.3. When $h(x_1,x_2)=x_1*x_2$, then $\epsilon_i^h=1$, $i=1,2$, and when $h(x_1,x_2)=x_1\div x_2$, then $\epsilon_1^h=1$ and $\epsilon_2^h=-1$, as it must.

14.2 Growth Accounting Approach

It follows from (14.1) and (14.3) that

$$\gamma_y \;=\; \gamma_A \;+\; \sum_i \epsilon_i^{\phi} \gamma_{x_i} \;. \tag{14.4}$$

This immediately implies that productivity growth

$$\gamma_A \;=\; \gamma_y \;-\; \sum_i \epsilon_i^{\phi} \gamma_{x_i} \tag{14.5}$$

can be measured as a residual, as we originally claimed. It remains to estimate the growth rates of the inputs and output from data and to pin down the elasticities, to which we now turn.

Suppose at each time t a firm can sell all the output it can make for R_t per unit, and it faces factor prices p_t. It seeks to choose factor inputs x_t to maximize economic profit given by

$$\max_x \{R\left[A\phi(x)\right] \;-\; p \cdot x\}, \tag{14.6}$$

where for notational convenience we suppress the functional dependence on time. First-order optimality conditions imply that

$$p_i \;=\; RA \frac{\partial \phi}{\partial x_i} \;. \tag{14.7}$$

Since $\phi(\cdot)$ is linearly homogeneous, the economic profit must be zero if a maximum is to exist. (A reasonable return to capital is included in the firm's cost so the firm makes profit in the usual business sense.) This fact together with (14.7) imply that

$$\epsilon_i^{\phi} \;=\; \frac{x_i \, \frac{\partial \phi}{\partial x_i}}{\phi(x)} \;=\; \frac{p_i x_i}{R\left[A\phi(x)\right]} \;=\; \frac{p_i x_i}{p \cdot x} \;:=\; S_i, \tag{14.8}$$

where S_i denotes, as before, the cost share of input i, the proportion of total cost attributable to the i^{th} factor of production. Substituting (14.8) into (14.5) (and bringing back the time index), we have

$$\frac{\dot{A}(t)}{A(t)} \;=\; \frac{\dot{y}(t)}{y(t)} \;-\; \sum_i S_i(t) \frac{\dot{x}_i(t)}{x_i(t)} \;. \tag{14.9}$$

The weighted average of the growth rates of the factor inputs in (14.9) is a *convex combination* since the cost shares are non-negative and sum to one.

Let us examine (14.9) more closely. First, suppose there is only one factor of production. Then, the productivity growth rate is simply the growth rate of the ratio of output to input, which is commonly referred to as a *partial*

productivity measure. The most common partial productivity measure is labor productivity, the growth in the ratio of output per person-hour. When there are several factors of production, then an increase in one partial productivity measure could be the result of a decrease in another partial productivity measure. For example, output per person-hour could increase because the firm purchased significant increases in capital that was used to substitute for the labor input. To address these concerns, *total factor productivity* measures, such as (14.9), attempt to consider all relevant (variable) factors of production.

Suppose data on inputs, outputs, and cost shares from two distinct periods, labeled $t = 0$ (the base period) and $t = 1$, have been collected. We shall assume the cost share for factor i is constant over the time interval $[0, 1]$ so that $S_i(t) = S_i$ for all $t \in [0, 1]$. Integrating both sides of (14.9) from 0 to 1,

$$\ln \frac{A(1)}{A(0)} \;=\; \ln \frac{y(1)}{y(0)} \;-\; \sum_i S_i \ln \frac{x_i(1)}{x_i(0)} \,,$$

or, equivalently,

$$\frac{A(1)}{A(0)} \;=\; \frac{y(1)}{y(0)} \prod_i \left(\frac{x_i(0)}{x_i(1)} \right)^{S_i} . \tag{14.10}$$

Given the present index of productivity, $A(0)$, the new index is obtained via (14.10). The cost shares need not be constant over the time frame, and so one often substitutes the average $(S_i(0) + S_i(1))/2$ for each S_i.

Example 14.4. Suppose at time 0 a firm used 100 units of capital and 100 units of labor to produce 1100 units of output. At time 1 the firm used 106 units of capital and 91 units of labor to produce 1133 units of output. The cost share of labor is two-thirds for both periods. (The price of labor relative to capital significantly increased from period 0 to period 1.) The growth in the index of productivity is

$$\frac{1133}{1100} \Big[\Big(\frac{100}{106} \Big)^{1/3} \Big(\frac{100}{91} \Big)^{2/3} \Big] = 1.0757.$$

When the percentage changes in the inputs or output are small, there is a back-of-the-envelope way to accurately estimate the growth in this index of productivity. Observe that output increased by 3%, the capital input increased by 6% and the labor input decreased by 9%. Given the cost shares, the percentage growth in the aggregate input is approximately $(+6)(1/3) + (-9)(2/3) = -4$. Since output increased by 3 percent, the overall percentage growth in productivity is approximately $3 - (-4) = 7$. (The accuracy will be better when the percentage changes in input and output are closer to zero.)

14.3 Multi-Output Productivity Measurement

The Tornqvist index can be used to develop an index of productivity when there are multiple outputs (as well as multiple inputs). In this section, we make the following assumption:

Assumption 9 *The technology $\mathcal{T}^t$ at time t is characterized by two index functions, one with respect to input and the other with respect to output, such that*

$$\mathcal{T}^t = \{(x, y) : \ g(y) \leq A(t)f(x)\},$$

where $f(\cdot)$ and $g(\cdot)$ are linearly homogeneous, translog functions with $f(\cdot)$ concave and $g(\cdot)$ convex.

We seek to measure A^1/A^0.

We assume firms are profit maximizers. Let w^t and p^t denote, respectively, the vector of prices on outputs and inputs in period $t = 0, 1$. In each period $t = 0, 1$, the firm solves the following optimization problem:

$$\max\{w^t \cdot y^t - p^t \cdot x^t : \ g(y^t) \leq A^t f(x^t)\}. \tag{14.11}$$

The proof of Theorem 13.16, p. 232, shows that

$$\ln \frac{f(x^1)}{f(x^0)} = 1/2 \sum_k \left[\frac{x_k^1 \frac{\partial f(x^1)}{\partial x_k}}{f(x^1)} + \frac{x_k^0 \frac{\partial f(x^0)}{\partial x_k}}{f(x^0)} \right] \left(\ln \frac{x_k^1}{x_k^0} \right), \tag{14.12}$$

$$\ln \frac{g(y^1)}{g(y^0)} = 1/2 \sum_k \left[\frac{y_k^1 \frac{\partial g(y^1)}{\partial y_k}}{g(y^1)} + \frac{y_k^0 \frac{\partial g(y^0)}{\partial y_k}}{g(y^0)} \right] \left(\ln \frac{y_k^1}{y_k^0} \right). \tag{14.13}$$

Profit maximization implies both cost minimization in x for fixed y and revenue maximization in y for fixed x. First, consider cost minimization. Think of y^i as fixed and $\Phi(x) = af(x)$. We have previously shown in (14.8) that the elasticity of $\Phi(\cdot)$ with respect to x_i equals the cost share for constant returns-to-scale technologies. Since the elasticity of $\Phi(\cdot)$ equals the elasticity of $f(\cdot)$, it follows from (14.12) that

$$\ln \frac{f(x^1)}{f(x^0)} = 1/2 \sum_k (S_k^1 + S_k^0) \left(\ln \frac{x_k^1}{x_k^0} \right). \tag{14.14}$$

Using an exactly analogous argument applied to revenue maximization, we have

$$\ln \frac{g(y^1)}{g(y^0)} = 1/2 \sum_k (R_k^1 + R_k^0) \left(\ln \frac{y_k^1}{y_k^0} \right), \tag{14.15}$$

where R_k^t denotes the revenue share of output k in period $t = 0, 1$. Since both $f(\cdot)$ and $g(\cdot)$ are homogeneous of degree one, it must be the case that $g(y^t) = A^t f(x^t)$ for each $t = 0, 1$. It directly follows that

$$\frac{A(1)}{A(0)} = \frac{g(y^1)/g(y^0)}{f(x^1)/f(x^0)} = \frac{\prod_k (\frac{y_k^1}{y_k^0})^{(R_k^1+R_k^0)/2}}{\prod_k (\frac{x_k^1}{x_k^0})^{(S_k^1+S_k^0)/2}} \,. \tag{14.16}$$

An extensive example of the use of multi-output productivity measurement is provided in Chapter 15.

14.4 Nonparametric Approach

Measures of productivity introduced so far can be calculated by each firm. They are firm-specific and absolute since there is no comparison of how well the firm is improving its productivity *relative* to other firms. If data from representative firms are available, it is possible to assess changes in productivity and efficiency relative to the industry as a whole.

We describe a nonparametric approach for (i) measuring productivity change between periods t and $t+1$, and (ii) decomposing productivity change into its "technical change" and "efficiency change" components. The measures are defined and computed via distance functions.

14.4.1 Input Productivity Change

Let $\mathcal{T}^t$, $\mathcal{D}^t(x,u)$ and $\mathcal{T}^{t+1}$, $\mathcal{D}^{t+1}(x,u)$ denote, respectively, the technology sets and its associated distance function in periods t and $t+1$. The input possibility sets characterizing these technologies can be defined using any of the nonparametric constructions introduced so far.

Let (x^t, u^t) and (x^{t+1}, u^{t+1}) denote the input-output data for a particular firm in periods t and $t+1$. We begin by showing how to measure the input productivity change due *solely* to the input technical change—that is, we exclude the input productivity change due to the (possible) improvement in input efficiency and concentrate on how to measure the degree to which input possibility sets have improved by "moving closer to the origin." To eliminate the efficiency effect, let

$$\hat{x}^k \;:=\; \frac{x^k}{\mathcal{D}^k(x^k, u^k)},\; k = t, t+1, \tag{14.17}$$

denote the *adjusted* input.

If the input possibility sets have improved from period t to $t+1$, less input will now be required to achieve the same level of output. From the definition of the input distance function, it will then follow that

(i) $\mathcal{D}^{t+1}(\hat{x}^t, u^t) > 1$.
(ii) $\mathcal{D}^t(\hat{x}^{t+1}, u^{t+1}) < 1$.

A measure of input technical change should be *less* than one to indicate a productivity improvement with respect to input.[2] Accordingly either

$$\mathcal{D}^t(\hat{x}^{t+1}, u^{t+1}) \text{ or } \mathcal{D}^t(\hat{x}^{t+1}, u^{t+1})^{-1}$$

could be used as a measure of the input technical change between periods t and $t+1$. In case (i), one is looking "forward in time;" that is, how period t's input-output pair would be assessed from the perspective of period $t+1$'s technology, very much in the spirit of the Laspeyres index discussed in Chapter 13. In case (ii), one is looking "backward in time;" that is, how period $t+1$'s input-output pair would be assessed from the perspective of period t's technology, very much in the spirit of the Paasche index discussed in Chapter 13. Since either measure could be appropriate, in the spirit of the Fisher ideal index, the measure of input technical change between periods t and $t+1$ "splits the difference" and takes the geometric mean

$$\sqrt{\mathcal{D}^t(\hat{x}^{t+1}, u^{t+1})\ \mathcal{D}^{t+1}(\hat{x}^t, u^t)^{-1}} \tag{14.18}$$

of these two measures. Substituting the definitions of $\hat{x}^t$ and $\hat{x}^{t+1}$ in (14.17) into (14.18), and using the fact that $\mathcal{D}(\cdot, u)$ is linearly homogeneous in x, we formally have:

Definition 14.5. *The* **measure of input technical change between periods t and t+1** *is*

$$\sqrt{\frac{\mathcal{D}^t(x^{t+1}, u^{t+1})}{\mathcal{D}^{t+1}(x^{t+1}, u^{t+1})}\ \frac{\mathcal{D}^t(x^t, u^t)}{\mathcal{D}^{t+1}(x^t, u^t)}}\ .$$

Example 14.6. Consider the single-output technology characterized by the production function defined in (14.1). In this special case, the measure of technical change must equal A^t/A^{t+1}. Pick a positive input vector x and a positive output u. By the definition and the linear homogeneity property of the distance function, for the functional form given in (14.1),

$$u = A^k \phi\Big(\frac{x}{\mathcal{D}^k(x,u)}\Big) \Longrightarrow \mathcal{D}^k(x,u) = \frac{A^k \phi(x)}{u},\ k = t, t+1. \tag{14.19}$$

Consequently, $\mathcal{D}^k(x,u)/\mathcal{D}^{k+1}(x,u) = A^t/A^{t+1}$ for each $k = t, t+1$. Thus, *each* measure of input technical change, $\mathcal{D}^t(\hat{x}^{t+1}, u^{t+1})$ or $\mathcal{D}^{t+1}(\hat{x}^t, u^t)^{-1}$, equals A^t/A^{t+1}, and obviously so will their geometric mean.

Let $\mathcal{RI}^k(x^k, u^k)$, $k = t, t+1$, denote the radial measures of input efficiency corresponding to each period (see Definition 9.1, p. 149). With respect to measuring the input productivity change due to the input efficiency change, a natural choice is to take the ratio $\mathcal{RI}^t(x^t, u^t)/\mathcal{RI}^{t+1}(x^{t+1}, u^{t+1})$. If the input

[2] This is a matter of convention.

efficiency has improved, this ratio will be less than one. *A priori*, however, the efficiency change could be less than, equal to, or greater than one. We have previously shown that the radial measure of input efficiency $\mathcal{RI}(x,u)$ is the reciprocal of the distance $\mathcal{D}(x,u)$.

Definition 14.7. *The* **measure of input efficiency change between periods t and t+1** *is*

$$\frac{\mathcal{RI}^t(x^t,u^t)}{\mathcal{RI}^{t+1}(x^{t+1},u^{t+1})} = \frac{\mathcal{D}^{t+1}(x^{t+1},u^{t+1})}{\mathcal{D}^t(x^t,u^t)} \quad .$$

Finally, the measure of the overall productivity change is given as follows:

Definition 14.8. *The* **measure of input productivity change between periods t and t+1** *is the product of the input technical change and the input efficiency change. Substituting the definitions for the input technical and efficiency change, 14.5 and 14.7, respectively, the formula for input productivity change is*

$$\sqrt{\frac{\mathcal{D}^{t+1}(x^{t+1},u^{t+1})}{\mathcal{D}^{t+1}(x^t,u^t)} \; \frac{\mathcal{D}^t(x^{t+1},u^{t+1})}{\mathcal{D}^t(x^t,u^t)}} \quad .$$

Remark 14.9. By definition, input productivity change is decomposable into its input technical change and input efficiency change components.

14.4.2 Output Productivity Change

In the single-output setting, a similar development undertaken in the previous section can be used to develop an analogous measure of output productivity change, and decompose it into its output technical change and output efficiency change components. We shall use the output distance function $\mathcal{O}(u,x)$ (see Definition 7.3, p. 110). Recall that the scaled output $u/\mathcal{O}(x,u)$ represents the maximum output the input vector x can achieve.

As before, we begin by showing how to measure the output productivity change due *solely* to the output technical change—that is, we exclude the output productivity change due to the (possible) improvement in output efficiency and concentrate on how to measure the degree to which output possibility sets have improved by "moving away from the origin." To eliminate the efficiency effect, let

$$\hat{u}^k \;:=\; \frac{u^k}{\mathcal{O}^k(x^k,u^k)},\; k=t,t+1, \tag{14.20}$$

denote the *adjusted* output.

If the output possibility sets have improved from period t to $t+1$, more output will be produced for the same level of input. From the definition of the output distance function, it will then follow that

(i) $\mathcal{O}^{t+1}(x^t, \hat{u}^t) < 1.$
(ii) $\mathcal{O}^t(x^{t+1}, \hat{u}^{t+1}) > 1.$

A measure of output technical change should be *greater* than one to indicate a productivity improvement with respect to output.[3] Accordingly either

$$\mathcal{O}^t(x^{t+1}, \hat{u}^{t+1}) \text{ or } \mathcal{O}^{t+1}(x^t, \hat{u}^t)^{-1}$$

could be used as a measure of the output technical change between periods t and $t+1$ depending on whether one is looking forward in time or backwards in time. Once again, since either measure could be appropriate, in the spirit of the Fisher ideal index, the measure of output technical change between periods t and $t+1$ "splits the difference" and takes the geometric mean

$$\sqrt{\mathcal{O}^t(\hat{x}^{t+1}, u^{t+1})\, \mathcal{O}^{t+1}(\hat{x}^t, u^t)^{-1}} \tag{14.21}$$

of these two measures. Substituting the definitions of $\hat{u}^t$ and $\hat{u}^{t+1}$ in (14.20) into (14.21), and using the fact that $\mathcal{O}(x, \cdot)$ is linearly homogeneous in u, we formally have:

Definition 14.10. *The* **measure of output technical change between periods t and t+1** *is*

$$\sqrt{\frac{\mathcal{O}^t(x^{t+1}, u^{t+1})}{\mathcal{O}^{t+1}(x^{t+1}, u^{t+1})} \frac{\mathcal{O}^t(x^t, u^t)}{\mathcal{O}^{t+1}(x^t, u^t)}} \ .$$

With respect to measuring the output productivity change due to the output efficiency change, a natural choice is to take the ratio of the output distance in period $t+1$ to the output distance in period t, since the output distance $\mathcal{O}(x, u)$ can be interpreted as the radial measure of output efficiency. If the output efficiency has improved, this ratio will be greater than one. *A priori*, however, the efficiency change could be less than, equal to, or greater than one.

Definition 14.11. *The* **measure of output efficiency change between periods t and t+1** *is*

$$\frac{\mathcal{O}^{t+1}(x^{t+1}, u^{t+1})}{\mathcal{O}^t(x^t, u^t)} \ .$$

Finally, the measure of the overall productivity change is given as follows:

Definition 14.12. *The* **measure of output productivity change between periods t and t+1** *is the product of the output technical change and the output efficiency change. Substituting the definitions for the output technical and efficiency change, (14.21) and (14.11), respectively, the formula for output productivity change is*

$$\sqrt{\frac{\mathcal{O}^{t+1}(x^{t+1}, u^{t+1})}{\mathcal{O}^{t+1}(x^t, u^t)} \frac{\mathcal{O}^t(x^{t+1}, u^{t+1})}{\mathcal{O}^t(x^t, u^t)}} \ .$$

[3] This is a matter of convention.

Remark 14.13. By definition, output productivity change is decomposable into its output technical change and output efficiency change components.

14.5 Exercises

14.1. What was the growth in the price of labor to price of capital ratio from period 0 to period 1 in Example 14.4?

14.2. Assume cost-minimizing behavior on the part of the producer and assume the technology exhibits constant returns-to-scale.

Table 14.1. Data for Exercise 14.2.

Period 0					Period 1				
K	L	p_K	p_L	y	K	L	p_K	p_L	y
60	400	1.5	0.4	2450	65	384	1.6	0.5	2989

(a) Use (14.9) and average cost shares to estimate the productivity growth of the firm for the data provided in Table 14.1.
(b) Use (14.10) and average cost shares to estimate the productivity growth of the firm for the data provided in Table 14.1.
(c) Compute the Laspeyres, Paasche, and Fisher ideal quantity indexes, and use these to provide another explanation for the estimate of productivity growth obtained in parts (a) and (b).

14.3. Use (14.16) to estimate the productivity growth for the data provided in Table 14.2.

Table 14.2. Data for Exercise 14.3.

Period 0								Period 1							
K	L	p_K	p_L	y_1	y_2	R_1	R_2	K	L	p_K	p_L	y_1	y_2	R_1	R_2
305	120	2	3	75	125	8.0	2.96	315	110	2.1	2.7	90	115	7.2	2.7

14.4. A firm's production function is $\Phi(K, L) = AK^{\alpha}L^{1-\alpha}$. The firm is a price-taker. The price of its output is R and the prices of the inputs are p_K and p_L, respectively.

(a) Determine the necessary conditions on the parameters R, A, α, p_K, and p_L to ensure an optimal profit exists.

(b) What are the optimal choices for factor inputs?
(c) Examine the setting when $\alpha = 0.5$, $p_K = 4$ and $p_L = 1$.

14.5. This problem extends the estimation of productivity for a more general production function of the form $y(t) = \Phi(x(t), t)$. Here, the symbol "t" denotes time, which is taken to be a nonnegative real number. Given an input vector $x = x(t)$ the *(primal) rate of technical change* is defined as

$$\tau := \frac{\partial \Phi / \partial t}{\Phi}.$$

We assume the producer is a profit-maximizer, namely, at each instant of time the producer optimizes the profit function

$$\max\{R(t)\Phi(x(t), t) - p(t) \cdot x(t)\},$$

where $R(t)$ is the net revenue per unit of output and $p(t)$ is the vector of prices at time t, which the producer takes as given. (Assume sufficient conditions on $\Phi(\cdot)$ to ensure that a unique solution exists, is differentiable, etc.) The *gross profit margin*, GPM, measures the profit in relation to net revenue, as is defined as

$$GPM := \frac{R(t)\Phi(x(t), t) - p(t) \cdot x(t)}{R(t)\Phi(x(t), t)} = 1 - \frac{p(t) \cdot x(t)}{R(t)\Phi(x(t), t)}.$$

(a) Approximate τ using Δt and interpret this approximation.
(b) Show that when $y(t) = \Phi(x(t), t) = A(t)\phi(x(t))$ that $\tau = \dot{A}/A$.
(c) Use the chain-rule to show that

$$\tau = \frac{\dot{y}}{y} - \sum_i \epsilon_i^{\Phi} \frac{\dot{x}_i}{x_i},$$

where ϵ_i^{Φ} is the elasticity of output with respect to factor i.
(d) Show that

$$\epsilon_i^{\Phi} = (1 - GPM)S_i,$$

where S_i represents the cost share of factor i.
(e) Use parts (c) and (d) to determine a formula for τ, and show that it reduces to the formula given in the text when the production function exhibits constant returns-to-scale at each point in time.

14.6. This problem assumes the setup of Exercise 14.5. Let $Q(y(t), p(t), t)$ denote the minimal cost function assumed to be appropriately differentiable. Given the output rate $y(t)$ and price vector $p(t)$, the *(dual) rate of cost dimunition* is defined as

$$\gamma := \frac{\partial Q / \partial t}{Q}.$$

(a) Approximate γ using Δt and interpret this approximation.

(b) Let ϵ denote the elasticity of cost with respect to output. Use the Envelope Theorem to show that $\gamma = -\epsilon \cdot \tau$, where τ is the rate of technical change.
(c) If one has access to input data and a means to parametrically estimate the production function, one can, in principle, estimate τ via Exercise 14.5. Often, however, input data are not available but price and cost data are. Suppose at each point in time the production function is known to exhibit constant returns-to-scale. Explain how to estimate τ given only price and cost data as well as a means to estimate a parametric form for the cost function.

14.6 Bibliographical Notes

Solow [1957] is the classic reference. Fare et. al. [1994] introduced the nonparametric approach to measuring productivity and decomposing it into its efficiency and technical change components. Their work is based in part on Caves et. al. [1982]. Grosskopf [1993] provides a thorough, accessible treatment of the material pertaining to the nonparametric approach discussed in this chapter.

Consult Christopher and Thor [1993] and P.T. Harker [1995] for discussion of practical issues pertaining to productivity measurement and implementation of productivity measures to a variety of industrial settings.

14.7 Solutions to Exercises

14.1 The total cost in period 0 is $100p_K + 100p_L$ of which 2/3 is due to labor cost. Therefore, $100p_L = 2(100p_K)$ or $p_K/p_L = 2$. The total cost in period 1 is $106p_K + 91p_L$ of which 2/3 is due to labor cost. Therefore, $91p_L = 2(106p_K)$ or $p_K/p_L = 2.3297$. The growth rate in the factor price ratio is thus $(2.3297 - 2)/2 = 16.48\%$.

14.2 (a) We have $\dot{y}/y \approx (2989 - 2450)/2450 = 0.22$, $\dot{K}/K \approx (65 - 60)/60 = 0.08\bar{3}$, and $\dot{L}/L \approx (384 - 400)/400 = -0.04$. The average cost share of capital is $0.5[1.5(60)/(1.5(60) + 0.4(400)) + 1.6(65)/(1.6(65) + 0.5(384))] = 0.3557$. Thus,

$$\dot{A}/A \approx 0.22 - [0.3557(0.08\bar{3}) + 0.6443(-0.04)] = 0.2161,$$

a 21/64% improvement in productivity.

(b) We have

$$\frac{A(1)}{A(0)} = \frac{2989}{2450}\left(\frac{60}{65}\right)^{0.3557}\left(\frac{400}{384}\right)^{0.6443} = 1.2174,$$

which suggests a 21.75% improvement in productivity.

(c) Here, $Q_L = p^0 \cdot x^1/p^0 \cdot x^0 = 251.1/250 = 1.0044$, $Q_P = p^1 \cdot x^1/p^1 \cdot x^0 = 296/296 = 1.000$, and $Q_F = \sqrt{(1.0044)(1.0000)} = 1.0022$. Thus, the quantity index suggests an almost imperceptible increase in input used, yet output grew by 22%. This leaves a productivity gain of 22% as the resulting explanation.

14.3 The total revenue equals total cost = 970 in period 0 and the total revenue equals total cost = 958.5 in period 1. The average revenue share for good 1 is $0.5[8(75)/970 + 7.2(90)/958.5] = 0.6267$ and the average cost share for capital is $0.5[2(305)/970 + 2.1(315)/958.5] = 0.6595$. Thus,

$$\frac{A(1)}{A(0)} = \frac{\left(\frac{90}{75}\right)^{0.6267}\left(\frac{115}{125}\right)^{0.3733}}{\left(\frac{315}{305}\right)^{0.6595}\left(\frac{110}{120}\right)^{0.3405}} = \frac{1.0867}{0.9917} = 1.0958,$$

suggesting a 9.58% productivity gain.

14.4 (a) The firm seeks to maximize its profit rate given by

$$RAK^{\alpha}L^{1-\alpha} - p_K K - p_L L.$$

Regardless of the output rate chosen, the firm must minimize cost to achieve that output rate. For a Cobb-Douglas technology, we know that each cost share equals the ratio of the exponent in the production function corresponding to that factor to the sum of the exponents, i.e., $p_K K/(p_K K + p_L L) = \alpha$. This implies that cost minimizing factor ratio is $L/K = [(1 - \alpha)/\alpha]p_K/p_L$ or that

$$L = \left[\frac{1 - \alpha}{\alpha}\frac{p_K}{p_L}\right]K.$$

Since the profit function is homogeneous of degree one (i.e., exhibits constant returns-to-scale), the optimal profit has to be zero (otherwise the firm could make infinite profit). Substituting the cost minimizing factor ratio identity into the profit equation yields

$$0 = \left(RA\Big[\frac{1-\alpha}{\alpha}\frac{p_K}{p_L}\Big]^{\alpha} - \frac{p_K}{\alpha} \right) K.$$

Consequently, the expression inside the parentheses must be zero.

(b) Assuming the condition of part (a) holds, then any choice of (K, L) that satisfies the cost minimizing factor ratio will achieve an optimal profit of zero. The scale of output is the one remaining decision for the firm.

(c) Here the condition simplifies to $[RA\sqrt{4} - 8] = 0$, which implies that $RA = 4$. The profit function is therefore $4\sqrt{KL} - 4K - L$. The minimum cost factor ratio is $L = 4K$, which results in a profit of zero. Suppose, for example, that $RA = 5$, then the optimal profit when $L = 4K$ is $2K$, which can be made arbitrarily high. On the other hand, if say $RA = 3$, then the optimal profit when $L = 4K$ (best) if $-2K$, and so the best choice is to set $K = L = 0$.

14.5 (a)

$$\tau \approx \frac{\Phi(x(t), t + \Delta t) - \Phi(x(t), t)}{\Phi(x(t), t)}.$$

For a fixed input vector $x = x(t)$, it locally measures the percentage change in output that is attributable to productivity (positive or negative) over time.

(b) Here,

$$\tau = \frac{\dot{A}(t)\phi(x(t))}{A(t)\phi(x(t))} = \frac{\dot{A}}{A}.$$

(c) By the chain rule,

$$\begin{aligned} \frac{\dot{y}}{y} &= \frac{d}{dt} \ln \Phi(x(t), t) = \frac{\sum_i (\partial \Phi / \partial x_i)\dot{x}_i \; + \; \partial \Phi / \partial t}{\Phi(x(t), t)} \\ &= \sum_i \Big(\frac{x_i(\partial \Phi / \partial x_i)}{\Phi} \Big) \frac{\dot{x}_i}{x_i} \; + \; \tau \\ &= \sum_i \epsilon_i^{\Phi} \frac{\dot{x}_i}{x_i} \; + \; \tau. \end{aligned}$$

(d) First-order optimality conditions imply that $R\partial \Phi / \partial x_i = p_i$. Thus, $R(\partial \Phi / \partial x_i)x_i = p_i x_i$, which yields

$$\epsilon_i^{\Phi} = \frac{(\partial \Phi / \partial x_i)x_i}{\Phi} = \frac{p_i x_i}{R\Phi} = (1 - GPM)\frac{p_i x_i}{p \cdot x} = (1 - GPM)S_i.$$

(e) Putting it all together, we have

$$\tau = \frac{\dot{y}}{y} - (1 - GPM) \sum_i S_i \frac{\dot{x}_i}{x_i}.$$

When the production function exhibits constant returns-to-scale, the profit must be zero (otherwise it would be infinite). Consequently, $GPM = 0$, and the formula for τ above reduces to the one developed in the chapter.

14.6 (a)

$$\gamma = \frac{Q(y(t), p(t), t + \Delta t) - Q(y(t), p(t), t)}{Q(y(t), p(t), t)}.$$

For a fixed output rate $y = y(t)$ and price vector $p = p(t)$, it locally measures the percentage change in minimal cost that is attributable to productivity (positive or negative) over time.

(b) Fix $y = y(t) = \Phi(x(t), t)$ and $p = p(t)$. The Lagrangian for the cost minimization problem at time t is

$$L(x, \lambda, p, y, t) = p \cdot x - \lambda(\Phi(x, t) - y).$$

Application of the Envelope Theorem F.2, p. 492, yields

$$\frac{\partial Q}{\partial t} = -\lambda^* \frac{\partial \Phi(x^*, t)}{\partial t},$$
$$\frac{\partial Q}{\partial y} = \lambda^*.$$

Thus,

$$\frac{\partial Q}{\partial t} = -\frac{\partial Q}{\partial y} \frac{\partial \Phi(x^*, t)}{\partial t}.$$

Substituting this identity into the definition of γ and using the fact that $y = \Phi(x^*, t)$, we have

$$\gamma = -\frac{\partial Q/\partial t}{Q} = \left[\frac{y}{Q}\frac{\partial Q}{\partial y}\right] \cdot \left[\frac{\partial \Phi/\partial t}{\Phi}\right] = \epsilon \cdot \tau.$$

(c) When the production function exhibits constant returns-to-scale, $\epsilon = 1$. Thus, given a parametric form for the cost function, one may differentiate it to obtain an estimate of γ and hence τ.

15

Performance Measurement

Index numbers can be used to measure and decompose a firm's performance from one period to the next. We shall illustrate with an example from the furniture industry.

15.1 A Manufacturing Example

Furniture manufacturing is labor intensive. The stages of production typically involve a wood cutting operation to set the pattern, a trimming operation to smooth the edges, an assembly operation to connect the different component pieces of wood, a sanding operation, and a finishing operation to stain the wood. The machinery used include lathes, sanders, and jigsaws. As for the material categories, wood (e.g., cherry, poplar) is obviously needed for both sofas and tables, and sofas need fabric and cushioning. A majority of the production labor is used for the finishing operation, and the finisher commands the highest wage.

Steeples Furniture[1] is a small company that manufactures sofas and tables for the high-end market. The input-output data for the base year ("Year 0") and the subsequent year ("Year 1") are shown in Table 15.1. A few observations are in order. In the base year:

- The sofa product line generated 75% of total revenue.
- Labor cost accounted for 50% of the total cost, with materials and capital accounting for approximately 30% and 20%, respectively.
- Average inventory equates to approximately a five week time supply. Steeples Furniture used a 20% cost per dollar of inventory per year to account for the handling, storage, and financing.
- Steeples Furniture used a 15% cost per dollar of book value of machinery to account for maintenance and economic depreciation.

[1] Not a real firm.

In the subsequent year:

Table 15.1. Input-output price-quantity data.

	Year 0			Year 1		
Output/Input	Value	Quantity	Price	Value	Quantity	Price
OUTPUT						
Sofas	600,000	300	2,000.00	720,800	340	2,120.00
Tables	200,000	80	2,500.00	235,200	84	2,800.00
Total Revenue	800,000			956,000		
INPUT						
Materials						
wood	64,000	4,000	16.00	85,000	4,250	20.00
fabric	105,000	3,500	30.00	126,000	4,000	31.50
cushioning	31,000	12,400	2.50	37,365	14,100	2.65
Total Materials	200,000			248,365		
Labor						
woodcutter	48,000	2,400	20.00	56,175	2,675	21.00
lathe operator	8,000	400	20.00	8,820	420	21.00
assembler	48,000	2,400	20.00	56,175	2,675	21.00
sander	8,100	450	18.00	8978	475	18.90
finisher	207,900	7,425	28.00	273,280	8,540	32.00
Total Labor	320,000			403,428		
Capital						
inventory	15,000	75,000	0.20	7,500	50,000	0.15
machinery	105,000	700,000	0.15	110,500	850,000	0.13
Total Capital	120,000			118,000		
Total Cost	640,000			769,793		
PROFIT	160,000			186,207		

- Average price of wood jumped to $20 per board-feet from $16, an increase of 25%. The average price increase of fabric and cushioning was more modest at 5% and 6%, respectively.
- Shortage of skilled labor drove up the price of labor. Wood cutters, lathe operators, assemblers, and sanders all received a 5% increase in their real (gross) hourly wage. To maintain its reputation for quality, Steeples Furniture wished to retain the services of its finishers, who commanded an approximate 15% increase in their real (gross) hourly wage.
- Due to their excellent reputation for quality, Steeples Furniture was able to increase the price of their sofas and tables by an average of 6% and 12%, respectively, to partially offset the increases in the cost of labor and materials. (The increase in the cost of wood affects the cost of a table far more than the cost of a sofa.) Due to the price increase of wood,

the company chose to aggressively market its sofa product lines in an attempt to partially shift its output mix. Even with the price increases, the company saw an increase of demand of almost 15% for their sofa product lines and 5% for their table product lines.

- Steeples Furniture spent \$255,000 = \$850,000 - (\$700,000)(0.85) to upgrade its machinery. The capital cost was lowered to 13% from 15% due to the improved quality of the machinery.
- Steeples Furniture worked hard to lower its average inventory to less than a three week time supply. With less inventory, there was less storage cost, and coupled with lower financing costs, the company lowered its inventory cost to 15% from 20%.

15.2 Performance Indexes

So how well did Steeples Furniture perform year-to-year? As a first step, Table 15.2 records the performance indexes with respect to value, quantity, and price for the outputs and inputs. Here is how the indexes are computed.

Table 15.2. Performance indexes.

Output/Input	V^1/V^0	Q^1/Q^0	P^1/P^0
OUTPUT			
Sofas	1.2013	1.1333	1.0600
Tables	1.1760	1.0500	1.1200
Total Revenue	1.1950	1.1125	1.0742
INPUT			
Materials			
wood	1.3281	1.0625	1.2500
fabric	1.2000	1.1429	1.0500
cushioning	1.2053	1.1371	1.0600
Total Materials	1.2418	1.1163	1.1125
Labor			
woodcutter	1.1703	1.1146	1.0500
lathe operator	1.1025	1.0500	1.0500
assembler	1.1703	1.1146	1.0500
sander	1.1084	1.0556	1.0500
finisher	1.3145	1.1502	1.1429
Total Labor	1.2607	1.1346	1.1111
Capital			
inventory	0.5000	0.6667	0.7500
machinery	1.0524	1.2143	0.8667
Total Capital	0.9833	1.1458	0.8582
Total Cost	1.2028	1.1310	1.0635

- *Column V^1/V^0.* On the revenue (output) side, this column records the ratio of the revenue in period 1 to the revenue in period 0. For example, total revenue in period 1 was 956,000, it was 800,000 in period 0, and so the ratio is 956,000/800,000 = 1.1950. On the input side, this column records the ratio of the expenditures in period 1 to the expenditures in period 0. For example, for the materials input category, the expenditure in period 1 was 248,365, it was 200,000 in period 0, and so the ratio is 248,365/200,000 = 1.2418.
- *Column Q^1/Q^0.* This column records the *Laspeyres quantity index*,

$$\mathcal{Q}_{\mathcal{L}} := \frac{p^0 \cdot x^1}{p^0 \cdot x^0}, \tag{15.1}$$

which is the ratio of the revenue/cost of an individual output/input or output/input category in period 1 *using period 0's prices* to the revenue/cost of the output/input in period 0. For example, the total revenue in period 1 using period 0's prices is 2,000(340) + 2,500(84) = 890,000, which when divided by 800,000 (the revenue received in period 0) yields 1.1125. For the materials input category,

$$p^0 \cdot x^1 = 16.00(4,250) + 30.00(4,000) + 2.50(14,100) = 223,250,$$

which when divided by $p^0 \cdot x^0 = 200,000$ yields 1.1163.

Remark 15.1. The Laspeyres quantity index associated with a group of categories is the weighted sum of the Laspeyres quantity indexes for each group with the weights corresponding to the cost shares in the base period. For example, the Laspeyres quantity index for total cost is 1.1310, which can be calculated as

$$\left(\frac{200,000}{640,000}\right) 1.1163 + \left(\frac{320,000}{640,000}\right) 1.1346 + \left(\frac{120,000}{640,000}\right) 1.1458.$$

- *Column P^1/P^0.* This column records the *implicit Laspeyres price index*, $\mathcal{P}_{\mathcal{L}}$, which by definition satisfies the following equality:

$$\mathcal{P}_{\mathcal{L}} \cdot \mathcal{Q}_{\mathcal{L}} = \frac{p^1 \cdot x^1}{p^0 \cdot x^0}. \tag{15.2}$$

Since $\mathcal{Q}_{\mathcal{L}} = p^0 \cdot x^1 / p^0 \cdot x^0$,

$$\mathcal{P}_{\mathcal{L}} = \frac{x^1 \cdot p^1}{x^1 \cdot p^0}, \tag{15.3}$$

and so the implicit Laspeyres price index is the *Paasche price index*. Since

$$\frac{x^1 \cdot p^1}{x^1 \cdot p^0} = \frac{p^1 \cdot x^1}{p^0 \cdot x^0} \Big/ \frac{p^0 \cdot x^1}{p^0 \cdot x^0},$$

one computes an entry in the P^1/P^0 column by simply taking the ratio of the entries in the V^1/V^0 and Q^1/Q^0 columns. For example, the price index for total revenue is calculated as $1.195/1.1125 = 1.0742$. Similarly, the price index for the materials input category is $1.2418/1.1163 = 1.1125$.

With regard to total revenue, there was a 19.5% growth, of which $100(0.1125/0.195) = 58\%$ is attributable to output growth and 42% is attributable to price growth. With regard to the materials input category, there was a 24.2% growth, of which $100(0.1163/0.2418) = 48\%$ is attributable to input growth and 52% is attributable to price growth.

For each column in the table, an index of greater than one shows an increase, whereas an index of less than one shows a decrease. Generally speaking, one prefers indexes of greater than one for output and less than one for input. Naturally, if output is increasing, one may expect an increase in input. If the input increase is less than the output increase, a productivity gain takes place. While productivity gains are desirable, the firm must respond to price changes in both output and inputs, perhaps shifting production to more profitable product lines.

15.3 Productivity Assessment

A productivity index

$$\text{Productivity Index} = \frac{\text{Output Index}}{\text{Input Index}} \tag{15.4}$$

takes the general form of the ratio of an Output Index to an Input Index. The theory of index numbers suggests several possibilities for the output/input indexes: Laspeyres quantity index, Paasche quantity index, Fisher ideal quantity index, and the Tornqvist quantity index.

Consider first the use of the Laspeyres quantity index. It has already been calculated—the index for each output/input is recorded in column Q^1/Q^0 in Table 15.2. With respect to output it is 1.1125 and with respect to input it is 1.1310. Using the Laspeyres quantity index, the productivity assessment for Steeples Furniture is $1.1125/1.1310 = 0.9836$.

Next, consider the use of the Tornqvist quantity index. It takes the general form

$$\prod_k \left(\frac{z_k^1}{z_k^0}\right)^{(S_k^1+S_k^0)/2}, \tag{15.5}$$

where z_k^t denotes the quantity of (input, output) factor k used in period t and S_k^t denotes the cost or revenue share for factor k in period $t = 0, 1$. Table 15.3 displays the relevant calculations. Using the Tornqvist quantity index, the productivity assessment for Steeples Furniture is $1.1121/1.1304 = 0.9838$, which is remarkably close to the productivity assessment obtained

using the Laspeyres quantity index. Note how the capital inventory input ratio is 0.6667 but its productivity factor is 0.9933, which contributes ever so slightly to increasing the productivity ratio. The reason for this is that the Tornqvist index, rightfully so, accounts for the fact that the huge reduction in capital inventory input used in the subsequent year corresponds to an input whose average cost share is only 1%.

Table 15.3. Calculation of the Tornqvist quantity index.

	Output/Input ratio	S^0	S^1	Subindex
OUTPUT				
Sofas	1.1333	0.7500	0.7540	1.0987
Tables	1.0500	0.2500	0.2460	1.0122
OUTPUT INDEX				1.1121
INPUT				
Materials				
wood	1.0625	0.1000	0.1104	1.0064
fabric	1.1429	0.1641	0.1637	1.0221
cushioning	1.1371	0.0484	0.0485	1.0062
MATERIALS INDEX				1.0350
Labor				
woodcutter	1.1146	0.0750	0.0730	1.0081
lathe operator	1.0500	0.0125	0.0115	1.0006
assembler	1.1146	0.0750	0.0730	1.0081
sander	1.0556	0.0127	0.0117	1.0007
finisher	1.1502	0.3248	0.3550	1.0487
LABOR INDEX				1.0671
Capital				
inventory	0.6667	0.0234	0.0097	0.9933
machinery	1.2143	0.1641	0.1435	1.0303
CAPITAL INDEX				1.0234
INPUT INDEX				1.1304

15.4 Performance Ratios

In what follows, we let x_k^t, p_k^t and E_k^t denote, respectively, the input, price and expenditure of input k in period $t = 0, 1$; we drop the superscript when referring to an input category. We let $\mathcal{R}^*$, $\mathcal{Q}^*$ and $\mathcal{P}^*$ denote the performance indexes, respectively, associated with total revenue—for Steeples Furniture, these ratios are 1.1950, 1.1125 and 1.0742, respectively. Keep in mind

$$\mathcal{R}^* = \mathcal{Q}^* \cdot \mathcal{P}^*.$$

Table 15.4 records **performance ratios** to measure the extent to which an input factor or input category contributed to Steeples Furniture's profitability year-to-year, and to decompose the measure of profitability into its productivity and price recovery components. Here is how the columns are computed.

Table 15.4. Performance ratios.

Input	Profitability	Productivity	Price Recovery
Materials			
wood	0.8998	1.0471	0.8594
fabric	0.9958	0.9734	1.0230
cushioning	0.9914	0.9784	1.0134
Total Materials	0.9623	0.9966	0.9656
Labor			
woodcutter	1.0211	0.9981	1.0230
lathe operator	1.0839	1.0595	1.0230
assembler	1.0211	0.9981	1.0230
sander	1.0781	1.0539	1.0230
finisher	0.9091	0.9672	0.9399
Total Labor	0.9479	0.9805	0.9668
Capital			
inventory	2.3900	1.6687	1.4323
machinery	1.1355	0.9162	1.2394
Total Capital	1.2153	0.9709	1.2517
Total Cost	0.9935	0.9836	1.0101

15.4.1 Profitability Ratio

For input k the **change in profitability** is the ratio

$$\frac{\mathcal{R}^*}{E_k^1/E_k^0}, \tag{15.6}$$

which divides the growth in total revenue by the growth in the expenditure for this input. With regard to an input category, the denominator in (15.6) is the ratio of the total expenditure on inputs in this category in period 1 to the total expenditure on inputs in this category in period 0. For the material input category, the change in profitability is $1.195/1.2418 = 0.9623$. The growth in revenue was 19.5%, but the growth in the material input expenditure was 24.2%. The index is less than one, indicating a negative change in profitability for this input category. The change in profitability can also be calculated as

$$\frac{\mathcal{R}^* \, E_k^0}{E_k^1}, \tag{15.7}$$

which divides the *projected* expenditure in period 1 to accommodate the observed growth in total revenue by the *actual* expenditure in period 1. For example, 200,000 was spent on materials in period 0. Given the 19.5% increase in total revenue, i.e., $100(\mathcal{R}^* - 1)$, the projected expenditure on materials in period 1 to accommodate this increase in total revenue would be 239,000. The actual expenditure was 248,365 (which is higher) and so the ratio is $239{,}000/248{,}365 = 0.9623$.

15.4.2 Productivity Ratio

For input k the **change in productivity** is the ratio

$$\frac{\mathcal{Q}^*}{x_k^1/x_k^0}, \tag{15.8}$$

which divides the growth in aggregate output (as measured by the quantity index) to the growth in the usage for this input. With regard to an input category, the denominator in (15.8) is the Laspeyres quantity index (15.1) for that category. For the material input category, the change in productivity is $1.1125/1.1163 = 0.9966$. The growth in aggregate output is measured as 11.3%, but the growth in the material input expenditure was 11.6%, slightly higher. The index is less than one, indicating a negative change in productivity for this input category. The change in productivity can also be calculated as

$$\frac{\mathcal{Q}^* \, x_k^0}{x_k^1}, \tag{15.9}$$

which divides the *projected* usage in period 1 to accommodate the observed growth in total output by the *actual* quantity in period 1. For example, 4,000 board-ft of wood was used in period 0. Given the 11.3% increase in aggregate output, i.e., $100(\mathcal{Q}^* - 1)$, the projected usage of wood in period 1 to accommodate this increase in total output would be 4,450. The actual usage was 4,250 (which is lower) and so the ratio is $4{,}450/4{,}250 = 1.0471$. Here, the index is greater than one, indicating a productivity increase for this input.

15.4.3 Price Recovery Ratio

For input k the **change in price recovery** is the ratio

$$\frac{\mathcal{P}^*}{p_k^1/p_k^0}, \tag{15.10}$$

which divides the growth in aggregate price (as measured by the price index) by the growth in the prices for this input. With regard to an input category,

the denominator in (15.10) is the implicit Laspeyres price index (15.3) for that category, which we have shown equals the Paasche price index. For the material input category, the change in price recovery is $1.0742/1.1125 = 0.9656$. The growth in aggregate price is measured as 7.4%, but the growth in the material input category price was 11.3%. The index is less than one, indicating a negative change in price recovery for this input category. The change in price recovery can also be calculated as

$$\frac{\mathcal{P}^* \, p_k^0}{p_k^1}, \tag{15.11}$$

which divides the *projected* price in period 1 to keep it in line with the observed growth in total price by the *actual* price in period 1. For example, the unit price of wood was 16.00 in period 0. Given the 7.4% increase in aggregate price, i.e., $100(\mathcal{P}^* - 1)$, the projected unit price of wood in period 1 to keep it in line with this increase in total price would be 17.19. The actual unit price was 20.00 (which is much higher) and so the ratio is $17.19/20.00 = 0.8594$. Here, the index is less than one, indicating a negative price recovery for the wood input.

Remark 15.2. The product of the change in productivity and the change in price recovery always equals the change in profitability. This is not an accident. Indeed, for input k the product of (15.8) and (15.10) yields

$$\frac{\mathcal{Q}^* \cdot \mathcal{P}^*}{p_k^1 x_k^1 / p_k^0 x_k^0} = \frac{\mathcal{R}^*}{E_k^1 / E_k^0}, \tag{15.12}$$

which equals (15.6). For an input category, the product of the change in productivity and the change in price recovery equals

$$\frac{\mathcal{Q}^*}{p^0 \cdot x^1 / p^0 \cdot x^0} \, \frac{\mathcal{P}^*}{p^1 \cdot x^1 / p^0 \cdot x^1} = \frac{\mathcal{R}^*}{E^1 / E^0}, \tag{15.13}$$

which equals the change in profitability for the input category.

When analyzing the performance ratios for Steeples Furniture with regard to profitability, the material and labor input categories contributed negatively, which is attributable for the most part to the lack of price recovery as opposed to lack of productivity. The capital input category, on the other hand, contributed very positively to profitability, which is attributable to price recovery. (Recall Steeples Furniture took steps to lower its cost of inventory and invested to upgrade the quality of its machinery, thereby lowering its cost of maintenance/economic depreciation.)

15.5 Distribution of Net Gain

Steeples Furniture had a profit of 160,000 in the base year, which increased to 186,207 in the subsequent year. Given the overall revenue growth of 19.5%, if

profits had stayed in line, i.e., experienced the same rate of growth, the profit in year 1 would have been 1.195(160,000) = 191,200. There is a net gain of 186,207 - 191,200 = -4993.

- How should one distribute this net gain (loss) among the different inputs?
- How much of an input's contribution to net gain is attributable to productivity versus price recovery?

Table 15.5 shows how to do this. Here is how each column is calculated.

Table 15.5. Distribution of net gain.

Input	Net Gain	Productivity	Price Recovery
Materials			
wood	(8,520)	3,200	(11,720)
fabric	(525)	(3,188)	2663
cushioning	(320)	(763)	443
Total Materials	(9,365)	(751)	(8,614)
Labor			
woodcutter	1,185	(100)	1,285
lathe operator	740	500	240
assembler	1,185	(100)	1,285
sander	702	461	241
finisher	(24,840)	(7,831)	(17,009)
Total Labor	(21,028)	(7,070)	(13,958)
Capital			
inventory	10,425	6,688	3,737
machinery	14,975	(10,688)	25,663
Total Capital	25,400	(4,000)	29,400
Total Cost	(4,993)	(11,821)	6,828

15.5.1 Net Gain

The **net gain** due to input k or input category is

$$\mathcal{R}^* E_k^0 - E_k^1, \tag{15.14}$$

which merely subtracts the actual expenditure in period 1 on this input (input category) from the projected expenditure to accommodate the growth in total revenues. For the material input category, the projected expenditure in year 1 would be 200,000(1.195) = 239,000. The actual expenditure was 248,365. The difference is 239,000 - 248,365 = 9,365.

15.5.2 Net Gain Due to Productivity

The **net gain due to productivity** for input k is

$$\mathcal{Q}^* E_k^0 - p_k^0 x_k^1, \tag{15.15}$$

which subtracts the expenditure in period 1 on this input *using the based period's prices* to the projected expenditure on this input to accommodate the growth in aggregate output. The net gain due to productivity for an input category is

$$\mathcal{Q}^* E^0 - p^0 \cdot x^1. \tag{15.16}$$

The use of base period prices attempts to separate out the price effect to focus just on the quantity effect. With regard to the wood input, at base period prices the expenditure on wood in period 1 would have been 16.00(4,250) = 68,000. The projected increase in the expenditure on wood to accommodate the 11.3% increase in aggregate output is (1.1125)64,000 = 71,200. The difference is 71,200 - 68,000 = 3,200. For the wood input the net gain due to productivity is quite positive. With regard to the material input category, the net gain due to productivity is (1.1125)200,000 - [16.00(4,250) + 30.00(4,000) + 2.50(14,100)] = -751.

Remark 15.3. By multiplying and dividing (15.15) by E_k^0 or (15.16) by E^0, the net gain due to productivity can also be easily calculated using the performance indexes in Table 15.2 as either $E_k^0[\mathcal{Q}^* - (Q^1/Q^0)_k]$ for input k or $E^0[\mathcal{Q}^* - (Q^1/Q^0)]$ for an input category. (Here, the symbols $(Q^1/Q^0)_k$ or (Q^1/Q^0) refer to the entries in Table 15.2 corresponding to input k or the input category. Recall this column records the Laspeyres quantity index.) For example, for the wood input, we calculate 64,000[1.1125 - 1.0625] = 3,200, and for the materials input category, we calculate 200,000[1.1125 - 1.1163] = -751.

15.5.3 Net Gain Due to Price Recovery

The **net gain due to price recovery** is *defined* so that the *sum* of it and the net gain due to productivity *equals* the net gain. For example, for the materials input category, -9,365 = -751 + -8,614.

Remark 15.4. One could calculate the net gain due to price recovery analogously to how the net gain due to productivity is calculated, but then the net gain cannot be decomposed into its productivity and price recovery components. Consider, for example, the wood inputs. Using the method of calculation described in Remark 15.3, the net gain due to price recovery could be calculated as 64,000[1.0742 - 1.2500] = -11,251, which is close to -11,720. For the materials input category, the net gain due to price recovery could be calculated as 200,000[1.0742 - 1.1125] = -7660, which is less close to -8,614. The reason for this discrepancy is that the indexes are by design *multiplicative*, i.e., $\mathcal{R}^* = \mathcal{Q}^* \cdot \mathcal{P}^*$, not additive, i.e., $\mathcal{R}^* \neq \mathcal{Q}^* + \mathcal{P}^*$.

With regard to the distribution of the net gain, the wood material input and the finisher labor input contributed quite negatively. On the other hand, Steeples Furniture did reasonably well on the other labor inputs, and its efforts with regard to the capital input were essential to only losing 4993 in projected profits.

15.6 Exercises

15.1. Provide a precise statement of the aggregation of the Laspeyres quantity indexes discussed in Remark 15.1 and prove this aggregation is valid.

15.2. Show how to obtain the numbers recorded for the capital inputs and total capital in Table 15.2.

15.3. How would the numbers recorded in Table 15.2 change if the P^1/P^0 column were computed using the Paasche price index and the Q^1/Q^0 were computed as the implicit Paasche quantity index?

15.4. Show how to obtain the numbers recorded for the material inputs and materials index in Table 15.3.

15.5. Compute the Tornqvist price index for output and each factor category for Steeples Furniture.

15.6. Show how to obtain the profitability, productivity, and price recovery numbers recorded for the capital inputs and total capital in Table 15.4.

15.7. Show how to obtain the net gain, productivity, and price recovery numbers recorded for the material inputs and total materials in Table 15.5.

15.7 Bibliographical Notes

The performance measurement discussed here is an adaptation of the American Productivity Center's Productivity Measurement System. See Kendrick [1984].

15.8 Solutions to Exercises

15.1 Let $x^t = (x_1^t, x_2^t, \ldots, x_M^t)$ and $p^t = (p_1^t, p_2^t, \ldots, p_M^t)$ denote the input and price vectors in periods $t = 0, 1$, respectively. Here, x_j^t and p_j^t denote the sub-vector of inputs and prices corresponding to category j, $j = 1, 2, \ldots, M$. The Laspeyres quantity index P_L is $p^0 \cdot x^1 / p^0 \cdot x^0$. The Laspeyres quantity index for category j is $P_L^j = p_j^0 \cdot x_j^1 / p_j^0 \cdot x_j^0$, and the cost share in period 0 for category j is $S_j^0 = p_j^0 \cdot x_j^0 / p^0 \cdot x^0$. Thus, the Laspeyres quantity index can be written as

$$\begin{aligned}
\frac{p^0 \cdot x^1}{p^0 \cdot x^0} &= \frac{\sum_{i=1}^n p_i^0 x_i^1}{\sum_{i=1}^n p_i^0 x_i^0} \\
&= \frac{\sum_{j=1}^M p_j^0 \cdot x_j^1}{\sum_{j=1}^M p_j^0 \cdot x_j^0} \\
&= \sum_{j=1}^M \Big(\frac{p_j^0 \cdot x_j^0}{\sum_{j=1}^M p_j^0 \cdot x_j^0} \Big) \Big(\frac{p_j^0 \cdot x_j^1}{p_j^0 \cdot x_j^0} \Big) \\
&= \sum_{j=1}^M S_j^0 P_L^j,
\end{aligned}$$

as claimed in the numerical example in the Remark.

15.2 For column V^1/V^0: inventory $= 7,500/15,00 = 0.5000$, machinery $= 110,500/105,000 = 1.0524$ and Total Capital $= 118,000/120,000 = 0.9833$. For column Q^1/Q^0: inventory $= 50,000/75,000 = 0.6667$ and machinery $= 850,000/700,00 = 1.2143$. To compute the numbers for Total Capital, we first compute the cost shares in the base period, which are $[0.20(75,000)/(0.20(75,000) + 0.15(700,000))] = 0.125$ for inventory and $1 - 0.125 = 0.875$ for machinery. For Total Capital we then compute $0.125(0.6667) + 0.875(1.2143) = 1.1458$. For column P^1/P^0: inventory $= 0.5000/0.6667 = 0.7500$, machinery $= 1.0524/1.2143 = 0.8667$, and Total Capital $= 0.9833/1.1458 = 0.8582$.

15.3 The Paasche price index is equivalent to the implicit Laspeyres price index and the implicit Paasche quantity index is the Laspeyres quantity index. So the numbers will not change.

15.4 For the Output/Input ratio column: wood $= 4,250/4,000 = 1.0625$, fabric $= 4,000/3,500 = 1.1429$, and cushioning $= 14,100/12,400 = 1.1371$. For column S^0: wood $= 64,000/640,000 = 0.1000$, fabric $= 105,000/640,000 = 0.1641$, and cushioning $= 31,000/640,000 = 0.0484$. For column S^1: wood $= 85,000/769,793 = 0.1104$, fabric $= 126,000/\ 769,793 = 0.1637$, and cushioning $= 37,365/769,793 = 0.0485$. For the Subindex column, we first compute the average cost shares, which are 0.1052, 0.1639 and 0.04845 for wood, fabric and cushioning, respectively. Thus, wood $= (1.0625)^{0.1052} = 1.0064$, fabric

$= (1.1429)^{0.1639} = 1.0221$ and cushioning $= (1.1371)^{0.04845} = 1.0062$. The Materials Index $= (1.0064)(1.0221)(1.0062) = 1.0350$.

15.5 The materials Tornqvist price index $=$

$$\left(\frac{20.00}{16.00}\right)^{0.1052}\left(\frac{31.50}{30.00}\right)^{0.1639}\left(\frac{2.65}{2.50}\right)^{0.04845} = 1.0349.$$

(The average cost shares are computed for the previous exercise.) The other indexes are computed similarly: the output price index $= 1.0746$, the price index for labor $= 1.0463$, the price index for capital $= 0.9736$, and the input price index $= 1.0542$.

15.6 For the Profitability column: $\mathcal{R}^*$, the V^1/V^0 ratio in Table 15.2, is 1.1950. Thus, inventory $= 1.1950/0.5000 = 2.3900$, machinery $= 1.1950/0.9833 = 1.1355$, and Total Capital $= 1.1950/0.9833 = 1.2153$. For the Productivity column: $\mathcal{Q}^*$, the Q^1/Q^0 ratio in Table 15.2, is 1.1125. Thus, inventory $= 1.1125/0.6667 = 1.6687$, machinery $= 1.1125/1.2143 = 0.9162$, and Total Capital $= 1.1125/1.1458 = 0.9709$. For the Price Recovery column, by definition $(P^1/P^0)(Q^1/Q^0) = V^1/V^0$. Thus, inventory $= 2.3900/1.6687 = 1.4323$, machinery $= 1.1355/0.9162 = 1.2394$ and Total Capital $= 1.2153/0.9709 = 1.2517$.

15.7 For the Net Gain column: $\mathcal{R}^* = 1.1950$, and so wood $= 1.1950(64,000) - 85,000 = -8,520$, fabric $= 1.1950(105,000) - 126,000 = -525$, cushioning $= 1.1950(31,000) - 37,365 = -320$, and Total Materials $= -8,520 - 525 - 320 = -9,365$. For the Productivity column: $\mathcal{Q}^* = 1.1125$, and so wood $= 1.1125(64,000) - 16.00(4,250) = 3,200$, fabric $= 1.1125(105,000) - 30.00(4,000) = -3,188$, cushioning $= 1.1125(31,000) - 2.50(14,100) = -763$, and Total Materials $= 3,200 - 3,188 - 763 = -751$. By definition, the entries recorded in the Price Recovery column are such that the sum of productivity and price recovery equals the net gain. Thus, wood $= -8,520 - 3,200 = 11,720$, fabric $= -525 - (-3,188) = 2,663$, cushioning $= -320 - (-763) = 443$, and Total Materials $= -9,365 - (-751) = -8,614$.

16

Economic Analysis

Up to now, we have ignored the market aspects of the story. Our focus has been on different ways to model technology and assess a firm's efficiency and productivity. We argue this is a critical first step to improving a firm's productivity. Firms are constantly looking for ways to innovate, either in the product market or in the means of production. In this chapter, we consider how a firm's price, output, labor employed, revenue, profit, and market share are affected by the market in which the firm competes. In the parlance of economics, we will analyze a specific market with consumers and producers, and derive the market's general equilibrium under a variety of different market structures. We use the analysis to quantitatively assess how a "productivity laggard" or a "productivity leader" would fare against its competition.

16.1 Market Structure and Equilibrium

Our market consists of N firms, each one producing a unique good or service. Labor is the sole variable factor of production. The labor elasticity of output is constant across firms, but the marginal product can be different due to different levels of productivity. To make matters concrete, we assume producer i's production function is

$$\Phi_i(L_i) \;=\; A_i L_i^a, \tag{16.1}$$

where elasticity of output parameter a is positive and less than one. The parameter A_i can be viewed as a productivity index, which is permitted to be different for each firm.

Loosely worded, a *market structure* defines what information each producer is assumed to incorporate in his decision-making. In all market structures we study here, producer i maximizes his profit, $\pi_i(p, w)$, which depends on the prevailing prices for the goods or services, p, and the prevailing wage rate, w, which influences his cost of production. In doing so, producer i determines his supply to the market, $S_i(p, w)$, termed the *supply schedule*, and his

corresponding demand for labor, $L_i(p, w)$, termed the *labor demand schedule.* The supply and labor demand schedules will depend on the market structure.

In this chapter, we examine three market structures:

- *Competitive market structure.* Here, the producer is a *price-taker* in that he does not consider how his price affects consumer demand or is affected by what other producers are charging. He merely assumes he can sell as much output as he desires at the prevailing price. Thus, his supply and labor demand schedules are only functions of his own price and the wage.
- *Monopolistic competitive market structure.* Here, each producer considers how his price affects consumer demand for his product. He does so only indirectly, since he assumes his product market represents a small fraction of the overall market.
- *Oligopoly market structure.* Here, there are few producers, so each producer considers how all prices directly affect his demand. In this market structure, a *Nash equilibrium* in prices will be determined, and the supply and labor demand schedules of each producer are now (non-trivial) functions of all prices and the wage.

Analysis of the producer side of the market determines supply and demand for labor as a function of prices and wage. To determine the equilibrium level of prices, wage, and outputs of each good, we have to examine the consumer side of the market. Our market consists of M consumers, each one of which inelastically supplies one unit of labor.[1] Each consumer has identical preferences for the goods or services given by the CES utility function

$$U(x_1, x_2, \ldots, x_N) \;:=\; \left[\sum_{i=1}^{N} (q_i x_i)^r \right]^{1/r}. \tag{16.2}$$

The parameter r is assumed positive and less than one. Each q_i represents a quality index, and so $q_i x_i$ can be thought of as *quality-adjusted* consumption. Given the prevailing prices p for the goods or services and her assumed budget (or income) I, each consumer determines her demand for each good or service i to maximize her overall utility. When summed over all consumers, this determines the *aggregate demand*, $D_i(p, w, I)$ for product i. Each consumer owns an equal share of each firm. Consequently, each consumer's income derives from two sources: her wage and the dividends she receives.

Definition 16.1. *An* **equilibrium** *is a vector of prices p and a wage rate w such that the following properties hold:*

A. Each producer maximizes his profit given p and w and the prevailing market structure.
B. Each consumer maximizes her utility given p and w.

[1] We shall not consider the "labor/leisure" choice here.

C. Supply equals aggregate demand for each good or service i:

$$S_i(p,w) \;=\; D_i(p,w,I); \tag{16.3}$$

that is, the product markets clear.

D. The labor market clears, namely,

$$\sum_{i=1}^{N} L_i(p,w) \;=\; M. \tag{16.4}$$

E. Each consumer's income matches expectations, namely,

$$I \;=\; w \;+\; (1/M)\sum_i \pi_i(p,w). \tag{16.5}$$

16.2 Competitive Market Structure

16.2.1 Consumers

Each consumer maximizes her utility subject to her budget constraint. Formally, the consumer's optimization problem is

$$\Gamma(p,I) \;:=\; \max_x \left\{ \left[\sum_{i=1}^{N} (q_i x_i)^r \right]^{1/r} : \sum_{i=1}^{N} p_i x_i \le I \right\}. \tag{16.6}$$

As mentioned before, the q coefficients can be thought of as indexes of quality associated with each product. Since the utility function exhibits constant returns-to-scale,

$$\Gamma(p,I) = \Gamma(p,1)I \text{ and } x(p,I) = x(p,1)I.$$

To ease subsequent notational burdens, we use **quality-adjusted** units and prices, respectively given by

$$\hat{x}_i \;:=\; q_i x_i, \;\; \hat{p}_i \;:=\; p_i/q_i. \tag{16.7}$$

It is sufficient to optimize U^r in lieu of U. Accordingly, first-order optimality conditions imply existence of a positive μ such that

$$r\hat{x}_i^{r-1} \;=\; \mu\hat{p}_i, \tag{16.8}$$

which in turn implies that

$$\hat{x}_i \;=\; \left(\frac{\mu}{r}\right)^{1/(r-1)} \hat{p}_i^{1/(r-1)}. \tag{16.9}$$

Using (16.9), and the fact that the budget must be tight at the optimum,

$$I \;=\; \sum_i \hat{p}_i \hat{x}_i \;=\; (\frac{\mu}{r})^{1/(r-1)} \sum_i \hat{p}_i^{\,r/(r-1)}. \tag{16.10}$$

Together, (16.9) and (16.10) show that consumer demand for product i is

$$\hat{x}_i \;=\; \left[\frac{\hat{p}_i^{1/(r-1)}}{\sum_i \hat{p}_i^{r/(r-1)}} \right] I\,. \tag{16.11}$$

The **demand schedule**, aggregated over all M consumers, is therefore

$$D_i(\hat{p}, w, I) \;=\; \left[\frac{\hat{p}_i^{1/(r-1)}}{\sum_i \hat{p}_i^{r/(r-1)}} \right] (IM). \tag{16.12}$$

16.2.2 Producers

In quality-adjusted units, the producer's profit maximization problem is

$$\pi_i(\hat{p}, w) \;:=\; \max_{L_i} \{\; \hat{p}_i[\hat{A}_i L_i^a] \;-\; wL_i \;\},$$

where we let $\hat{A}_i := q_i A_i$ denote the *quality-adjusted* productivity index. First-order optimality conditions imply that

$$\hat{p}_i \hat{A}_i a L_i^{a-1} \;=\; w. \tag{16.13}$$

This in turn implies the **labor employed and supply schedules** are

$$L_i(\hat{p}, w) \;=\; (\hat{p}_i \hat{A}_i a / w)^{1/(1-a)}, \tag{16.14}$$

$$S_i(\hat{p}, w) \;=\; \hat{A}_i^{1/(1-a)} [\hat{p}_i a / w]^{a/(1-a)}. \tag{16.15}$$

Let C_i, R_i and $\pi_i = R_i - C_i$ denote the optimal cost, revenue and profit. Multiplying both sides of (16.13) by L_i shows that it will *always* be true in this setting that

$$aR_i \;=\; C_i \tag{16.16}$$

$$\pi_i \;=\; (1-a)R_i \;=\; \Big[\frac{(1-a)}{a}\Big] C_i. \tag{16.17}$$

16.2.3 Equilibrium

Equilibrium conditions (A) and (B) are automatically met if the supply and demand schedules derived in the previous sections are maintained. Since labor is supplied inelastically and all income I is spent,

$$\begin{aligned} \sum_i \pi_i \;&=\; \sum_i R_i \;-\; \sum_i C_i \\ &=\; MI \;-\; w \sum_i L_i \\ &=\; MI \;-\; wM, \end{aligned} \tag{16.18}$$

and so condition (E) is automatically met. In our special setting, we can say more. In light of (16.17), conditions (D) and (E) imply that

$$I \;=\; w + (1/M)\Big[\frac{(1-a)}{a}\sum_i wL_i\Big] \;=\; w/a. \tag{16.19}$$

In particular, (16.19) shows that the equilibrium wage to income ratio *always* equals a, and the equilibrium dividend to income ratio *always* equals $(1-a)$. We shall now make use of these facts when analyzing condition (C).

Condition (C) says the product markets must clear in equilibrium. Let

$$\mathcal{P}_i := \hat{p}_i/I \tag{16.20}$$

denote the **price-income ratio**. Equating the supply schedule (16.15) to the demand schedule (16.12), and using (16.19),

$$\mathcal{P}_i \;=\; \frac{(M/\Delta)^{(1-a)(1-r)/(1-ar)}}{A_i^{(1-r)/(1-ar)}}, \tag{16.21}$$

where

$$\Delta \;:=\; \sum_i \mathcal{P}_i^{r/(r-1)}. \tag{16.22}$$

Substituting (16.21) into (16.22), we obtain that

$$\Delta \;=\; \frac{\mathcal{E}^{(1-ar)/(1-r)}}{M^{(1-a)r/(1-r)}}, \tag{16.23}$$

where

$$\mathcal{E} \;:=\; \sum_i \hat{A}_i^{r/(1-ar)}. \tag{16.24}$$

Substituting (16.23) into (16.21), the **equilibrium price-income ratios** are

$$\mathcal{P}_i^* \;=\; (M/\mathcal{E})^{1-a} \;*\; \xi_i, \tag{16.25}$$

where

$$\xi_i := \;\frac{1}{\hat{A}_i^{(1-r)/(1-ar)}}. \tag{16.26}$$

The relative equilibrium prices are fixed *a priori*, given by the vector ξ. Substituting (16.26) into (16.14), and using (16.19) and (16.20), the **equilibrium labor employed** by producer i is

$$L_i^* \;=\; \hat{A}_i^{r/(1-ar)} M/\mathcal{E}. \tag{16.27}$$

It follows that the **equilibrium (quality-adjusted) output** by producer i is

$$S_i^* \;=\; q_i A_i {L_i^*}^a = \hat{A}_i {L_i^*}^a = \hat{A}_i^{1/(1-ar)}(M/\mathcal{E})^a. \tag{16.28}$$

Using (16.17) and (16.19), the **equilibrium profit-income ratios** are

$$\pi_i^*/I \;=\; (1-a)L_i^*. \tag{16.29}$$

Remark 16.2. Equations (16.29) and (16.27) show that in equilibrium the ratio of the profits of firm i to firm j is the *a priori* constant

$$(\hat{A}_i/\hat{A}_j)^{r/(1-ar)}. \tag{16.30}$$

Since $\hat{A}_i$ is the product of the firm's quality and productivity indexes, *as a firm's relative product quality and/or productivity increases, its relative profits increase* (as one would expect).

16.2.4 Comparative Statics

We begin by examining the following question: What happens when the number of firms grows? To examine this question, we imagine the existing firms are replicated so that the constant $\mathcal{E}$ in (16.24) grows linearly in the number of firms. It is clear from (16.26)-(16.29) that the *price-income ratio, labor employment levels, outputs, and the profit-income ratio all decline when $\mathcal{E}$ increases.* But is our consumer worse off? It might be tempting to say so, but we must be careful: the number of products the consumer purchases may increase with the number of firms and, as we shall see, consumers like variety.

To investigate, we examine the consumer's *equilibrium level of utility.* The per-capita demand for good i is the aggregate demand S_i^* (16.28) divided by the number of consumers M. Substituting S_i^*/M into the consumer's utility function (16.2) and using (16.24),

$$U^* \;=\; \frac{\mathcal{E}^{(1-ar)/r}}{M^{1-a}}\;, \tag{16.31}$$

which is *increasing* in $\mathcal{E}$, and so the consumer is in fact better off. The elasticity of U^* with respect to $\mathcal{E}$ depends critically on the elasticity of output parameter a and the degree of substitution parameter r. When $r \to 0$ the substitution among products is virtually nil, and so the consumer is far better off (in relative terms) with an increase in product variety. When $r \to 1$ products are virtual substitutes for one another, and so the consumer is much less better off (in relative terms) with an increase in product variety. When $a \to 0$, output elasticity is so low the consumer needs an increase in product variety to see any increase in utility.

What happens when the number of firms remains the same but the number of consumers increases? It is clear from (16.26)-(16.29) that the *price-income ratio, labor employment levels, outputs, and the profit-income ratio all increase when M increases.* However, from (16.31) we see that consumer utility *decreases* since $0 < a < 1$.

What happens when the number of firms and consumers both increase at the same rate? All levels remain the same, but since there is an increase in product variety, the consumer is far better off. In particular, the elasticity of U^* with respect to the scale factor is $(1 - ar)/r \;-\; (1 - a) = (1 - r)/r$.

16.3 Monopolistic Competitive Market Structure

In a monopolistic competition, *producers incorporate the shape of the consumer's demand function when they seek to maximize their profits.*

Suppose producer i employs L_i units of labor at a cost of wL_i. The (quality-adjusted) supply is $\hat{A}_i L_i^a$. The producer knows the form of the consumer's demand function, as specified in (16.12), and therefore knows that the equilibrium price that will clear his market is

$$\hat{A}_i L_i^a = \Lambda \hat{p}_i^{1/(r-1)}, \tag{16.32}$$

$$\Lambda := \frac{IM}{\sum_i \hat{p}_i^{r/(r-1)}} . \tag{16.33}$$

We assume there are enough firms so that each producer does not consider how his decision will affect Λ, and therefore each producer determines his labor demand schedule as a function of Λ. Given that the labor market must clear, a unique equilibrium value for Λ will emerge.

In this setting, the producer's profit function is

$$\begin{aligned} \pi_i(\Lambda, L_i) &= \hat{p}_i[\hat{A}_i L_i^a] - wL_i \\ &= \Lambda \hat{p}_i^{\frac{r}{r-1}} - wL_i \\ &= [\Lambda^{1-r} \hat{A}_i^r] L_i^{ar} - wL_i. \end{aligned} \tag{16.34}$$

Since the form of the profit function

$$\beta_i L_i^b - wL_i$$

is identical to the competitive case, it can be readily verified that in equilibrium

$$I = w/ar . \tag{16.35}$$

Moreover, the labor employed by producer i is

$$L_i = \left[\frac{\Lambda^{1-r} \hat{A}_i^r}{w/ar} \right]^{1/(1-ar)} \propto \hat{A}_i^{r/(1-ar)}, \tag{16.36}$$

where the symbol $\propto$ means "proportional up to a positive constant." Since

$$\sum_i L_i = M, \tag{16.37}$$

it follows from (16.36) that the **equilibrium labor employed** by producer i is

$$L_i^* = \hat{A}_i^{r/(1-ar)} (M/\mathcal{E}), \tag{16.38}$$

where $\mathcal{E}$ is defined as before in (16.24). Since (16.38) is identical to (16.27), the labor employed and thus the equilibrium outputs for producer i remain *unchanged*. Equating (16.36) to (16.38) and using (16.35),

$$\Lambda^{1-r} \;=\; (M/\mathcal{E})^{1-ar} I\,. \tag{16.39}$$

From (16.32),

$$\hat{p}_i \;=\; \left(\frac{\Lambda}{\hat{A}_i L_i^a}\right)^{1-r}, \tag{16.40}$$

and so it now follows from (16.38) and (16.39) that the **equilibrium price-income ratios** are

$$\mathcal{P}_i^* \;=\; \frac{(M/\mathcal{E})^{1-a}}{\hat{A}_i^{(1-r)/(1-ar)}}. \tag{16.41}$$

These ratios are identical to (16.26). Consequently, the equilibrium price-income ratios remain *unchanged*, too.

It can be readily verified that the form of equation (16.17) remains the same, *except* that here the constant a must be replaced with ar. The **equilibrium profit-income ratios** are therefore

$$\pi_i^*/I \;=\; (1-ar)L_i^*. \tag{16.42}$$

Since $r < 1$, in equilibrium the incomes, prices, and firm profit's all *increase* (relative to the competitive setting), but outputs and hence utility remain *unchanged* since the price-income ratios do *not* change. What does change here is the relative proportion of income attributable to wage, which now declines from a to ar.

16.4 Social Planner's Perspective

In this section, we dispense with price mechanisms altogether and consider how a *social planner* would allocate resources to maximize consumer utility. A social planner must decide how to divide the consumer's unit of labor among the different firms so as to produce the vector of quantities that maximize her utility. Formally, the social planner solves

$$\max\left\{\left[\sum_i (A_i L_i^a)^r\right]^{1/r} \;:\; \sum_i L_i = 1\right\}.$$

It is sufficient to optimize U^r in lieu of U. For this modified problem, the Lagrangian is

$$\mathcal{L}(L,w) \;:=\; \sum_i (A_i L_i^a)^r \;-\; w\left(\sum_i L_i - 1\right).$$

Since the social planner's problem is a well-behaved concave optimization problem, it can be solved by first fixing the Lagrange multiplier w, optimizing $\mathcal{L}$ to obtain the solution vector $L(w)$, and then finding w for which

$\sum_i L_i(w) = 1$. Since the Lagrangian is additively-separable in the L_i's, each product market can be solved separately. An inspection of each product subproblem reveals that it is precisely equivalent to the producer's problem in the non-competitive setting (when $M = 1$). Thus, the optimal labor allocations (demands) selected by the social planner are *identical* to what is obtained in both market structures. Since the social planner will achieve the maximum utility for each consumer, we conclude that *the utility previously obtained in the competitive or monopolistically competitive market structure is at the highest possible level.*

16.5 Oligopoly Market Structure

In this setting, each producer knows the demand schedule given in (16.12). Producer i takes the other producer's prices as given and selects his price to maximize his profits. In contrast to the monopolistic competitive environment previously analyzed, however, here *each producer specifically accounts for the fact that the price he selects affects the denominator of the expression in brackets in (16.12).*

16.5.1 Profit Maximization Formulation

It will be convenient for the developments to follow to parameterize the profit optimization problem in a slightly different way. (As before, all units are quality-adjusted.) Define

$$\Lambda_i \;:=\; \sum_{j \neq i} \hat{p}_j^{\,r/(r-1)}, \tag{16.43}$$

$$\mathcal{H}_i \;:=\; \frac{\hat{p}_i^{\,r/(r-1)}}{\hat{p}_i^{\,r/(r-1)} \;+\; \Lambda_i} \;=\; \frac{1}{1 \;+\; \Lambda_i \hat{p}_i^{\,r/(1-r)}}\,. \tag{16.44}$$

Knowledge of all other prices is subsumed in the parameter Λ_i, which the producer takes as given. The variable $\mathcal{H}_i$ represents the producer's share of income.

Since $\mathcal{H}_i$ uniquely determines $\hat{p}_i$ (for given value of Λ_i), we shall use it in lieu of $\hat{p}_i$ as our producer's decision variable. The equilibrium prices must be positive, which ensures that $\mathcal{H}_i \in (0, 1)$. In particular,

$$\hat{p}_i(\Lambda_i, \mathcal{H}_i) \;=\; \left[\frac{1 - \mathcal{H}_i}{\mathcal{H}_i \Lambda_i}\right]^{(1-r)/r}. \tag{16.45}$$

The producer's revenue, R_i, is $(IM)\mathcal{H}_i$, and his supply, S_i, is R_i/p_i. Using (16.45), and the fact that $S_i = \hat{A}_i L_i^a$, the labor employed is

$$L_i(\Lambda_i, \mathcal{H}_i) \;=\; \frac{\mathcal{H}_i^{1/ar} \Lambda_i^{(1-r)/ar}}{\hat{A}_i^{1/a} (1-\mathcal{H}_i)^{(1-r)/ar}} \, (IM)^{1/a}. \tag{16.46}$$

The producer's profit function is

$$\pi_i(\Lambda_i, \mathcal{H}_i) \;:=\; (IM)\mathcal{H}_i \;-\; wL_i(\Lambda_i, \mathcal{H}_i). \tag{16.47}$$

16.5.2 Equilibrium

Using the fact that

$$\frac{d \ln L_i}{d\mathcal{H}_i} \;=\; (1/ar)\left[\frac{1}{\mathcal{H}_i} \;+\; \frac{1-r}{1-\mathcal{H}_i}\right],$$

the first-order optimality condition $\partial \pi_i / \partial \mathcal{H}_i = 0$ implies that

$$(ar/w)(IM) \;=\; L_i(\Lambda_i, \mathcal{H}_i)\left[\frac{1}{\mathcal{H}_i} \;+\; \frac{1-r}{1-\mathcal{H}_i}\right]. \tag{16.48}$$

The equilibrium values for Λ_i and $\mathcal{H}_i$ are *not* independent, a fact we shall use shortly. Substituting (16.45) into (16.43),

$$\Lambda_i \;=\; \sum_{j \neq i} \frac{\mathcal{H}_j \Lambda_j}{1-\mathcal{H}_j}. \tag{16.49}$$

Add $\mathcal{H}_i \Lambda_i / (1-\mathcal{H}_i)$ to both sides of (16.49) to obtain that

$$\Lambda_i \;=\; (1-\mathcal{H}_i)\mathcal{G}, \tag{16.50}$$

where

$$\mathcal{G} \;:=\; \sum_k \frac{\mathcal{H}_k \Lambda_k}{1-\mathcal{H}_k}\,. \tag{16.51}$$

Substituting (16.50) into (16.46) shows that

$$L_i(\Lambda_i, \mathcal{H}_i) \;\propto\; \mathcal{H}_i^{1/ar} \hat{A}_i^{-1/a}. \tag{16.52}$$

Since the labor market must clear, the **equilibrium labor demand** is thus

$$L_i^*(\Lambda_i, \mathcal{H}_i) \;=\; \frac{\mathcal{H}_i^{1/ar} \hat{A}_i^{-1/a}}{\Delta} \, M, \tag{16.53}$$

where

$$\Delta \;:=\; \sum_i \mathcal{H}_i^{1/ar} \hat{A}_i^{-1/a}. \tag{16.54}$$

Substituting (16.53) into (16.48) shows that *the equilibrium values for the $\mathcal{H}$'s must satisfy the following three sets of equations:*

$$(Iar/w)\Delta = \hat{A}_i^{-1/a}\mathcal{H}_i^{1/ar}\left[\frac{1}{\mathcal{H}_i} + \frac{1-r}{1-\mathcal{H}_i}\right], \tag{16.55}$$

$$\sum_i \mathcal{H}_i = 1, \tag{16.56}$$

$$\Delta = \sum_i \mathcal{H}_i^{1/ar}\hat{A}_i^{-1/a}. \tag{16.57}$$

It remains to show that a solution to (16.55)-(16.57) exists. We shall show there is a *unique* solution. Define

$$\kappa := (Iar/w). \tag{16.58}$$

The constant κ is inversely proportional to the wage-income ratio. The right-hand side of (16.55) is a strictly increasing, continuous function of $\mathcal{H}_i$ and its value is zero at zero. Hence, for each value of

$$\eta := \kappa\Delta, \tag{16.59}$$

there is a unique solution to (16.55), which we shall denote by $\mathcal{H}_i(\eta)$. Since $\mathcal{H}_i(\cdot)$ is strictly increasing and continuous, there exists a unique positive value, η^*, for which

$$\sum_i \mathcal{H}_i(\eta^*) = 1. \tag{16.60}$$

Consequently, $\mathcal{H}_i(\eta^*)$, $i = 1, 2, \ldots, N$, and

$$\Delta(\eta^*) := \sum_i \hat{A}_i^{-1/a}\mathcal{H}_i(\eta^*)^{1/ar}, \tag{16.61}$$

$$\kappa(\eta^*) := \eta^*/\Delta(\eta^*) \tag{16.62}$$

define the unique solution to (16.55)-(16.57).

Not surprisingly, the equilibrium wage-income and price-income ratios remain unchanged. That is, as the income scales, the prices and wage scale in exactly the same proportion. From (16.45), the equilibrium price-income ratio can be expressed as

$$\mathcal{P}_i^* = \hat{p}_i^*/I = \Big[\frac{1-\mathcal{H}_i}{\mathcal{H}_i(\Lambda_i I^{r/(1-r)})}\Big]^{(1-r)/r}, \tag{16.63}$$

and from (16.46), the labor employed by producer i can be expressed as

$$L_i^*(\Lambda_i, \mathcal{H}_i) = \frac{\mathcal{H}_i^{1/ar}[\Lambda_i I^{r/(1-r)}]^{(1-r)/ar}}{A_i^{1/a}(1-\mathcal{H}_i)^{(1-r)/ar}}\, M^{1/a}. \tag{16.64}$$

As I changes, the wage adjusts so that κ remains unchanged. This in turn implies the values for $\mathcal{H}_i$'s and hence the L_i^*'s remain unchanged. Thus, the $\Lambda_i I^{r/(1-r)}$ term in (16.64) remains constant, which implies the $\mathcal{P}_i^*$'s remain

unchanged, too. It follows then that one of the prices can be taken as the *numeraire* and set equal to one, which will then completely determine all other prices, incomes, and the other economic variables of interest.

It remains to devise a procedure to find the $\mathcal{H}_i$'s, Δ, and κ, the subject of the next subsection. Once these variables have been determined, the economic variables of interest are easily obtained, as follows:

- Take the wage rate as the *numeraire* and set its value to 1.
- The income I is κ/ar via (16.58).
- Δ determines labor employed by producer i, L_i^*, via (16.53).
- Labor employed by producer i, L_i^*, determines the firm's output as $A_i {L_i^*}^a$.
- The market share for producer i, $\mathcal{H}_i$, determines its revenue as $(IM)\mathcal{H}_i$, which in turn determines the firm's price, p_i, by dividing it by the output it produces.

16.5.3 Algorithm to Compute the Equilibrium

To find the unique equilibrium, one performs a bisection search on η until (16.60) is satisfied (to a desired degree of accuracy). Informally, if

$$\sum_i \mathcal{H}_i(\eta) \; < \; 1,$$

then increase the lower bound for η; if the opposite holds true, then lower the upper bound for η. The monotonicity and continuity of $\mathcal{H}_i(\cdot)$ ensures that this procedure will converge to the unique solution.

For each candidate value of η, it still remains to find the value for $\mathcal{H}_i(\eta)$ that solves (16.55). Again, bisection search can be used. Informally, if the right-hand side of (16.55) exceeds η for a particular value of $\mathcal{H}_i$, then increase the lower bound for $\mathcal{H}_i$; if the opposite holds true, then increase the lower bound for $\mathcal{H}_i$. This bisection search must be performed for each of the N product markets. For the special case $ar = 1/2$, the N bisection searches are *unnecessary*, since the nonlinear equation (16.55) can be transformed into a *quadratic* equation by multiplying both sides by $1 - \mathcal{H}_i$. There will be two positive roots, but since both $\mathcal{H}_i$ and $1-\mathcal{H}_i$ must be positive, the smaller root will be the correct solution. To see this, note that the form of the quadratic equation (dropping the subscript i) is

$$0 \; = \; \eta - (e+\eta)\mathcal{H} + er\mathcal{H}^2 \; , \tag{16.65}$$

where $e = A^{-1/a}$. The larger positive root is

$$\frac{(e+\eta) \; + \; \sqrt{(e+\eta)^2 \; - \; 4er\eta}}{2er} \; . \tag{16.66}$$

Since $r < 1$ the expression inside the square root exceeds $(e-\eta)^2$ (which shows that the smaller root will be indeed positive). It may then be verified that the numerator exceeds $2e$, which shows that the larger root will exceed one.

Remark 16.3. When $ar = 1/3$, the nonlinear equation (16.55) may be transformed into a *cubic* equation for which a closed form solution also exists.

16.5.4 Comparison to Competitive and Monopolistic Competitive Market Structures

In the competitive and monopolistic competitive market structures the labor employed by producer i is

$$L_i^* \;=\; \frac{\hat{A}_i^{r/(1-ar)}}{\sum_i \hat{A}_i^{r/(1-ar)}}\, M. \tag{16.67}$$

Comparing (16.67) to (16.53), in order for the labor employment levels to be identical it must be the case that

$$\mathcal{H}_i \;=\; \frac{\hat{A}_i^{r/(1-ar)}}{\sum_i \hat{A}_i^{r/(1-ar)}}\;, \tag{16.68}$$

which implies that

$$\mathcal{H}_i = L_i^*/M\;. \tag{16.69}$$

In the competitive market structure, identity (16.69) must hold because

$$\mathcal{H}_i \;=\; \frac{R_i^*}{\sum_i R_i^*} \;=\; \frac{wL_i^*/a}{\sum_i wL_i^*/a} \;=\; \frac{L_i^*}{\sum_i L_i^*} \;=\; L_i^*/M\;. \tag{16.70}$$

In general, identity (16.69) will *not* hold for an oligopoly market structure, since it is no longer the case that the producer's first-order optimality condition implies that

$$aR_i \;=\; wL_i. \tag{16.71}$$

In particular, substituting (16.68) into (16.55) does not render (16.55) immediately true.

Remark 16.4. In the special case when the firms are identical, so that $A_i = A$, the solution for the oligopoly setting will trivially satisfy (16.69), as the labor employed for each firm must be identical.

In conclusion, when the producers incorporate all of the information about prices into their respective profit maximization problem, the equilibrium solution will be different and must lead to an *inferior* allocation of resources from the consumer's perspective. (The previous allocations were the best, as they coincided with the social planner's solution.)

16.6 Productivity Analysis

We briefly examine the importance of productivity to a firm in a market place. To ease the burdens of computing the equilibrium in the oligopoly market structure, we fix $a = 0.8$ and $r = 0.625$ so that $ar = 0.5$. We set $M = 1$ so that labor employed by a firm equals its labor share. We analyze markets with $N = 3$, $N = 5$ and $N = 10$ firms. We shall think of firm 1 as the reference firm. For simplicity all remaining firms are identical. The quality index for each firm is set to one. The productivity index for all remaining firms is also set to one.

16.6.1 Analysis of a Productivity Laggard

In this subsection, we analyze the situation in which firm 1 is a "productivity laggard." Here, we set $A_1 = 0.5$. Table 16.1 shows the relevant market analysis.

With respect to firm 1, its

- profit is 40% of its competition,
- labor and market share are half of what it would be if it could match the productivity of its competition, and
- output is 25% of its competition.

These trends *persist* even when the number of firms decline. A sizable productivity deficit, therefore, cannot be overcome even if there are fewer competitors in the market place.

16.6.2 Analysis of a Productivity Leader

In this subsection, we analyze the situation in which firm 1 is a "productivity leader." Here, we set $A_1 = 2$. Table 16.2 shows the relevant market analysis.

With respect to firm 1, its

- its profit, output, labor, and market shares are approximately two and one-half times that of its competitors, and
- its output is four times that of its competitors.

These trends persist as the number of firms grows. A sizable productivity advantage, therefore, maintains profitability even when there is increased competition.

16.7 Exercises

16.1. (An example of a constant returns-to-scale technology.) An economy consists of two producers and two consumers. Each producer converts raw material m into its output. The production functions for the firms are $Y_1 = 2m$ and $Y_2 = 3m$, respectively. Consumer 1 owns firm 1 and consumer 2 owns

Table 16.1. Market analysis when $A_1 = 0.5$.

$N = 3$	Competitive		Monopolistic Competition		Oligopoly	
	Firm 1	Other Firms	Firm 1	Other Firms	Firm 1	Other Firms
Labor	0.174	0.413	0.174	0.413	0.217	0.391
Output	0.123	0.493	0.123	0.493	0.147	0.472
Price	1.762	1.047	2.819	1.676	3.214	2.077
Revenue	0.217	0.516	0.347	0.826	0.474	0.981
Profit	0.043	0.103	0.174	0.413	0.257	0.589
Market Share	0.174	0.413	0.174	0.413	0.195	0.403
$N = 5$	Competitive		Monopolistic Competition		Oligopoly	
	Firm 1	Other Firms	Firm 1	Other Firms	Firm 1	Other Firms
Labor	0.095	0.226	0.095	0.226	0.106	0.223
Output	0.076	0.305	0.076	0.305	0.083	0.301
Price	1.562	0.929	2.499	1.486	2.662	1.643
Revenue	0.119	0.283	0.190	0.452	0.222	0.495
Profit	0.024	0.057	0.095	0.226	0.115	0.272
Market Share	0.095	0.226	0.095	0.226	0.101	0.225
$N = 10$	Competitive		Monopolistic Competition		Oligopoly	
	Firm 1	Other Firms	Firm 1	Other Firms	Firm 1	Other Firms
Labor	0.045	0.106	0.045	0.106	0.047	0.106
Output	0.042	0.166	0.042	0.166	0.043	0.166
Price	1.342	0.798	2.148	1.277	2.208	1.333
Revenue	0.056	0.133	0.089	0.212	0.095	0.221
Profit	0.011	0.026	0.045	0.106	0.049	0.115
Market Share	0.045	0.106	0.045	0.106	0.046	0.106

firm 2. Each consumer also owns 10 units of raw material. The utility functions for the consumers are $U_1(x_1, x_2) = x_1^{0.4} x_2^{0.6}$ and $U_2(x_1, x_2) = x_1^{0.5} x_2^{0.5}$, respectively. Determine the competitive equilibrium for this economy, namely, the market-clearing prices for each of the two goods; the profit-maximizing outputs, revenues, costs, and profits for each of the two firms; the incomes and quantities purchased of the two goods that maximizes each consumer's utility. (Take the price of the raw material as the *numeraire* and set its value to be 1.)

16.2. Answer Exercise 16.1 when the production functions exhibit decreasing returns-to-scale and are given by $Y_1 = 2\sqrt{m}$ and $Y_2 = 3\sqrt{m}$, respectively.

16.3. Consider the competitive economy examined in this chapter, except that consumer's utility is now given by the Cobb-Douglas function

$$U(x_1, \ldots, x_N) \;=\; \Pi_i \, x_i^{\beta_i}, \; 0 < \beta_i < 1, \; \sum_i \beta_i = 1.$$

Table 16.2. Market analysis when $A_1 = 2$.

$N = 3$	Competitive		Monopolistic Competition		Oligopoly	
	Firm 1	Other Firms	Firm 1	Other Firms	Firm 1	Other Firms
Labor	0.543	0.228	0.543	0.228	0.447	0.277
Output	1.227	0.307	1.227	0.307	1.050	0.358
Price	0.553	0.930	0.885	1.489	1.164	1.743
Revenue	0.679	0.285	1.086	0.457	1.222	0.623
Profit	0.136	0.057	0.543	0.228	0.775	0.347
Market Share	0.543	0.228	0.543	0.228	0.495	0.253
$N = 5$	Competitive		Monopolistic Competition		Oligopoly	
	Firm 1	Other Firms	Firm 1	Other Firms	Firm 1	Other Firms
Labor	0.373	0.157	0.373	0.157	0.322	0.169
Output	0.908	0.227	0.908	0.227	0.809	0.242
Price	0.513	0.863	0.821	1.381	0.957	1.505
Revenue	0.466	0.196	0.746	0.314	0.774	0.364
Profit	0.093	0.039	0.373	0.157	0.451	0.194
Market Share	0.373	0.157	0.373	0.157	0.347	0.163
$N = 10$	Competitive		Monopolistic Competition		Oligopoly	
	Firm 1	Other Firms	Firm 1	Other Firms	Firm 1	Other Firms
Labor	0.209	0.088	0.209	0.088	0.192	0.089
Output	0.572	0.143	0.572	0.143	0.534	0.145
Price	0.457	0.769	0.731	1.230	0.786	1.280
Revenue	0.261	0.110	0.418	0.176	0.420	0.186
Profit	0.052	0.022	0.209	0.088	0.228	0.096
Market Share	0.209	0.088	0.209	0.088	0.200	0.089

(a) Characterize the competitive equilibrium.
(b) Show that the competitive equilibrium coincides with the social planner's allocation.

16.4. (A more general version of Exercise 16.3.) In this problem, we examine a competitive economy with *non-identical* consumers and producers. The ownership shares of the consumers are also permitted to be different. The economy has two producers with respective production functions $\Phi_1(L_1) = 10L_1^{0.25}$ and $\Phi_2(L_2) = 20(L_2)^{2/3}$. There are three consumers in the economy with respective utility functions $U_1(x_1, x_2) = x_1^{0.8}x_2^{0.2}$, $U_2(x_1, x_2) = x_1^{0.5}x_2^{0.5}$ and $U_3(x_1, x_2) = x_1^{0.3}x_2^{0.7}$. The ownerships shares are as follows: consumer 1 owns 20% of firm 1 and 40% of firm 2, consumer 2 owns 50% of firm 1 and 10% of firm 2, and consumer 3 owns 30% of firm 1 and 50% of firm 2. Determine the competitive equilibrium for this economy, namely, the market-clearing prices for each of the three goods; the profit-maximizing outputs, revenues, costs,

and profits for each of the two firms; the incomes and quantities purchased of the three goods that maximizes each consumer's utility. (Take the wage rate as the *numeraire* and set its value to be 1.)

16.5. (The general version of Exercise 16.4.) The production function for producer i, $i = 1, 2, \ldots, N$, is

$$\Phi_i(L_i) \;=\; A_i L_i^{a_i}.$$

(In Exercise 16.4, $a_1 = 1/4$, $a_2 = 2/3$, $A_1 = 10$ and $A_2 = 20$.) Each of the M consumers has a (possibly different) Cobb-Douglas utility function given by

$$U_c(x_1, \ldots, x_N) \;=\; \Pi_i \; x_i^{\beta_i^c},\; 0 < \beta_i^c < 1,\; \sum_i \beta_i^c = 1$$

for consumer $c = 1, 2, \ldots, M$. (In Exercise 16.4, $\beta_1^1 = 0.8$, $\beta_2^1 = 0.5$ and $\beta_3^1 = 0.3$.) The ownership shares of the consumers are also permitted to be different, i.e., the ownership share of consumer c, $c = 1, 2, \ldots, M$, in firm i, $i = 1, 2, \ldots, N$, is T_i^c. (In Exercise 16.4, $(T_1^1, T_1^1) = (0.2, 0.4)$, $(T_2^2, T_2^2) = (0.5, 0.1)$, $(T_1^3, T_3^3) = (0.3, 0.5)$.) In what follows, take the wage rate as the *numeraire* and set its value to be 1.

(a) Define a set of equations (in terms of generic parameters) that can be solved to determine the competitive equilibrium.
(b) There are N market clearing conditions, one for each product market, and 1 market clearing condition for the labor market for a total of $N + 1$ equations. However, there are only N prices to be determined. Verify that the system of market clearing conditions contains a redundant equation.

16.6. (Open-ended computational exercise.) Analyze the competitive, monopolistic competition, and oligopoly market structures for a concrete numerical example of the economy analyzed in this chapter. Set the wage rate to be 1 and pick values for a and r so that $ar = 0.5$. Is it true that the allocations of labor are the same for the competitive and monopolistic competition market structures, but are different for the oligopoly market structure? If so, how much worse off is the consumer? How do the outputs change?

16.8 Bibliographical Notes

General equilibrium theory can be found in graduate-level microeconomic texts such as Varian [1992] and Mas-Collel et. al. [1995]. Starr [1997] provides an accessible treatment of all aspects to this topic. Debreu [1959] remains a classic on this subject. Discussion of market structures appears in industrial organization texts such as Tirole [1988] and Shy [1995] and the handbook edited by Schmalensee and Willig [1989].

16.9 Solutions to Exercises

16.1 Because the production functions are linear in m, the profit functions are also linear in m. (Since there is only one factor of production, linearity implies constant returns-to-scale.) That is,

$$\begin{aligned}\Pi(p_1, m_1) &:= p_1(2m_1) - m_1 = (2p_1 - 1)m_1,\\ \Pi(p_2, m_2) &:= p_2(3m_2) - m_2 = (3p_2 - 1)m_2.\end{aligned}$$

Profits cannot be infinite; in fact they must be zero. Thus, it must be the case that $p_1 = 0.5$ and $p_2 = 0.\bar{3}$. Since profits are zero, all income of each consumer derives from the value of their raw material, which is 10. Since each consumer has a Cobb-Douglas utility, their expenditures shares are given by the respective exponents. See (5.14), p. 75. That is, consumer 1 will spend 40% of his income or 4 on good 1 and 60% or 6 on good 2, whereas consumer 2 will spend 50% of his income or 5 on both goods 1 and 2. Thus, total expenditure by the consumers on good 1 is 9 and 11 on good 2. This immediately implies that firm 1's revenue is 9 and firm 2's revenue is 11. Since $p_1 = 0.5$ and $p_2 = 0.\bar{3}$, it follows that the output of firm 1 is 18 and the output of firm 2 is 33, and that consumer 1 purchases $4/0.5 = 8$ units of good 1 and $6/0.\bar{3} = 18$ units of good 2, whereas consumer 2 purchases $5/0.5 = 10$ units of good 1 and $5/0.\bar{3} = 15$ units of good 2.

16.2 In this problem the profit functions for each producer are now

$$\begin{aligned}\Pi(p_1, m_1) &:= p_1(2\sqrt{m_1}) - m_1,\\ \Pi(p_2, m_2) &:= p_2(3\sqrt{m_2}) - m_2.\end{aligned}$$

First-order optimality conditions imply that $(2p_1)/(2\sqrt{m_1}) = 1$ or $m_1 = p_1^2$ and $(3p_2)/(2\sqrt{m_2}) = 1$ or $m_2 = 2.25p_2^2$. Substituting these expressions back into the profit function, we have that $\Pi_1 = p_1^2$, $R_1 = 2p_1^2$, $C_1 = p_1^2$ and $\Pi_2 = 2.25p_2^2$, $R_2 = 4.5p_1^2$, $C_2 = 2.25p_2^2$, where R_i and C_i denote the revenue and cost, respectively, for firm $i = 1, 2$. Consumer 1's income is now $10 + p_1^2$ and consumer 2's income is now $10 + 2.25p_2^2$. Consumer 1 spends 40% of his income on good 1 and 60% on good 2, whereas consumer 2 spends 50% of his income on each good. Hence, the total expenditure on good 1 is

$$0.4(10 + p_1^2) + 0.5(10 + 2.25)p_2^2 = 0.4p_1^2 + 1.125p_2^2 + 9,$$

which must equal the revenue $2p_1^2$ of firm 1, and the total expenditure on good 2 is

$$0.6(10 + p_1^2) + 0.5(10 + 2.25)p_2^2 = 0.6p_1^2 + 1.125p_2^2 + 11,$$

which must equal the revenue $4.5p_2^2$ of firm 2. We conclude that

$$\begin{aligned}1.6p_1^2 - 1.125p_2^2 &= 9,\\ -0.6p_1^2 + 3.375p_2^2 &= 11.\end{aligned}$$

The solution is $p_1^2 = 9.0476$ and $p_2 = 4.8677$. The incomes for consumers 1 and 2 are 19.0476 and 20.9524, respectively. With the exponent of 0.5 for the production function and setting the price of the raw material to one, we have profit = cost = labor employed. Since the cost equals the total raw material of 20, profit equals 20, too, which further implies that total incomes (the sum of the value of raw material and profit) must equal 40, which it does. Table 16.3 provides the rest of the answers.

Table 16.3. Equilibrium results for Exercise 16.2.

	Firm 1	Firm 2
Output	6.0158	9.9283
Price	3.0079	2.2063
Revenue	18.0952	21.9048
Cost	9.0476	10.9524
Profit	9.0476	10.9524
Units purchased by consumer 1	2.5330	5.1800
Units purchased by consumer 2	3.4829	4.7483

16.3 (a) We shall take the wage rate as numeraire and set its value to 1. Equation (16.19) is still valid, and so each consumer's income equals $1/a$. For the Cobb-Douglas utility function, we know that the consumer's expenditure share on good i is β_i. See (5.14), p. 75. Hence, each consumer spends $\beta_i(1/a)$ on good i. Since there are M identical consumers, the total expenditure by all consumers on good i is $\beta_i(M/a)$. This total expenditure must equal the revenue, R_i, obtained by firm i, which by (16.16) equals $C_i/a = L_i/a$. (Keep in mind $w = 1$.) Thus, the total labor employed by firm i, L_i, equals $\beta_i M$. (Since consumers are identical, each producer employs $100\beta_i\%$ of each consumer's unit of labor.) This quantity must also equal $L_i(p, 1) = (p_i A_i a)^{1/(1-a)}$ given in (16.14). (Here, there are no quality-adjusted units.) Thus,

$$\beta_i M = (p_i A_i a)^{1/(1-a)},\ i = 1, 2, \ldots, N.$$

This can be used to solve for the unique set of market-clearing prices. Given that $L_i = \beta_i M$, it is straightforward to determine the profit-maximizing output, revenue, cost, and profit for each firm, as well as the units purchased by each (identical) consumer for each good (= aggregate output divided by M).

(b) We set $M = 1$. The social planner's optimization problem is

$$\max_{L \geq 0} \left\{ \prod_i (AL_i^a)^{\beta_i} : \sum_i L_i = 1 \right\}.$$

To solve for the optimal L_i, one may equivalently solve the optimization problem

$$\max_{L \geq 0} \left\{ \prod_i L_i^{\beta_i} : \sum_i L_i = 1 \right\}.$$

The optimal value equals $\Gamma_\Phi(p,1)$, where $\Phi(L_1,L_2,\ldots,L_N) = \prod_i L_i^{\beta_i}$ and $p = (1,1,\ldots,1) \in \mathbb{R}_+^N$. Since the optimal choice for the L_i must also be a cost-minimizing choice, and since the production function is Cobb-Douglas, the amount expended on labor input i must be equal to β_i. (This conclusion can be reached, of course, by directly using Lagrange multipliers.) And, this is the solution obtained in part (a).

16.4 Let R_i^c denote the expenditure on good i by consumer c. (It also equals the revenue obtained by firm i from consumer c.) Let $R_i = \sum_c R_i^c$ denote the revenue obtained by firm i (aggregating over all consumers). From (16.17), $\pi_i = (1-a_i)R_i$, and so $\pi_1 = (0.75)R_1$ and $\pi_2 = (0.\bar{3})R_2$. Since the consumer's utility function is Cobb-Douglas, the expenditure on good i by consumer c will equal β_i^c of the consumer's income I^c. Thus,

$$\begin{aligned}
R_1^1 &= (0.8)\{1 + (0.2)(0.75)R_1 + (0.4)(0.\bar{3})R_2\},\\
R_1^2 &= (0.5)\{1 + (0.5)(0.75)R_1 + (0.1)(0.\bar{3})R_2\},\\
R_1^3 &= (0.3)\{1 + (0.3)(0.75)R_1 + (0.5)(0.\bar{3})R_2\},\\
R_1 = \sum_c R_1^c &= 1.6 + 0.375R_1 + 0.17\bar{3}R_2,\\
R_2^1 &= (0.2)\{1 + (0.2)(0.75)R_1 + (0.4)(0.\bar{3})R_2\},\\
R_2^2 &= (0.5)\{1 + (0.5)(0.75)R_1 + (0.1)(0.\bar{3})R_2\},\\
R_2^3 &= (0.7)\{1 + (0.3)(0.75)R_1 + (0.5)(0.\bar{3})R_2\},\\
R_2 = \sum_c R_2^c &= 1.4 + 0.375R_1 + 0.\bar{3}R_2.
\end{aligned}$$

Thus,

$$\begin{aligned}
0.625R_1 - 0.17\bar{3}R_2 &= 1.6,\\
-0.375R_1 + 0.840R_2 &= 1.4,
\end{aligned}$$

from which we obtain $R_1 = 3.4493$ and $R_2 = 3.2065$. Using (16.16) and $w = 1$, we have $L_i = a_i R_i$, from which we compute the L_i. Given the L_i, we compute the outputs $Y_i = A_i L_i^{a_i}$ and the market-clearing prices $p_i = R_i/Y_i$. The revenues, costs, and profits are computed easily from these. Consumer incomes are 1.9449, 2.4004, and 2.3105, respectively. Table 16.4 provides the rest of the answers.

16.5 (a) We use the notation of the solution to Exercise 16.4. By the arguments provided there, we have

$$R_i^c = \beta_i^c \left\{ 1 + \sum_j T_j^c (1-a_j) R_j \right\},$$

Table 16.4. Equilibrium results for Exercise 16.4.

	Firm 1	Firm 2
Output	9.6364	33.1888
Price	0.3579	0.0966
Revenue	3.4493	3.2065
Cost	0.8623	2.1377
Profit	2.5869	1.0688
Units purchased by consumer 1	4.3469	4.0262
Units purchased by consumer 2	3.3530	12.4224
Units purchased by consumer 3	1.9365	16.7403

which implies that

$$R_i = \sum_c R_i^c = \sum_c \beta_i^c \; + \; \sum_j \left(\sum_c \beta_i^c T_j^c \right) (1 - a_j) R_j. \tag{16.72}$$

This is a *linear* system of N equations in the N unknowns, the R_i, which has a unique solution.

(b) By (16.16), $R_i = L_i / a_i$, which implies that $\sum_i a_i R_i = \sum_i L_i = M$. Now summing up both sides of (16.72) over i, we have

$$\begin{aligned}
\sum_i R_i &= \sum_i \left(\sum_c \beta_i^c \right) \; + \; \sum_i \left\{ \sum_j \left(\sum_c \beta_i^c T_j^c \right) (1 - a_j) R_j \right\} \\
&= \sum_c \left(\sum_i \beta_i^c \right) \; + \; \sum_j \left(\sum_i \sum_c \beta_i^c T_j^c \right) (1 - a_j) R_j \\
&= M \; + \; \sum_j \left(\sum_c \left(\sum_i \beta_i^c \right) T_j^c \right) (1 - a_j) R_j \\
&= M \; + \; \sum_j \left(\sum_c T_j^c \right) (1 - a_j) R_j \\
&= M \; + \; \sum_j (1 - a_j) R_j.
\end{aligned}$$

In this derivation we have used the fact that $\sum_i \beta_i^c = 1$ for each consumer c and that $\sum_c T_j^c = 1$ for each firm j. Subtracting $\sum_k R_k$ from both sides of the last equation yields $\sum_i a_i R_i = M$, a redundant equation.

Part IV

Engineering Models of Technology

17

Index-Based Dynamic Production Functions

An economist typically works with aggregate data that record the cumulative amounts of inputs and outputs in some predetermined period of time (e.g., quarterly, yearly). With today's information systems, detailed shop-floor data are becoming increasingly available, which opens the door to a refined description of technology.

At a micro-level, the exact shape of the input curve must be known to project realized output rates over time. Within an activity or stage of production, this dynamic input-output process is conveniently encapsulated by a **dynamic production function**

$$x = (x_1(\cdot), x_2(\cdot), \ldots, x_n(\cdot)) \quad \xrightarrow{f} \quad y = (y_1(\cdot), y_2(\cdot), \ldots, y_m(\cdot)).$$

Each $x_i(t)$ represents the quantity of input i at time t, and each $y_j(t)$ represents the quantity of output j realized at time t. A dynamic production function $f(\cdot)$ is a *functional*, since both its domain and range are vectors of functions, not vectors of numbers. A dynamic production function defines a recipe with more flexible elements than a steady-state production function.

We begin by describing a motivating example of non-instantaneous behavior. Next, we define the class of functions used to model all flows of goods and services. The simplest description of dynamic production assumes instantaneous transformations. This assumption can be relaxed to incorporate constant lead times. Index-based dynamic production functions will be used to model these processes, and three practical ways of indexing to incorporate constant lead times will be described.

17.1 A Motivating Example

Consider a production system that takes two time periods to transform a unit of input into a unit of output. In the first three time periods, system input was observed to be 24, 48 and 96 units, respectively. What would be your

answer to the following question: "How much total output has emerged by the end of the first, second, and third time periods?"

The production system description is (purposely) ambiguous, but we begin with an "obvious" answer, namely, no units of output will be realized in each of the first two periods and 24 units of output will be realized by the end of the third period. Even this simple answer makes the implicit assumption that there was *no* input in the two periods *prior* to the first period; otherwise, these two input numbers should be included as output in the first two periods. For simplicity, we shall assume that there was no input prior to the first period.

Further investigation reveals that the system operates one 8-hour shift per day, say from 9:00am to 5:00pm, and the first three time periods correspond to Monday through Wednesday. The process involves a heating (light manufacturing) operation that requires a cooling period of exactly 16 hours before the semi-finished part is completed as a finished product. (Another example is a painting operation that requires parts to dry.) All semi-finished parts are stored in a room while cooling, and this room is available 24 hours a day. Since the production process occurs round-the-clock, we define a time period to correspond to a single 24-hour day, say 9:00am to 9:00am.

The length of each time period is 24 hours. Let $x_i(\tau)$, $\tau \in [0, 24]$, denote the input curve in each period (day) $i = 1, 2, 3$. The total or cumulative input in each period is

$$x_i := \int_0^{24} x_i(\tau) d\tau.$$

The **shape** of the input in each period i is

$$s_i(\tau) := x_i(\tau)/x_i, \quad \tau \in [0, 24).$$

It represents normalized input in that $\int_0^{24} s_i(\tau) d\tau = 1$ for each i. We know that $x_1 = 24$, $x_2 = 48$ and $x_3 = 96$ and that $s_i(\tau) = 0$ if $8 \leq t \leq 24$ for each i, but, as yet, no further information about the shapes of the input curves are known. Let $Y(t)$ denote the cumulative output obtained due to input in the time interval $[0, t]$. (Here, $t \in [0, \infty)$ represents a point in time.) From what we have learned so far,

$$Y(8) = 0,\ Y(24) = Y(32) = 24,\ Y(48) = Y(56) = 72,\ Y(72) = 168.$$

Suppose competitive pressures and transportation lead times dictate that shipping occurs 24 hours a day. In particular, suppose a sizeable percentage of the shipping activity occurs during the third shift, 1:00am - 9:00am, each day. (A one-day time period consisting of three 8-hr shifts is consistent with the original description.) It is useful now to view the overall production process as consisting of three *stages* in series:

$$\text{light manufacturing} \longrightarrow \text{cooling} \longrightarrow \text{shipping}.$$

Shipping cannot ship a semi-finished part and so it will be necessary to determine $Y(t)$ for $t \in [16, 24)$, i.e., between 1am–9am. However, it will not be

possible to obtain these values *unless* the actual shapes, $s_i(\tau)$, of each input curve are known. If each $s_i(\tau)$ is "front-loaded," so that most input occurred early in each shift, then

$$Y(16) \approx Y(24) = 24,\ Y(40) \approx Y(48) = 72,\ Y(56) \approx Y(72) = 168.$$

If, on the other hand, each $s_i(\tau)$ is "back-loaded," so that most input occurred later in each shift, then

$$Y(0) \approx Y(16) = 0,\ Y(24) \approx Y(40) = 24,\ Y(48) \approx Y(56) = 72.$$

The shape (or distribution) of input over time can have an enormous effect on how output emerges over time.

At an atomic level, the production process, although relatively short, is not instantaneous. The follow-on cooling operation takes 16 hours, which is too significant to ignore. To simplify matters, we have conveniently assumed it took *exactly* 16 hours for each part to cool. Suppose all parts belong to a product family, each part is identical from the light manufacturing perspective, but parts require different times to cool with the maximum time to cool being 16 hours. To keep track of exact inventories of completed parts, the input curves associated with each part within the family must be known. If the number of parts in the product family is large and if customer demands exhibit a high degree of substitution, then one may wish to model only the aggregate input curve associated with all parts in the family and keep track of only the aggregate inventory of completed parts. For example, it may be computationally necessary to reduce the number of variables in a formulation used for planning purposes (when considering all of the product families and the many other aspects to the process). If this is indeed the case, then it will be appropriate to view aggregate output as emerging *continuously* over a 16-hour period, the distribution of which depends on how the aggregate input function disaggregates into components associated with each part in the family.

Output emerging continuously over time can also arise when modeling processing times that are random. For example, one possible outcome of a testing or inspection process is a failed part that requires rework. The testing or inspection time typically depends on the type of diagnosis. Estimating the output of failed parts over time is required for planning resource requirements for a rework activity.

17.2 Input-Output Domain

We describe the class of functions we use to model the flows of goods and services. Points in time are modeled by the interval $(-\infty, \infty)$. Unless otherwise stated, each function of time is (i) finite-valued and nonnegative, and (ii) has compact support, i.e., the points in time where the function is positive is contained in a closed and bounded interval of time. There are two fundamental types of functions of time, event-based and rate-based flows. We describe each below.

17.2.1 Event-Based Flows

For discrete-parts manufacturing systems, the flows of inputs and outputs are event-based at the microscopic level. An **event-based flow** $z(\cdot)$ associates a nonnegative real number $z(\tau)$ to an event that occurs at time τ. For example, $z(\tau)$ might be the quantity of parts initiated into a production process at time τ or a value associated with a job completed at time τ. Event-based flows only take on positive values at those times when events occur. We assume the number of event occurrences in a bounded interval of time is finite.

We shall integrate event-based functions, but cannot use the ordinary (Riemann) integral, since the integral of an event-based function is always zero. For an event-based function $z(\cdot)$ a summation takes the place of an integral. For instance, the integral of $z(\cdot)$ on the interval $(-\infty, t]$ simply adds up all the values associated with the events that occur on or before time t, namely,

$$Z(t) := \sum_{\tau_i \leq t} z(\tau_i). \tag{17.1}$$

The function $Z(\cdot)$ is a step-function whose "jumps" occur at the times τ_i. The integral of $z(\cdot)$ on the time interval $(s, t]$ is the difference $Z(t) - Z(s)$.

17.2.2 Rate-Based Flows

For a **rate-based flow** $z(\cdot)$, the nonnegative real number $z(\tau)$ represents the rate (quantity per unit time) of flow at time τ. Rate-based flows sometimes represent a fluid approximation to event-based flows, and also arise quite naturally when modeling physical processes. We shall insist a rate-based flow is *piecewise continuous*, and the number of its discontinuities in any bounded time interval is finite. The intervals between adjacent discontinuity points define the pieces on which the rate-based flow is continuous. We shall often refer to the *cumulative* flow associated with a rate-based flow $z(\cdot)$, which is defined as the (Riemann) integral

$$Z(t) := \int_{-\infty}^{t} z(\tau) d\tau.$$

While it is certainly possible to imagine a "mixed" flow, in the developments to follow, a flow is either rate-based or event-based. Let D denote the set of all functions of time that are either event- or rate-based. A dynamic production function is a map from $\mathcal{X} \subset \mathsf{D}^n$ into D^m.

17.3 Instantaneous Processes

A dynamic production function is *instantaneous* if the outputs at time t are solely a function of the inputs at time t, and possibly other exogenous information that is t-dependent. It has the form

$$y(t) = [f(x)](t) = \Phi(x(t), t) := (\Phi_1(x(t), t), \ldots, \Phi_m(x(t), t)).$$

Example 17.1. Consider a single-input, single-output process characterized by

$$y(t) = ax(t),$$

where a unit of input instantaneously results in a units of output. The constant a could be less than one to model yield loss common to industries such as semiconductor manufacture.

Example 17.2. Another example from productivity measurement is a single-input, single-output process characterized by

$$y(t) = A(t)\phi(x(t)) = A(0)e^{gt}x(t);$$

here, output at time t is proportional to the input rate at time t, and the proportionality constant changes over time to reflect productivity improvements.

17.4 Index-Based Processes

17.4.1 Definition

Typically, when there are several inputs (e.g., multiple materials, subassemblies, machine and labor services), there is definite linkage in their use, especially in discrete-parts manufacturing. For example, in an assembly process, there is a well-defined recipe for the number of parts or subassemblies needed to make a finished product. The following definition captures this notion.

Definition 17.3. *A dynamic production function is* **index-based** *if each input vector $x(\cdot)$ in its domain $\mathcal{X}$ has the form*

$$x(t) = (x_1(t), x_2(t), \ldots, x_n(t)) = ([\xi_1(z)](t), [\xi_2(z)](t), \ldots, [\xi_n(z)](t)),$$

and the corresponding output vector has the form

$$\begin{aligned}[f(x)](t) = y(t) &= (y_1(t), y_2(t), \ldots, y_n(t)) \\ &= ([\psi_1(z)](t), [\psi_2(z)](t), \ldots, [\psi_n(z)](t)),\end{aligned}$$

where $\xi_i : \mathsf{D} \to \mathsf{D}$ and $\psi_i : \mathsf{D} \to \mathsf{D}$ are each one-to-one. That is, the components of each input vector and resulting output vector are uniquely determined by a single function $z(\cdot)$ called the **index***; the shape of any one input or output curve completely determines the shape of all remaining input and output curves.*

It is, of course, possible to define processes with several indexes, but we shall not explore this generalization.

The computational advantage of an index-based process can be considerable, since the structure of the input-output transformation reduces to specifying $n + m$ independent transformations that often possess relatively simple forms.

17.4.2 Fixed Proportions, Instantaneous Model

The simplest, classic example of an index-based process is the straightforward extension of a simple Leontief process to the dynamic setting. We call it a **fixed proportions dynamic model**. The production process is instantaneous and the inputs and outputs are in constant proportions. The input, output vectors are, respectively,

$$x(t) = (a_1, a_2, \ldots, a_n)z(t), \tag{17.2}$$
$$y(t) = (u_1, u_2, \ldots, u_m)z(t). \tag{17.3}$$

The vectors $a = (a_1, a_2, \ldots, a_n)$ and $u = (u_1, u_2, \ldots, u_m)$ are the **technical coefficients** that characterize this technology. The index $z(\cdot)$ is called the **intensity** of this process.

Example 17.4. An example from semiconductor manufacturing illustrates the use for vector-valued output. In semiconductor wafer manufacturing, each wafer consists of many die. In an ideal world, all die on the wafer would have identical characteristics. Due to (random) fluctuations, die on a single wafer are not identical and must be classified into different "bins" based on key operating characteristics. In this setting, $z(\cdot)$ indexes the amount of wafer starts and the u^k represent the (expected) proportion of die that will be classified into bin k, after accounting for yield loss. See Leachman et. al. [1996] and Leachman [2002] for a detailed description.

More generally, the technical coefficients could be functions of time, namely, each $a_i = a_i(\cdot)$ and $u_j = u_j(\cdot)$, in which case the input and output vectors are, respectively,

$$x(t) = (a_1(t), a_2(t), \ldots, a_n(t))z(t), \tag{17.4}$$
$$y(t) = (u_1(t), u_2(t), \ldots, u_m(t))z(t). \tag{17.5}$$

For this more general description, the inputs and output are in constant proportions at each point in time; however, these proportional constants may vary over time.

Example 17.5. Technical coefficients can change over time due to productivity improvements. Required inputs per unit of intensity can decline due to learning, operational improvements, etc., and the outputs per unit of intensity can increase due to better yields.

17.4.3 Fixed Proportions, Constant Lead Time Models

We describe three practical ways to use indexing to incorporate constant lead times into the fixed proportions dynamic model characterized by (17.2) and (17.3).

Indexing Non-storable Services

Here, $z(\cdot)$ indexes the non-storable labor and machine services, which are simultaneously used to produce a single output. The storable inputs, such as materials and subassemblies, are assumed to be withdrawn from inventory "just-in-time" for their use, but each may require a constant lead time for transportation, inspection, etc.

Let $\ell_k \geq 0$ denote the lead time for the k^{th} storable input. In this setting we have $x_i(t) = a_i z(t)$ for the i^{th} non-storable service input, and $x_k(t) = a_k z(t + \ell_k)$ for the k^{th} storable input, since its withdrawal from inventory occurs exactly ℓ_k time units before its use. As in the motivating example of Section 17.1, we assume output emerges a constant lead time $\rho \geq 0$ after use of the non-storable services; consequently, $y(t) = z(t - \rho)$.

Example 17.6. A production process uses two raw (storable) materials and one machine service to produce a single output. The relevant data are:

- 12 units of material 1 are required per unit of output, and it takes 2 hours to transport this material to the machine station. Here, $a_1 = 12$ and $\ell_1 = 2$.
- 18 units of material 2 are required per unit of output, and it takes 3 hours to transport this material to the machine station. Here, $a_2 = 18$ and $\ell_2 = 3$.
- 2 hours of machine service are required per unit of output. Here, $a_3 = 2$.
- After machining has taken place, it takes 5 hours to inspect the semi-finished output, after which the completed output is available to service demand. Here, $\rho = 5$.

Between hours 100 and 102, a total of 32 hours of machine services has been consumed, uniformly spread over this period. Here, $x_3(t) = 16$ for $t \in [100, 102]$, and $z(t) = 16$ for $t \in [100, 102]$ since $x_3(t) = z(t)$. In words, 8 units are being machined at a constant rate during this two-hour period of time. Given a lead time of 2 hours for material 1 (and the just-in-time assumption), the withdrawal rate of material 1 input is $x_1(t) = 12(8) = 96$ for $t \in [98, 100]$. Similarly, given a lead time of 3 hours for material 2 (and the just-in-time assumption), the withdrawal rate of material 2 input is $x_2(t) = 18(8) = 144$ for $t \in [97, 99]$. Given that it takes 5 hours to inspect the semi-finished output as it emerges from the machining process, the final output rate is $y(t) = 8$ for $t \in [105, 107]$. In terms of this index,

$$\begin{aligned} z(t) &= 16 \text{ for } t \in [100, 102], \\ x_1(t) &= 6z(t + \ell_1) = 6z(t + 2), \\ x_2(t) &= 9z(t + \ell_2) = 9z(t + 3), \\ x_3(t) &= z(t), \\ y(t) &= 0.5z(t - \rho) = 0.5z(t - 5). \end{aligned}$$

Indexing "Outs"

A second approach to indexing is to let $z(\cdot)$ index the output or "outs" of the process in which case $y(t) = z(t)$. On the input side, $x_i(t) = a_i z(t+\rho)$ for the i^{th} non-storable service input, since its usage occurs exactly ρ time units before the product emerges as output, and $x_k(t) = a_k z(t+\ell_k+\rho)$ for the k^{th} storable input, since its withdrawal from inventory occurs exactly $\rho+\ell_k$ time units before the product emerges as output.

Example 17.7. Consider the same data provided in Example 17.6. Between hours 105 and 107, a total of 16 units of completed output has emerged, uniformly spread over this period. Here, $z(t) = 8$ for $t \in [105, 107]$. The function $z(\cdot)$ here does *not* equal the $z(\cdot)$ function in the previous example since here it is being used to index output and not machine services. The consumption of resources has not changed, i.e., it is still the case that

$$\begin{aligned} x_1(t) &= 12(8) = 96,\ t \in [98, 100], \\ x_2(t) &= 18(8) = 144,\ t \in [97, 99], \\ x_3(t) &= 16,\ t \in [100, 102]. \end{aligned}$$

What has changed, however, is how these functions relate to the chosen index. In terms of this index,

$$\begin{aligned} z(t) &= 8 \text{ for } t \in [105, 107], \\ x_1(t) &= 12z(t+\ell_1+\rho) = 12z(t+7), \\ x_2(t) &= 18z(t+\ell_2+\rho) = 18z(t+8), \\ x_3(t) &= 2z(t+\rho) = 2z(t+5), \\ y(t) &= z(t). \end{aligned}$$

Example 17.8. Consider the same data provided in Example 17.6, except that now *two* simultaneous semi-finished outputs emerge after machining in a 3:1 ratio, i.e., a total of 12 units of semi-finished output 1 and a total of 4 units of semi-finished output 2 emerge uniformly over the interval [100, 102]. In this example, it takes 5 hours to inspect output 1 (as before), whereas now it takes 9 hours to inspect output 2.

The notion of outs in this example cannot apply simultaneously to both outputs, since there are two nonequal ρ's, i.e., $\rho^1 = 5$ and $\rho^2 = 9$. It has to apply to one of the outputs, from which the other output and inputs can be determined. If output 1 is chosen as the index, then

$$\begin{aligned} z(t) &= 6,\ t \in [105, 107], \\ x_1(t) &= 16z(t+7),\ t \in [98, 100], \\ x_2(t) &= 24z(t+8),\ t \in [97, 99], \\ x_3(t) &= (8/3)z(t+5),\ t \in [100, 102], \\ y^1(t) &= z(t),\ t \in [105, 107], \\ y^2(t) &= (1/3)z(t-7),\ t \in [109, 111]. \end{aligned}$$

On the other hand, if output 2 is chosen as the index, then

$$\begin{aligned} z(t) &= 2,\ t \in [97, 99], \\ x_1(t) &= 48z(t),\ t \in [97, 99], \\ x_2(t) &= 72z(t-1),\ t \in [98, 100], \\ x_3(t) &= 8z(t-3),\ t \in [100, 102], \\ y^1(t) &= 3z(t+4),\ t \in [105, 107], \\ y^2(t) &= z(t),\ t \in [109, 111]. \end{aligned}$$

Indexing "Starts"

For the special case when the storable lead times are all identical, say $\ell_i = \ell$, then a third approach to indexing is to let $z(\cdot)$ index the "starts" of the process in which case $x_k(t) = a_k z(t)$ for the k^{th} storable input and $x_i(t) = a_i z(t-\ell)$ for the i^{th} non-storable service input, since the usage of the non-storable service occurs exactly ℓ units after the withdrawal of the storable inputs. On the output side, $y(t) = z(t-\ell-\rho)$.

Example 17.9. Suppose the data in Example 17.6 is changed so that $\ell_1 = \ell_2 = 2.5$ (the average of 2 and 3). The machine services are still consumed uniformly over the interval [100, 102]. Since the function $z(\cdot)$ now indexes starts,

$$\begin{aligned} z(t) &= 8 \text{ for } t \in [97.5, 99.5], \\ x_1(t) &= 12z(t),\ t \in [97.5, 99.5], \\ x_2(t) &= 18z(t),\ t \in [97.5, 99.5], \\ x_3(t) &= 2z(t-2.5),\ t \in [100, 102], \\ y(t) &= z(t-7.5),\ t \in [105, 107]. \end{aligned}$$

When the storable lead times are not all identical, then the notion of a start cannot apply to all inputs simultaneously—it has to apply to one of the inputs, from which the other inputs and output can be determined. Suppose the lead times $\ell_1 = 2$ and $\ell_2 = 3$, as before. If input 1 is chosen as the index, then

$$\begin{aligned} z(t) &= 96,\ t \in [98, 100], \\ x_1(t) &= z(t),\ t \in [98, 100], \\ x_2(t) &= 1.5z(t+1),\ t \in [97, 99], \\ x_3(t) &= (1/6)z(t-2),\ t \in [100, 102], \\ y(t) &= (1/12)z(t-7),\ t \in [105, 107]. \end{aligned}$$

On the other hand, if input 2 is chosen as the index, then

$$\begin{aligned} z(t) &= 144,\ t \in [97, 99], \\ x_1(t) &= (2/3)z(t-1),\ t \in [98, 100], \end{aligned}$$

$$x_2(t) = z(t),\ t \in [97, 99],$$
$$x_3(t) = (1/9)z(t-3),\ t \in [100, 102],$$
$$y(t) = (1/18)z(t-8),\ t \in [105, 107].$$

Remark 17.10. Under restrictive assumptions, the notion of starts and outs for the whole process can apply and can be used to index the consumption of non-storable resources, the withdrawal of storable inputs, and subsequent final output. In general, one can still speak of starts and outs and use them to index the process: in the case of starts, one of the inputs is chosen as the index; in the case of outs, one of the outputs is chosen as the index. When the consumption of non-storable resources is chosen as the index, all inputs and outputs can be related to it (as long as lead times are constant).

17.5 Exercises

17.1. Consider the data of Example 17.6. Suppose between the hours of 205 and 209, a total of 192 hours of machine services has been consumed, uniformly spread over this period. Ignore the previous input described in this example.

(a) What are the specific values for $z(\cdot)$, $x_1(\cdot)$, $x_2(\cdot)$, $x_3(\cdot)$ and $y(\cdot)$ if $z(\cdot)$ is used to index the non-storable services?
(b) What are the specific values for $z(\cdot)$, $x_1(\cdot)$, $x_2(\cdot)$, $x_3(\cdot)$ and $y(\cdot)$ if $z(\cdot)$ is used to index the outs?
(c) Suppose, as in Example 17.8, $\rho_1 = 5$ and $\rho_2 = 9$.
  (i) What are the specific values for $z(\cdot)$, $x_1(\cdot)$, $x_2(\cdot)$, $x_3(\cdot)$ and $y(\cdot)$ if $z(\cdot)$ is used to index output 1?
  (ii) What are the specific values for $z(\cdot)$, $x_1(\cdot)$, $x_2(\cdot)$, $x_3(\cdot)$ and $y(\cdot)$ if $z(\cdot)$ is used to index output 2?
(d) Suppose, as in Example 17.9, $\ell_1 = \ell_2 = 2.5$. What are the specific values for $z(\cdot)$, $x_1(\cdot)$, $x_2(\cdot)$, $x_3(\cdot)$ and $y(\cdot)$ if $z(\cdot)$ is used to index the starts?
(e) Suppose, as in Example 17.9, $\ell_1 = 2$ and $\ell_2 = 3$.
  (i) What are the specific values for $z(\cdot)$, $x_1(\cdot)$, $x_2(\cdot)$, $x_3(\cdot)$ and $y(\cdot)$ if $z(\cdot)$ is used to index input 1?
  (ii) What are the specific values for $z(\cdot)$, $x_1(\cdot)$, $x_2(\cdot)$, $x_3(\cdot)$ and $y(\cdot)$ if $z(\cdot)$ is used to index input 2?

17.2. A production process uses two raw (storable) materials and one machine service to produce two outputs. The relevant data are:

- Two simultaneous semi-finished outputs emerge after machining in a 2:1 ratio.
- 9 units of material 1 are required per unit of aggregate output, and it takes 1 hour to transport this material to the machine station. Here, $a_1 = 9$ and $\ell_1 = 1$.

- 27 units of material 2 are required per unit of aggregate output, and it takes 4 hours to transport this material to the machine station. Here, $a_2 = 27$ and $\ell_2 = 4$.
- 3 hours of machine service are required per unit of aggregate. Here, $a_3 = 3$.
- After machining has taken place, it takes 8 hours to inspect semi-finished output 1 and 12 hours to inspect semi-finished output 2. Here, $\rho_1 = 8$ and $\rho_2 = 12$.
- Between hours 100 and 110, a total of 900 hours of machine services has been consumed, uniformly spread over this period.

(a) What are the specific values for $x_1(\cdot)$, $x_2(\cdot)$, $x_3(\cdot)$, $y_1(\cdot)$ and $y_2(\cdot)$?
(b) What is the general form for the functions $x_1(\cdot)$, $x_2(\cdot)$, $x_3(\cdot)$, $y_1(\cdot)$ and $y_2(\cdot)$ in terms of the index $z(\cdot)$ when the index is chosen to represent the
(i) non-storable machine service?
(ii) output 1?
(iii) output 2?
(iv) input 1?
(v) input 2?

17.6 Bibliographical Notes

See Hackman [1990] for an in-depth discussion of acceptable properties of dynamic production functions.

17.7 Solutions to Exercises

17.1 (a) We have

$$\begin{aligned} z(t) &= 48,\ t \in [205, 209], \\ x_1(t) &= 288,\ t \in [203, 207], \\ x_2(t) &= 432,\ t \in [202, 206], \\ x_3(t) &= 48,\ t \in [205, 209], \\ y(t) &= 24,\ t \in [210, 214]. \end{aligned}$$

(b) All functions remain unchanged, except that now $z(t) = 24$, $t \in [210, 214]$.
(c) For (i),

$$\begin{aligned} z(t) &= 18,\ t \in [210, 214], \\ y_1(t) &= 18,\ t \in [210, 214], \\ y_2(t) &= 6,\ t \in [214, 218]. \end{aligned}$$

All other functions remain unchanged. As for (ii), now $z(t) = 6$, $t \in [210, 214]$. All other functions remain unchanged from their values in (i).
(d) Here,

$$\begin{aligned} z(t) &= 24,\ t \in [202.5, 206.5], \\ x_1(t) &= 288,\ t \in [202.5, 206.5], \\ x_2(t) &= 432,\ t \in [202.5, 206.5]. \end{aligned}$$

All other functions remain the same.
(e) Only the $z(\cdot)$ function changes. For part (i), $z(t) = 288$, $t \in [203, 207]$, whereas for part (ii), $z(t) = 432$, $t \in [202, 206]$.

17.2 (a) The specific values are:

$$\begin{aligned} x_1(t) &= 270,\ t \in [99, 109], \\ x_2(t) &= 810,\ t \in [96, 106], \\ x_3(t) &= 90,\ t \in [100, 110], \\ y_1(t) &= 20,\ t \in [108, 118], \\ y_2(t) &= 10,\ t \in [112, 122]. \end{aligned}$$

(b) For part (i):

$$\begin{aligned} x_1(t) &= 3z(t+1), \\ x_2(t) &= 9z(t+4), \\ x_3(t) &= z(t), \\ y_1(t) &= (2/9)z(t-8), \\ y_2(t) &= (1/9)z(t-12). \end{aligned}$$

For part (ii):

$$
\begin{aligned}
x_1(t) &= (27/2)z(t+9),\\
x_2(t) &= (81/2)z(t+12),\\
x_3(t) &= (9/2)z(t+8),\\
y_1(t) &= z(t),\\
y_2(t) &= (1/2)z(t-4).
\end{aligned}
$$

For part (iii):

$$
\begin{aligned}
x_1(t) &= 27z(t+13),\\
x_2(t) &= 81z(t+16),\\
x_3(t) &= 9z(t+12),\\
y_1(t) &= 2z(t+4),\\
y_2(t) &= z(t).
\end{aligned}
$$

For part (iv):

$$
\begin{aligned}
x_1(t) &= z(t),\\
x_2(t) &= 3z(t+3),\\
x_3(t) &= (1/3)z(t-1),\\
y_1(t) &= (2/27)z(t-9),\\
y_2(t) &= (1/27)z(t-13).
\end{aligned}
$$

For part (v):

$$
\begin{aligned}
x_1(t) &= 3z(t-3),\\
x_2(t) &= z(t),\\
x_3(t) &= (1/9)z(t-4),\\
y_1(t) &= (2/81)z(t-12),\\
y_2(t) &= (1/81)z(t-16).
\end{aligned}
$$

18

Distribution-Based Dynamic Production Functions

Distribution-based processes can be used to model the non-instantaneous behavior described in Section 17.1. To ease notational burdens, we shall confine our attention to single-input, single-output dynamic production functions. Accordingly, we drop the subscripts i and j and furthermore take $\Phi(z(\tau), \tau) = \phi(z(\tau), \tau) = z(\tau)$.

18.1 Description

18.1.1 Overview

Let the rate-based index $z(\tau)$ denote the rate of starts at time τ so that $z(\tau)\Delta\tau$ represents an input batch of starts in the interval $[\tau, \tau + \Delta\tau]$. Assume that an input batch at time τ will eventually result in a total of $\Phi_j(z(\tau), \tau)\Delta\tau$ units of output j, $1 \leq j \leq m$, and will consume a total of $\phi_i(z(\tau), \tau)\Delta\tau$ units of input i, $1 \leq i \leq n$.

Output realizations and input consumptions are not instantaneous but occur over time, as follows. The proportion of the total quantity of output j that will emerge in the next t' units of time is $W_j^y(t', \tau)$, and the proportion of the total input i consumed over the next t' units of time is $W_i^x(t', \tau)$. (These proportions are allowed to depend on τ.) Thus, for $t > \tau$, the input batch at time τ will result in

$$[\Phi_j(z(\tau), \tau)\Delta\tau]W_j^y(t - \tau, \tau)$$

units of output j by time t and will consume

$$[\phi_i(z(\tau), \tau)\Delta\tau]W_i^x(t - \tau, \tau)$$

units of input i by time t. Total outputs produced and total inputs consumed up to time t are the respective sums of the outputs produced and inputs consumed due to all such input batches up to time t. They are given, respectively, as

$$\sum_{\tau} [\Phi_j(z(\tau), \tau)\Delta\tau] W_j^y(t-\tau, \tau), \tag{18.1}$$

$$\sum_{\tau} [\phi_i(z(\tau), \tau)\Delta\tau] W_i^x(t-\tau, \tau), \tag{18.2}$$

where the sum is over $\tau = k\Delta\tau,\ k = \ldots, -2, -1, 0, 1, 2, \ldots,\ k \le t/\Delta\tau$.

The limits of (18.1) and (18.2) as $\Delta\tau \to 0$, assuming they exist, are

$$Y_j(t) := \int_{-\infty}^{t} \Phi_j(z(\tau), \tau) W_j^y(t-\tau, \tau) d\tau, \tag{18.3}$$

$$X_i(t) := \int_{-\infty}^{t} \phi_i(z(\tau), \tau) W_i^x(t-\tau, \tau) d\tau, \tag{18.4}$$

where $Y_j(t)$ and $X_i(t)$ denote, respectively, the cumulative amount of output j and input i up to time t.

When the index $z(\cdot)$ is event-based, $z(\tau)$ denotes a batch of starts at time τ that will result in a total of $\Phi_j(z(\tau), \tau)$ units of output j and consume a total of $\phi_i(z(\tau), \tau)$ units of input i. The integrals (18.3) and (18.4) are replaced by the corresponding summations

$$Y_j(t) := \sum_{\tau_k \le t} \Phi_j(z(\tau_k), \tau_k) W_j^y(t-\tau_k, \tau_k), \tag{18.5}$$

$$X_i(t) := \sum_{\tau_k \le t} \phi_i(z(\tau_k), \tau_k) W_i^x(t-\tau_k, \tau_k). \tag{18.6}$$

18.1.2 Definition

The functions $\Phi_j(\cdot)$ and $\phi_i(\cdot)$ are called **transforms**. Given $z(\cdot)$ and a transform $\xi : R_+^2 \to R_+$ the function of time $\xi(z(\cdot),\ \cdot)$ is assumed to preserve the flow type of the index $z(\cdot)$. Each function W_j^y, $1 \le j \le m$, and W_i^x, $1 \le i \le n$, is called a **cumulative lead time distribution**, which will be formally defined below.

Definition 18.1. *A dynamic production function is* **distribution-based** *if there are cumulative lead time distributions $W_j^y(\cdot)$ and $W_i^x(\cdot)$ and transforms $\Phi_j(\cdot)$ and $\phi_i(\cdot)$ such that the cumulative output and input functions have the forms (18.3) and (18.4) when the index $z(\cdot)$ is rate-based or have the forms (18.5) and (18.6) when the index $z(\cdot)$ is event-based.*

Example 18.2. Consider the transforms

$$\Phi_j(z(\tau), \tau) = \alpha_j(\tau) z(\tau),\ \phi_i(z(\tau), \tau) = \beta_i(\tau) z(\tau). \tag{18.7}$$

It follows from (18.3)-(18.6) that the resulting production functionals $Y_j := F_j(z)$, $X_i := \Xi_i(z)$ are each *linear* in $z(\cdot)$. That is, for nonnegative scalars c_1 and c_2,

$$F_j(c_1z_1 + c_2z_2) = c_1F_j(z_1) + c_2F_j(z_2),$$
$$\Xi_i(c_1z_1 + c_2z_2) = c_1\Xi_i(z_1) + c_2\Xi_i(z_2).$$

The linearity of $F_j(\cdot)$ and $\Xi_i(\cdot)$ with respect to $z(\cdot)$ is *not* synonymous with linearity of the output $Y_j(\cdot)$ and input $X_i(\cdot)$ functions of *time* given $z(\cdot)$. Linear production functionals are, in effect, the continuous-time extension of the basic nonparametric steady-state models we developed and analyzed in Parts I and II.

Definition 18.3. *A distribution-based dynamic production function is* **linear** *if its transforms satisfy (18.7).*

18.1.3 Lead Time Density

Loosely speaking, a lead time density is the derivative of a cumulative lead time distribution. A lead time density is either rate- or event-based. A rate-based (event-based) lead time density can be viewed as a family $\{w(\cdot,\tau),\ \tau \geq 0\}$ of probability density (mass) functions on $\mathbb{R}_+$.

Below, additional properties are imposed to ensure $w(\cdot,\tau)$ is close to $w(\cdot,\tau')$ when τ is sufficiently close to τ', which will be sufficient to ensure the integrals (18.3)-(18.6) exist. These technical properties and the remarks that follow can be skipped on first reading.

Definition 18.4. *A* **rectangular partition of** $\mathbb{R}_+^2$ *is a collection of disjoint rectangles whose union is* $\mathbb{R}_+^2$ *such that each rectangle in the partition has positive area and the number of rectangles in the partition that intersect any compact (i.e., bounded and closed)* $C \subset \mathbb{R}_+^2$ *is finite.*

Definition 18.5. *A function* $w : \mathbb{R}_+^2 \to \mathbb{R}_+$ *is a* **rate-based lead time density** *if it satisfies the following properties:*

(i) $w(\cdot,\tau)$ *is piecewise continuous and* $\int_0^\infty w(t,\tau)dt = 1$ *for each* $\tau \geq 0$.
(ii) There is a rectangular partition of $\mathbb{R}_+^2$ *such that* $w(\cdot)$ *is continuous on each rectangle in the partition.*

Definition 18.6. A function $w : \mathbb{R}_+^2 \to \mathbb{R}_+$ is an **event-based lead time density** if it satisfies the following properties:

(i) $w(\cdot,\tau)$ is event-based and $\sum_{t_i} w(t_i,\tau) = 1$ for each $\tau \geq 0$.
(ii) The set of occurrence times of all events of the family of functions $\{w(\cdot,\tau),\ \tau \geq 0\}$ in any bounded time interval is finite.
(iii) $w(t,\cdot)$ is piecewise continuous for each $t \geq 0$.

For a lead time density $w(\cdot,\cdot)$, we denote its integral on $[0,u]$ by

$$W(u,\tau) := \begin{cases} \int_0^u w(t,\tau)dt & \text{if } w(\cdot) \text{ is rate-based,} \\ \sum_{0\leq t_i\leq u} w(t_i,\tau) & \text{if } w(\cdot) \text{ is event-based.} \end{cases}$$

This is the cumulative lead time distribution corresponding to $w(\cdot,\cdot)$.

Remark 18.7. When $w(\cdot,\tau)$ is rate-based, $W(\cdot,\tau)$ is differentiable everywhere except at the discontinuity points of $w(\cdot,\tau)$ (of which there are not too many), and when $w(\cdot,\tau)$ is event-based, $W(\cdot,\tau)$ is piecewise continuous (more specifically, it is right-continuous with continuous left limits).

The technical conditions (ii) of (18.5) and (ii) and (iii) of (18.6) are automatically satisfied for a piecewise constant lead time density, a special but still quite general case.

Definition 18.8. *A lead time density is* **piecewise constant** *if there is a partition of* $\mathbb{R}_+$ *into nonempty intervals for which the subfamily of functions* $\{w(\cdot,\tau),\ \tau \in I\}$ *are all identical for each interval* I *in the partition. The number of intervals that intersect any bounded interval is finite. When the family of functions* $\{w(\cdot,\tau),\ \tau \geq 0\}$ *are all identical the lead time density is* **constant**.

For most of our numerical examples, we assume a constant lead time density $w(\cdot,\cdot)$, and to ease notational burdens we shall, with a slight abuse of notation, suppress the irrelevant τ, and simply refer to $w(u)$ and its integral $W(u)$ on the interval $[0,u]$. For a constant lead time density, the function $W(t-\tau)$ of τ will be piecewise continuous.

18.1.4 Technical Remarks

Remark 18.9. Fix a lead time density $w(\cdot,\cdot)$. For a fixed $t \geq 0$, define

$$\theta_w(\tau) := \begin{cases} W(t-\tau,\tau), & \text{if } \tau \in [0,t], \\ 0, & \text{if } \tau > t. \end{cases}$$

Condition (ii) of (18.5) and conditions (ii) and (iii) of (18.6) ensure that $\theta_w(\tau)$ is piecewise continuous, which guarantees (along with the compact support) that the integrals (18.3)-(18.6) exist. When $w(\cdot,\cdot)$ is event-based, $\theta_w(\tau)$ will be a step-function whose jumps consist of all $\tau \in [0,t]$ for which τ is a jump of the function $w(t,\cdot)$. If no conditions are imposed on the relationship between the $w(\cdot,\tau)$'s, then $\theta_w(\tau)$ could be pathological; in particular, when $z(\cdot)$ is rate-based the integrals (18.3)-(18.6) may not exist. For example, when $w(\cdot,\cdot)$ is event-based, depending on the spatial arrangement of all times corresponding to all events, the function $\theta_w(\tau)$ can be event-based, rate-based, or even mixed.

Remark 18.10. When $w^y(\cdot)$, $w^x(\cdot)$ and $z(\cdot)$ are all rate-based both a generic output function $Y(t)$ and input function $X(t)$ will be differentiable everywhere except at a finite number of points. At the differentiable points, it may be shown that the output and input *rates* are given by

$$y(t) = \frac{dY}{dt} = \int_{-\infty}^{t} \Phi(z(\tau),\tau) w^y(t-\tau,\tau) d\tau, \tag{18.8}$$

$$x(t) = \frac{dX}{dt} = \int_{-\infty}^{t} \Phi(z(\tau),\tau) w^x(t-\tau,\tau) d\tau. \tag{18.9}$$

When $z(\cdot)$ is rate-based and $w^y(\cdot)$ is event-based, $y(t)$ can be rate- or event-based depending once again on the spatial arrangement of all times corresponding to all events. For example, in Chapter 18 we shall show that when $z(\cdot)$ is rate-based and $w^y(\cdot)$ is event-based and constant, then $y(t)$ is

$$\sum_{\tau_j} \Phi(z(t-\tau_j), t-\tau_j) w^y(\tau_j), \tag{18.10}$$

where the summation is over all τ_j such that $t-\tau_j$ is an occurrence time of an event.

18.2 Constant Lead Time Processes

18.2.1 Description

Suppose ℓ is a positive constant lead time and $W(t) = 1$ if $t \geq \ell$ and 0 otherwise. For this choice of $W(\cdot)$, the cumulative output in (18.3) is

$$Y(t) = \int_{-\infty}^{t-\ell} z(\tau) d\tau = Z(t-\ell).$$

Here we assume that no input was consumed prior to time 0 so that $Y(t) = 0$ if $t \leq \ell$. The output curve

$$y(t) = z(t-\ell),$$

is an instantaneous process shifted by the *constant* lead time ℓ.

Example 18.11. Suppose $z(t) = 200t,\ t \in [0,1]$, and the lead time is constant at $\ell = 0.5$. Then:

$$Z(t) = \begin{cases} 100t^2, & 0 \leq t \leq 1, \\ 100, & 1 \leq t. \end{cases}$$

$$Y(t) = \begin{cases} 0, & 0 \leq t \leq 0.5, \\ 100(t-0.5)^2, & 0.5 \leq t \leq 1.5, \\ 100, & 1.5 \leq t. \end{cases}$$

Consider now constant, event-based lead time densities with more than one event. Specifically, there are positive constants $0 < \ell_1 < \ldots < \ell_n$ and $w_1, w_2, \ldots, w_n$ for which $\sum_{j=1}^n w_j = 1$ and

$$W(u) = \begin{cases} 0, & \text{if } u < \ell_1, \\ \sum_{j=1}^i w_j, & \text{if } u \in [\ell_i, \ell_{i+1}],\ 1 \leq i \leq n-1, \\ 1, & \text{if } u \geq \ell_n. \end{cases}$$

Fix a $t > \ell_n$. For this choice of $W(\cdot)$, the output in (18.3) on p. 310 is

$$\begin{aligned} Y(t) &= Z(t-\ell_n) + [Z(t-\ell_{n-1}) - Z(t-\ell_n)](1-W_1) \\ &\quad + [Z(t-\ell_{n-2}) - Z(t-\ell_{n-1})](1-W_2) \\ &\quad + \ldots + [Z(t-\ell_1) - Z(t-\ell_2)](1-W_{n-1}) \\ &= \sum_{j=1}^{n} w_j Z(t-\ell_j), \end{aligned} \tag{18.11}$$

where $W_i := \sum_{j=1}^{i} w_j$, $i = 1, 2, \ldots, n$, denotes the partial sums. The output rate is

$$y(t) = \sum_{j=1}^{n} w_j z(t-\ell_j), \tag{18.12}$$

which is a *convex combination* of input rates at various times *prior* to time t and is consistent with (18.10). A probabilistic interpretation of (18.12) will be provided in Chapter 20.

18.2.2 Integer Lead Times

Suppose there are three events corresponding to $\ell_1 = 2$, $\ell_2 = 5$, and $\ell_3 = 8$ and $w_1 = 0.25$, $w_2 = 0.70$, and $w_3 = 0.05$. The interpretation is that 25% of the time an entering part will emerge as output in exactly 2 units of time, 70% of the time an entering part will emerge as output in exactly 5 units of time, and 5% of the time an entering part will emerge as output in exactly 8 units of time. Assume that each unit of input produces a unit of output. Fix the input curve as

$$z(t) = 100 \cdot 1_{[0,1]}(t) + 200 \cdot 1_{[1,2]}(t) + 300 \cdot 1_{[2,3]}(t) + 400 \cdot 1_{[3,4]}(t),$$

so that no input is consumed prior to time 0 and 100, 200, 300, and 400 units are started uniformly within the periods 1, 2, 3 and 4, respectively. The output curve is

$$y(t) = 0.25z(t-2) + 0.70z(t-5) + 0.05z(t-8). \tag{18.13}$$

As a result of this input—and the fact that the longest lead time is 8—output will emerge over the first 12 intervals of time $[t-1, t]$, $1 \le t \le 12$. Let

$$\begin{aligned} z_t &:= \int_{t-1}^{t} z(\tau) d\tau, \\ y_t &:= \int_{t-1}^{t} y(\tau) d\tau, \end{aligned}$$

denote, respectively, the cumulative input and cumulative output consumed over the interval $[t-1, t]$ for positive integers t. In the example, $z_t = 100t$, $t = 1, 2, 3, 4$. It follows from (18.13) that

$$\begin{aligned} y_t &= 0.25 \int_{t-1}^{t} z(\tau - 2)d\tau + 0.70 \int_{t-1}^{t} z(\tau - 5)d\tau + 0.05 \int_{t-1}^{t} z(\tau - 8)d\tau \\ &= 0.25z_{t-2} + 0.70z_{t-5} + 0.05z_{t-8}, \quad t = 1, 2, \ldots. \end{aligned}$$

(Keep in mind $z(t) = 0$ for $t \leq 0$.) In matrix form, the cumulative outputs in each of the first 12 intervals $[t-1, t]$, $1 \leq t \leq 12$, are:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \\ y_{11} \\ y_{12} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0.25 & 0 & 0 & 0 \\ 0 & 0.25 & 0 & 0 \\ 0 & 0 & 0.25 & 0 \\ 0.70 & 0 & 0 & 0.25 \\ 0 & 0.70 & 0 & 0 \\ 0 & 0 & 0.70 & 0 \\ 0.05 & 0 & 0 & 0.70 \\ 0 & 0.05 & 0 & 0 \\ 0 & 0 & 0.05 & 0 \\ 0 & 0 & 0 & 0.05 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 25 \\ 50 \\ 75 \\ 170 \\ 140 \\ 210 \\ 285 \\ 10 \\ 15 \\ 20 \end{bmatrix}$$

Note that $\sum_{t=1}^{12} y_t = 1000 = \sum_{t=1}^{12} z_t$, as it should.

Suppose $t = 23$. All parts that enter by time 15 will emerge by time 23, 95% of the parts that enter between times 15 and 18 will emerge by time 23, and 25% of the parts that enter between times 18 and 21 will emerge by time 23. No part that enters between time 21 and 23 will emerge as output by time 23. Thus,

$$\begin{aligned} Y(23) &= Z(15) + 0.95[Z(18) - Z(15)] + 0.25[Z(21) - Z(18)] \\ &= 0.05Z(15) + 0.70Z(18) + 0.25Z(21). \end{aligned}$$

This is just the integral over $[0, 23]$ of both sides of (18.13).

18.2.3 Noninteger Lead Times

Consider the data provided in Section 18.2.2, except that the lead times for the three events are now $\ell_1 = 2.21$, $\ell_2 = 5.63$, and $\ell_3 = 8.42$. The output curve is

$$y(t) = 0.25z(t - 2.21) + 0.70z(t - 5.63) + 0.05z(t - 8.42).$$

It follows once again from (18.13) that

$$y_t = 0.25\int_{t-1}^{t} z(\tau-2.21)d\tau + 0.70\int_{t-1}^{t} z(\tau-5.63)d\tau + 0.05\int_{t-1}^{t} z(\tau-8.42)d\tau. \tag{18.14}$$

When the lead times are noninteger, the integrals in (18.14) can be calculated, as follows. For the first integral

$$\begin{aligned}\int_{t-1}^{t} z(\tau-2.21)d\tau &= \int_{t-1}^{(t-1)+0.21} z(\tau-2.21)d\tau + \int_{(t-1)+0.21}^{t} z(\tau-2.21)d\tau \\ &= \int_{t-1}^{(t-1)+0.21} z_{t-3}d\tau + \int_{(t-1)+0.21}^{t} z_{t-2}d\tau \\ &= 0.21z_{t-3} + 0.79z_{t-2}.\end{aligned}$$

A similar calculation yields

$$\begin{aligned}\int_{t-1}^{t} z(\tau-5.63)d\tau &= 0.63z_{t-6} + 0.37z_{t-5}, \\ \int_{t-1}^{t} z(\tau-8.42)d\tau &= 0.42z_{t-9} + 0.58z_{t-8},\end{aligned}$$

from which we obtain

$$\begin{aligned}y_t &= 0.25\{0.21z_{t-3} + 0.79z_{t-2}\} + 0.70\{0.63z_{t-6} + 0.37z_{t-5}\} \\ &\quad + 0.05\{0.42z_{t-9} + 0.58z_{t-8}\} \\ &= 0.021z_{t-9} + 0.029z_{t-8} + 0.441z_{t-6} + 0.259z_{t-5} \\ &\quad + 0.1975z_{t-3} + 0.0525z_{t-2}.\end{aligned}$$

The y_t are also a linear combination of the z_t. In matrix form, the cumulative outputs within each of the first 12 intervals $[t-1, t]$, $1 \le t \le 12$, are:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \\ y_{11} \\ y_{12} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0.1975 & 0 & 0 & 0 \\ 0.0525 & 0.1975 & 0 & 0 \\ 0 & 0.0525 & 0.1975 & 0 \\ 0.259 & 0 & 0.0525 & 0.1975 \\ 0.441 & 0.259 & 0 & 0.0525 \\ 0 & 0.441 & 0.259 & 0 \\ 0.029 & 0 & 0.441 & 0.259 \\ 0.021 & 0.029 & 0 & 0.441 \\ 0 & 0.021 & 0.029 & 0 \\ 0 & 0 & 0.021 & 0.029 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 19.75 \\ 44.75 \\ 69.75 \\ 120.65 \\ 116.90 \\ 165.90 \\ 238.80 \\ 184.30 \\ 12.90 \\ 17.90 \end{bmatrix}$$

The cumulative output up to time $t = 12$, $\sum_{t=1}^{12} y_t = 991.6$, equals the cumulative input up to time $t = 12$, $\sum_{t=1}^{12} z_t$, *less* 8.4. The gap between the

cumulative output and input is a result of the fact that 5% of the starts from time 3.58 to 4, i.e., $0.05\{(4 - 3.58)400\} = (0.021)400 = 8.4$, will not yet be realized as output by time 12. Another way to see this is that the sum of the elements in each column equals one, except that last one, which equals $0.979 = 1 - 0.021$.

Remark 18.12. If the column lengths extend to infinity, then the sum of the elements in each column *must* add to one.

Remark 18.13. In general, when $z(\cdot)$ is rate-based, constant over each interval $(t-1, t)$, then

$$\int_{t-1}^{t} z(\tau - \ell)d\tau = (\ell - \ell^-)z_{t-\ell^+} + (\ell^+ - \ell)z_{t-\ell^-} \ .$$

Here, ℓ^+ is the smallest integer no smaller than ℓ and ℓ^- is the largest integer no larger than ℓ. The expression $(\ell - \ell^-)$ is just the fractional part of the lead time ℓ and $(\ell^+ - \ell)$ is the remainder from one. For example, when $\ell = 5.63$, $\ell^- = 5$ and $\ell^+ = 6$ and so $(\ell - \ell^-) = 0.63$ and $(\ell^+ - \ell) = 0.37$.

Remark 18.14. There are two common proposals to model noninteger lead times. Suppose, for example, the natural lead time is 14 hours but the natural period of time is 8 hours, which represents a shift. Measured in units of shifts, the lead time is 1.75. One obvious fix is to simply subdivide each 8 hour block into four two-hour blocks. This proposal, however, will increase the number of variables and constraints by a factor of four. If the lead time were 13 hours, then hourly time periods will be used, thereby increasing the variables and constraints by a factor of 8. (In the example, if time is broken up in intervals of length 0.01 instead of 1, then the lead times will be 221, 563 and 842 units of time. The number of variables and constraints will increase by a factor of 100!) There is a second fundamental problem with this approach. On such a small scale, the variation of input rates over time it permits can be impractical on the shop floor. Constraints could be added to limit such variability, but again at the expense of complicating the model and increasing the computational burdens.

A second proposal is to round the lead time up to the nearest integer, in this case 1.75 becomes 2. This builds a buffer to ensure that promised inventory will be available to fulfill demand. Inflating lead times is not an uncommon practice of implementing *Manufacturing Resources Planning (MRP)* systems. Unfortunately, this proposal creates unnecessary work-in-process in the system. (If the difference between the actual and rounded-up lead times are sufficiently close, then this may be not be too costly.)

The example of this subsection shows that neither common proposal is necessary. After appropriate pre-processing, it will be *unnecessary* to complicate the model and increase computational burdens to accommodate noninteger lead times.

18.2.4 Non-Integer Lead Times with Unequal Length Periods

In this section, we fix time points, $t_1, t_2, \ldots$, not necessarily integer, such that the (rate-based) input curve is constant on each interval of time (t_{i-1}, t_i). The lead times ℓ_k need not be integer, either. Using unequal length intervals of time in a model is quite practical, especially if the model is intended to guide relatively short-term actions. For example, certain days may run two shifts, whereas other days may run only one shift. Natural time periods used for tactical planning models such as weeks may differ in capacity due to vacations, holidays, or planned shutdowns.

We illustrate the computations by modifying the data provided in Section 18.2.3, as follows. The constant rate on the interval $(0, 0.70)$ is 100, the constant rate on the interval $(0.70, 1.65)$ is $2.55.\bar{5}$, and the constant rate on the interval $(1.65, 4)$ is $327.\bar{6}$. The cumulative input curve is

$$Z(t) = \begin{cases} 100t, & 0 \le t \le 0.70, \\ 168.42t - 47.89, & 0.70 \le t \le 1.65, \\ 327.\bar{6}t - 310.65, & 1.65 \le t \le 4, \\ 1000, & 4 \le t. \end{cases}$$

These numbers were constructed so that $Z(1) = 100$, $Z(2) = 300$ and $Z(4) = 1000$, as in Section 18.2.3. We shall let z_1, z_2, and z_3 denote the constant rates within each period of time. Keep in mind that

(a) the periods no longer correspond to the intervals (0, 1), (1, 2), and (2, 3)—the period lengths here are 0.70, 0.95, and 2.35, respectively;
(b) $Z(0.7) = 0.70z_1 = 70$, $Z(1.65) = 0.70z_1 + 0.95z_2 = 230$, and $Z(4) = 0.70z_1 + 0.95z_2 + 2.35z_3 = 1000$; and,
(c) the slopes of the piecewise linear, increasing $Z(\cdot)$ curve for the three segments are $z_1 = 100$, $z_2 = 168.42$, and $z_3 = 327.\bar{6}$, respectively.

Once again, the cumulative output curve (18.11) is

$$Y(t) = 0.25Z(t - 2.21) + 0.70Z(t - 5.63) + 0.05Z(t - 8.42).$$

The computation of $Y(t)$ for integer t, $1 \le t \le 12$, requires the computation of the following $Z(\cdot)$ values:

$$Z(0.79) = Z(0.70) + (0.79 - 0.70)z_2 = 0.70z_1 + 0.09z_2 = 85.16.$$
$$Z(1.79) = Z(1.65) + (1.79 - 1.65)z_2 = 0.70z_1 + 0.95z_2 + 0.14z_3 = 275.87.$$
$$Z(2.79) = Z(1.65) + (2.79 - 1.65)z_2 = 0.70z_1 + 0.95z_2 + 1.14z_3 = 603.54.$$
$$Z(3.79) = Z(1.65) + (3.79 - 1.65)z_2 = 0.70z_1 + 0.95z_2 + 2.14z_3 = 931.21.$$
$$Z(4.79) = Z(5.79) = \cdots = Z(11.79) = 0.70z_1 + 0.95z_2 + 2.35z_3 = 1000.$$

$$Z(0.37) = 0.37z_1 = 37.$$
$$Z(1.37) = Z(0.70) + (1.37 - 0.70)z_2 = 0.70z_1 + 0.67z_2 = 182.84.$$

$$Z(2.37) = Z(1.65) + (2.37 - 1.65)z_2 = 0.70z_1 + 0.95z_2 + 0.72z_3 = 465.92.$$
$$Z(3.37) = Z(1.65) + (3.37 - 1.65)z_2 = 0.70z_1 + 0.95z_2 + 1.72z_3 = 793.59.$$
$$Z(4.37) = Z(5.37) = \cdots = Z(11.37) = 0.70z_1 + 0.95z_2 + 2.35z_3 = 1000.$$

$$Z(0.58) = 0.58z_1 = 58.$$
$$Z(1.58) = Z(0.70) + (1.58 - 0.70)z_2 = 0.70z_1 + 0.88z_2 = 218.21.$$
$$Z(2.58) = Z(1.65) + (2.58 - 1.65)z_2 = 0.70z_1 + 0.95z_2 + 0.93z_3 = 534.73.$$
$$Z(3.58) = Z(1.65) + (3.58 - 1.65)z_2 = 0.70z_1 + 0.95z_2 + 1.93z_3 = 862.40.$$
$$Z(4.58) = Z(5.58) = \cdots = Z(11.58) = 0.70z_1 + 0.95z_2 + 2.35z_3 = 1000.$$

Given the expressions for these $Z(\cdot)$ values, it is straightforward to determine expressions for the $Y(t)$ and y_t. For example,

$$\begin{aligned}
Y(6) &= 0.25Z(3.79) + 0.70Z(0.37) \\
&= 0.25[0.70z_1 + 0.95z_2 + 2.14z_3] + 0.70[0.37z_1] \\
&= 0.434z_1 + 0.2375z_2 + 0.535z_3 = 258.70, \\
Y(7) &= 0.25Z(4.79) + 0.70Z(1.37) \\
&= 0.25[0.70z_1 + 0.95z_2 + 2.35z_3] + 0.70[0.70z_1 + 0.67z_2] \\
&= 0.665z_1 + 0.7065z_2 + 0.5875z_3 = 377.99, \\
y_7 &= Y(7) - Y(6) \\
&= 0.231z_1 + 0.469z_2 + 0.0525z_3 = 119.29.
\end{aligned}$$

In matrix form, the cumulative outputs at each point in time t, $1 \le t \le 12$, are:

$$\begin{bmatrix} Y(1) \\ Y(2) \\ Y(3) \\ Y(4) \\ Y(5) \\ Y(6) \\ Y(7) \\ Y(8) \\ Y(9) \\ Y(10) \\ Y(11) \\ Y(12) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0.175 & 0.0225 & 0 \\ 0.175 & 0.2375 & 0.035 \\ 0.175 & 0.2375 & 0.285 \\ 0.434 & 0.2375 & 0.535 \\ 0.665 & 0.7065 & 0.5875 \\ 0.665 & 0.9025 & 1.0915 \\ 0.694 & 0.9025 & 1.7915 \\ 0.7 & 0.9465 & 2.2325 \\ 0.7 & 0.95 & 2.279 \\ 0.7 & 0.95 & 2.329 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 21.29 \\ 68.97 \\ 150.89 \\ 258.70 \\ 377.99 \\ 576.14 \\ 808.41 \\ 960.91 \\ 976.75 \\ 993.12 \end{bmatrix}$$

The cumulative output up to time $t = 12$, $Y(12) = 993.12$, equals the cumulative input up to time $t = 12$, $Z(12)$, *less* 6.68. The gap between the cumulative output and input is a result of the fact that 5% of the starts from time 3.58 to 4, i.e., $0.05\{(4 - 3.58)327.\bar{6}\} = 6.68$, will not yet be realized as output by time 12.

Remark 18.15. Fix time points, $t_1, t_2, \ldots$, not necessarily integer, such that the (rate-based) input curve is constant on each interval of time (t_{i-1}, t_i). For each real number x, let

$$\iota(x) := \max\{i : \ t_i \le x\} \tag{18.15}$$

denote the highest index associated with those t_i no larger than x. The cumulative curve $Z(\cdot)$ is a piecewise linear function; its slope on the interval $[t_{i-1}, t_i]$ is z_i. Consequently, for real numbers a and b, $a < b$,

$$\begin{aligned}\int_a^b z(\tau)d\tau &= Z(b) - Z(a)\\ &= [Z(t_{\iota(b)}) + (b - t_{\iota(b)})z_{\iota(b)+1}] \ - \ [Z(t_{\iota(a)}) + (a - t_{\iota(a)})z_{\iota(a)+1}]\\ &= [Z(t_{\iota(b)}) - Z(t_{\iota(a)})] \ + \ \{(b - t_{\iota(b)})z_{\iota(b)+1} - (a - t_{\iota(a)})z_{\iota(a)+1}\}\\ &= \sum_{k=\iota(a)+1}^{\iota(b)} z_k(t_k - t_{k-1}) \ + \ \{(b - t_{\iota(b)})z_{\iota(b)+1} - (a - t_{\iota(a)})z_{\iota(a)+1}\}.\end{aligned}$$

If a and b happen to belong to the same interval $[t_{i-1}, t_i]$ for some i, then $\iota(a) = \iota(b)$ and $\int_a^b z(\tau)d\tau = (b - a)z_{\iota(a)+1}$. If, for example, $a = 1.41$ and $b = 3.08$ in the example above, then $\iota(a) = 1$, $\iota(b) = 2$ and

$$\begin{aligned}\int_{1.41}^{3.08} z(\tau)d\tau &= Z(3.08) - Z(1.41)\\ &= [Z(1.65) + (3.08 - 1.65)z_3] - [Z(0.70) + (1.41 - 0.70)z_2]\\ &= [Z(1.65) - Z(0.70)] + \{1.43z_3 - 0.71z_2\}\\ &= z_2 + \{1.43z_3 - 0.71z_2\}\\ &= 0.29z_2 + 1.43z_3.\end{aligned}$$

18.3 Time-Dependent Lead Time Processes

18.3.1 Description

Let $\ell : \mathbb{R}_+ \to \mathbb{R}_+$ be a continuous function, and for each $\tau \in \mathbb{R}_+$ define $W(t, \tau) = 1$ if $t \ge \ell(\tau)$ and 0 otherwise. With this choice for the lead time distribution, an input that arrives at time τ will be completed exactly $\ell(\tau)$ units of time later. Using the fact that $W(t-\tau, \tau) = 1$ if and only if $t \ge \tau + \ell(\tau)$, the output in (18.3) is

$$Y(t) = \int_{\{\tau:\ \tau+\ell(\tau)\le t\}} z(\tau)d\tau.$$

While not required, inputs after time τ are often always completed after any input prior to time τ. This will occur when the **time-of-completion function**

$$\rho(\tau) := \tau + \ell(\tau)$$

is increasing in which case

$$Y(t) = \int_0^{\rho^{-1}(t)} z(\tau)d\tau = Z(\rho^{-1}(t)).$$

Constant lead time applies when $\ell(\tau) = \ell$, in which case $\rho^{-1}(t) = (t - L)$.

For the following two examples we fix the input curve as

$$z(t) = 100 \cdot 1_{[0,1]}(t) + 200 \cdot 1_{[1,2]}(t) + 300 \cdot 1_{[2,3]}(t) + 400 \cdot 1_{[3,4]}(t),$$

so that no input is consumed prior to time 0 and 100, 200, 300 and 400 units are started uniformly within the periods 1, 2, 3 and 4, respectively. As before, z_t denotes the cumulative input consumed over the interval $[t-1, t]$ for positive integers t so that $z_t = 100t$, $t = 1, 2, 3, 4$, for this example.

18.3.2 First-In, First-Out Example

In this example, we set the time-of-completion function to

$$\rho(\tau) = (t - 1) + \sqrt{\tau - t - 1},\ \tau \in [t - 1, t],\ t = 1, 2, \ldots\ .$$

If an input starts at time $\tau \in [0, 1]$, then it will emerge as output at time $\rho(\tau) = \sqrt{\tau}$ and its lead time $\ell(\tau)$ is $\sqrt{\tau} - \tau$. For example, if the input starts at time $\tau = 0.16$, then it will emerge as output at time $\rho(0.16) = 0.4$ and its lead time $\ell(0.16)$ is 0.24. If an input starts at time $\tau = 1.16 \in [1, 2]$, then it will emerge as output at time $\rho(1.16) = 1 + \sqrt{1.16 - 1} = 1.4$ and its lead time $\ell(1.4)$ is 0.24. With this time-of-completion function, all input that starts within the time interval $[t - 1, t]$ will emerge as output within this same time interval; moreover, the lead time for an input that starts at time τ will be identical to the lead time for starts at times $\tau - 1$, $\tau - 2$, etc.

Table 18.1 records a sample of start-times, time-of-completion, and lead times. The start-time and time-of-completion should be interpreted in relation to the interval $[t - 1, t]$, i.e., the start-time and time-of-completion represent the amount of time after time $t - 1$. The lead time for a start τ units after time $t-1$ is $\sqrt{\tau}-\tau$, which reaches its maximum at $\tau = 0.25$. Thus, lead times are increasing for the first quarter of the time interval and decrease thereafter.

To determine the cumulative output $Y(t)$ in (18.16), it remains to determine $\rho^{-1}(t)$. Pick a $t \in [0, 1]$. The unit of output that emerges at time t must have been started at the time $\tau = \rho^{-1}(t)$ for which $\sqrt{\tau} = t$. Obviously, this value is t^2. Pick a $t \in [1, 2]$. The unit of output that emerges at time t must have been started at the time $\tau = \rho^{-1}(t)$ for which $1 + \sqrt{\tau - 1} = t$. This value is $1 + (t - 1)^2$. The general formula is therefore

$$\rho^{-1}(t) = (n - 1) + (t - n - 1)^2,\ t \in [n - 1, n],\ n = 1, 2, \ldots\ .$$

For example, when $t = 2.9$, $\rho^{-1}(t) = 2 + (2.9 - 2)^2 = 2.81$, and so

$$Y(2.9) = Z(2.81) = 100 + 200 + (0.81)300 = 543.$$

Table 18.1. Start-times, time-of-completion, and lead times for example 18.3.2.

Start-time	Time-of-Completion	Lead time
0.00	0.00	0.00
0.01	0.10	0.09
0.04	0.20	0.12
0.09	0.30	0.21
0.16	0.40	0.21
0.25	0.50	0.25
0.36	0.60	0.24
0.49	0.70	0.21
0.64	0.80	0.16
0.91	0.90	0.09
1.00	1.00	0.00

18.3.3 Leapfrog Example

For this example, the time-of-completion function is

$$\rho(\tau) = \begin{cases} \tau + \tau^2, & 0 \le \tau \le 1, \\ \tau + (\tau - 1)^2, & 1 \le \tau \le 2, \\ \tau + (\tau - 2)^2, & 2 \le \tau \le 3, \\ \tau + (\tau - 3)^2, & 3 \le \tau \le 4. \end{cases}$$

For example, if an input starts at time 1.1, then its time-of-completion is $\rho(1.1) = 1.11$ and its lead time is 0.01; if an input starts at time 1.5, then its time-of-completion is $\rho(1.5) = 1.75$ and its lead time is 0.25; if an input starts at time 1.9, then its time-of-completion is $\rho(1.9) = 2.71$ and its lead time is 0.81. With this time-of-completion function, all input that starts within the time interval $[t-1, t]$ will emerge as output continuously within the time interval $[t-1, t+1]$. That is, it takes two units of time to complete all starts within a single period of time; moreover, the lead time for an input that starts at time τ will be identical to the lead time for starts at times $\tau - 1$, $\tau - 2$, etc.

In this example, starts early in the interval emerge quickly, whereas starts late in the interval emerge more slowly. Because of this, starts late in one interval will emerge *after* starts at the beginning of the next interval. For example, the time-of-completion for a start at time $\tau = 1.9$ is 2.71, whereas the time-of-completion for a start at time $\tau = 2.1$ is 2.11. This services system does not preserve the "First-In, First-Out (FIFO)" discipline.

Since the system is not FIFO, the cumulative output $Y(t)$ cannot be determined via (18.16). Instantaneous output at time $t \in [n-1, n]$ emerges as a result of *two* input streams: instantaneous starts some time ago within the current period $[n-1, n]$ *and* instantaneous starts some time ago within the previous period $[n-2, n-1]$. In effect, there are *two* $\rho^{-1}(t)$'s, which we label $\rho^{-1}_{current}(t)$ and $\rho^{-1}_{previous}(t)$. The calculations are best illustrated with a concrete example.

Pick a $t \in [0,1]$. The unit of output that emerges at time t must have been started at the time $\tau = \rho^{-1}_{current}(t)$ for which $\tau + \tau^2 = t$. (There is no starts prior to time 0, and so there is no $\rho^{-1}_{previous}(t)$ when $t \in [0,1]$.) This value is

$$\rho^{-1}_{current}(t) = -0.5 + 0.5\sqrt{4t+1} = 0.5\left(\sqrt{4t+1} - 1\right),\ t \in [0,1].$$

For example, if $t = 0.5$, then $\rho^{-1}_{current}(0.5) = 0.366$. The corresponding lead time $\ell(0.366) = 0.5 - 0.366 = 0.134$. Now pick a $t \in [1,2]$. The flow of output that emerges at time t *as a result of starts within* $[1,2]$ must have been started at the time $\tau = \rho^{-1}_{current}(t)$ for which $\tau + (\tau - 1)^2 = t$. This value is

$$\rho^{-1}_{current}(t) = 0.5 + 0.5\sqrt{4t-3} = 1 + 0.5\left(\sqrt{4t-3} - 1\right),\ t \in [1,2].$$

For example, if $t = 1.8$, then $\rho^{-1}_{current}(1.8) = 1.5247$. The corresponding lead time $\ell(1.5247) = 1.8 - 1.5247 = 0.2753$. The flow of output that emerges at time t *as a result of starts within* $[0,1]$, however, must have been started at the time $\tau = \rho^{-1}_{previous}(t)$ for which $\tau + \tau^2 = t$. This value is

$$\rho^{-1}_{previous}(t) = -0.5 + 0.5\sqrt{4t+1} = 0.5\left(\sqrt{4t+1} - 1\right),\ t \in [1,2].$$

For example, if $t = 1.8$, then $\rho^{-1}_{previous}(1.8) = 0.9318$. The corresponding lead time for this start is $\ell(0.9318) = 1.8 - 0.9318 = 0.8682$. One may repeat the calculation to show that the general formulae are

$$\rho^{-1}_{current}(t) = (n-1) + 0.5\left(\sqrt{4t - (4n-5)} - 1\right),\ t \in [n-1, n],\ n \geq 1,$$
$$\rho^{-1}_{previous}(t) = (n-2) + 0.5\left(\sqrt{4t - (4n-9)} - 1\right),\ t \in [n-1, n],\ n \geq 2.$$

For this production process, the lead times for units emerging exactly one time unit apart will be identical, and hence there corresponding start-times will be offset by exactly one time unit. For example, when $t = 2.8$,

$$\rho^{-1}_{current}(2.8) = 2 + 0.5(\sqrt{4.2} - 1) = 2.5247,$$
$$\rho^{-1}_{previous}(2.8) = 1 + 0.5(\sqrt{8.2} - 1) = 1.9318,$$

and so

$$\begin{aligned} Y(2.8) &= Z(1.9318) + 0.5247 z_3 \\ &= z_1 + (0.9318) z_2 + 0.5247 z_3 \\ &= 100 + (0.9318)200 + (0.5247)300 = 443.77. \end{aligned}$$

Similarly, when $t = 4$,

$$\rho^{-1}_{current}(4) = 3 + 0.5(\sqrt{5} - 1) = 3.618,$$
$$\rho^{-1}_{previous}(4) = 2 + 0.5(\sqrt{9} - 1) = 3,$$

and so

$$\begin{aligned} Y(4) &= Z(3) + 0.618 z_4 \\ &= z_1 + z_2 + z_3 + (0.618) z_4 \\ &= 100 + 200 + 300 + (0.618)400 = 847.21. \end{aligned}$$

It will not be until time 5 that all 1000 input starts will be realized as output.

Remark 18.16. Since it takes two time units to complete all starts within a single period, two $\rho^{-1}(\cdot)$ functions are required to represent cumulative output. They can be re-labeled $\rho_0^{-1}(\cdot)$ and $\rho_{-1}^{-1}(\cdot)$, respectively, where '0' and '-1' denote the current and previous periods. If instead it took three units of time to complete all starts within a single period of time, then three $\rho^{-1}(\cdot)$ functions, $\rho_0^{-1}(\cdot)$, $\rho_{-1}^{-1}(\cdot)$, $\rho_{-2}^{-1}(\cdot)$, will be required.

18.4 Continuous Lead Time Processes

18.4.1 Description

For computational purposes, all rate-based flows in previous examples were piecewise continuous; the points of discontinuity defined a natural time grid, which we now formally define.

Definition 18.17. *A* **time grid** $\mathcal{G} := \{t_i\}$ *is a finite or countably infinite collection of points in time such that*

$$\ldots t_{-2} < t_{-1} < t_0 := 0 < t_1 < t_2 < \ldots,$$

and the number of t_i *in any bounded interval is finite.* **Period i** *refers to the interval* $[t_{i-1}, t_i)$ *and its* **length** *is* $t_i - t_{i-1}$. *A* **uniform time grid** *has identical period lengths and a* **standard time grid** *has period lengths equal to 1.*

Fix a time grid $\mathcal{G}$, and let $z(\cdot)$ be a rate-based flow whose points of discontinuity belong to $\mathcal{G}$. For each interval $[t_{k-1}, t_k]$, define the *scale* (cumulative input) as

$$z_k := \int_{t_{k-1}}^{t_k} z(\tau) d\tau, \tag{18.16}$$

and the *shape* as

$$s_k(\tau) := z(\tau)/z_k. \tag{18.17}$$

By construction $\int_{t_{k-1}}^{t_k} s_k(\tau) d\tau = 1$ for all k. Conceptually, $z(\cdot)$ can be identified by the collection $\{(z_k, s_k(\cdot))\}$ of scales and shapes. In what follows, the shapes $\{s_k(\cdot)\}$ are *given* and remain fixed; the scales $\{z_k\}$ will be allowed to vary.

Let $W(\cdot,\cdot)$ be a cumulative lead time density as defined in Chapter 17. Replacing $\Phi(z(\tau),\tau)$ with $z(\tau)$ in (18.3) on p. 310, the cumulative output is

$$Y(t) = \int_{-\infty}^{t} z(\tau)W(t-\tau,\tau)d\tau. \tag{18.18}$$

Under the assumptions above, $Y(t)$ is a linear combination of the z_k and the weights can be *pre-computed*. To see this, using the definitions above,

$$\begin{aligned} Y(t) &= \sum_{k:t_k\le t} \int_{t_{k-1}}^{t_k} z(\tau)W(t-\tau,\tau)d\tau \\ &= \sum_{k:t_k\le t} z_k \int_{t_{k-1}}^{t_k} s_k(\tau)W(t-\tau,\tau)d\tau \\ &= \sum_{k:t_k\le t} z_k \Pi_k(t), \end{aligned} \tag{18.19}$$

where, for each k,

$$\Pi_k(t) := \int_{t_{k-1}}^{t_k} s_k(\tau)W(t-\tau,\tau)d\tau. \tag{18.20}$$

The $\Pi_k(\cdot)$ can be pre-computed.

Examples in the next section apply the following useful special case:

- The time grid $\mathcal{G}$ is uniform with common period length $L := t_k - t_{k-1}$.
- The shapes $s_k(\cdot)$ are identical in that there exists a *single* shape curve $s(\cdot)$ defined on $[0, L]$ such that $s_k(\tau) = s(\tau - t_{k-1})$ on $[t_{k-1}, t_k]$ for all k.
- The lead time distribution does not depend on the arrival time τ. Accordingly, the second argument of $W(\cdot)$ will be suppressed.

Under these assumptions, for $t \ge t_{k-1}$,

$$\Pi_k(t) = \int_0^L s(\tau)W((t-t_{k-1})-\tau)d\tau := \hat{\Pi}(t-t_{k-1}). \tag{18.21}$$

The weight $\Pi_k(t)$ on the variable z_k in (18.19) depends only on the *difference* $t - t_{k-1}$. When t coincides with a time grid point $t_i > t_{k-1}$, we let

$$\hat{\Pi}_{i-(k-1)} := \hat{\Pi}(t_i - t_{k-1}),$$

and write

$$Y(t_i) = \sum_{k\le i} z_k \hat{\Pi}_{i-(k-1)}. \tag{18.22}$$

Here, $\hat{\Pi}_{i-(k-1)}$ represents the percentage of total output that will emerge by the end of period $i \ge k$ as a result of the input in period k.

Example 18.18. Suppose it takes a maximum of three time units for output to continuously emerge. As a result of input in a period, output will emerge during this period, one period after, two periods after, and three periods after this period. Suppose the respective percentages are 0.10, 0.20, 0.30 and 0.40. Then, for example, $Y(t_7) = 0.10z_7 + 0.20z_6 + 0.30z_5 + 0.40z_4$ and $Y(t_2) = 0.10z_2 + 0.20z_1 + 0.30z_0 + 0.40z_{-1}$. As we have previously noted, when production is not instantaneous, output after time 0 could be a result of input prior to time 0, which is past history and must be pre-specified. We shall discuss this in detail in Chapter 19.

18.4.2 Examples

In this section, we use a continuous lead time model to answer the question posed in Section 17.1. We assume a standard time grid.

We shall compute the $\hat{\Pi}_k$ for the following concrete examples. For the shape curve $s(\tau)$, $\tau \in [0,1]$, we examine three cases:

- *Constant loading (C)*: $s(\tau) = 1$ for all $\tau \in [0,1]$.
- *Front loading (F)*: $s(\tau) = 2(1-\tau)$.
- *Back loading (B)*: $s(\tau) = 2\tau$.

As for the lead time distribution $W(\cdot)$, we assume the maximal lead time is 2 periods and examine three analogous cases:

- *Uniform (U)*: $W(t) = t/2$.
- *Early (E)*: $W(t) = 1 - (1 - t/2)^2$.
- *Late (L)*: $W(t) = t^2/4$.

In each case $W(t) = 1$ when $t \geq 2$. In this setting, there are only two numbers to compute for each of the $3 \cdot 3 = 9$ scenarios, namely, $\hat{\Pi}_1$ and $\hat{\Pi}_2$, since $\hat{\Pi}_3 = 1$. (This is because $W(3-\tau) = 1$ for each $\tau \in [0,1]$ and $\int_0^1 s(\tau)d\tau$ by definition equals one.)

Set $\hat{\Pi}_0 := 0$ and let $\hat{\pi}_k := \hat{\Pi}_k - \hat{\Pi}_{k-1}$, $k = 1,2,3$. Each $\hat{\pi}_k$ reflects the percentage of total output that will emerge in the k^{th} period following the input in any one period, and can be viewed as the probability mass function corresponding to the cumulative distribution given by $\hat{\Pi}$. Table 18.2 shows the $\hat{\pi}_k$'s for the 9 scenarios.

Table 18.2. Probability mass functions for each scenario.

	U			E			L		
	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$
C	1/4	1/2	1/4	5/12	1/2	1/12	1/12	1/2	5/12
F	1/3	1/2	1/6	13/24	5/12	1/24	1/8	7/12	7/24
B	1/6	1/2	1/3	7/24	7/12	1/8	1/24	5/12	13/24

Remark 18.19. Due to "time-reversibility," i.e., moving backwards in time from 1 to 0 instead of moving forward in time from 0 to 1, there are four pairs of scenarios in which the $\hat{\pi}_k$'s for one of the scenarios is the reverse of the $\hat{\pi}_k$'s for its matching pair. For example,

$$(\hat{\pi}_1^{BE}, \hat{\pi}_2^{BE}, \hat{\pi}_3^{BE}) = (\hat{\pi}_3^{FL}, \hat{\pi}_2^{FL}, \hat{\pi}_1^{FL}).$$

The same relationship holds for the scenario pairs "CE-CL," "FU-BU," "FE-BL." The scenario "CU" has no matching pair.

Example 18.20. Consider scenario "FL." We have

$$\begin{aligned}
\hat{\Pi}_1 &= \int_0^1 [2(1-\tau)] \frac{(1-\tau)^2}{4} d\tau \\
&= (1/2) \int_0^1 (1-\tau)^3 d\tau \\
&= -(1/8)(1-\tau)^4 \Big|_0^1 \\
&= 1/8, \\
\hat{\Pi}_2 &= \int_0^1 [2(1-\tau)] \frac{(2-\tau)^2}{4} d\tau \\
&= (1/2) \int_0^1 (4 - 8\tau + 5\tau^2 - \tau^3) d\tau \\
&= (1/2)[4\tau - 4\tau^2 + (5/3)\tau^3 - \tau^4/4 \Big|_0^1 \\
&= 17/24.
\end{aligned}$$

Thus $\hat{\pi}_1 = 1/8$, $\hat{\pi}_2 = \hat{\Pi}_2 - \hat{\Pi}_1 = 17/24 - 1/8 = 7/12$, and $\hat{\pi}_3 = \hat{\Pi}_3 - \hat{\Pi}_2 = 1 - 17/24 = 7/24$.

In the example of Section 17.1, the cumulative input in each of the first three periods is 24, 48 and 96, respectively. Using the $\hat{\pi}_k$'s in Table 18.2 and equation (18.22), the outputs in each period and the cumulative outputs at the end of each of the first three periods for each scenario are shown in Tables 18.3 and 18.4. There is a wide discrepancy (in percentage terms) in the projected output curve across scenarios, and there is a huge gap between the projected output curves shown in Table 18.3 and the instantaneous output curve given by $(24, 72, 168)$.

Example 18.21. For scenario "FL," the $\hat{\pi}$ vector is $(1/8, 7/12, 7/24)$. This means that 3 of the 24 starts in period 1 will emerge as output by time 1, 14 will emerge as output by time 2, and 7 units will emerge as output by time 3. Symbolically we represent this as

$$24 \text{ starts in period } 1 \longrightarrow 24(1/8, 7/12, 7/24) = (3, 14, 7, 0, 0, \cdots).$$

Furthermore, 9 of the 72 starts in period 2 will emerge as output by time 2, 42 units will emerge as output by time 3, and 21 units will emerge by time 4, i.e.,

$$72 \text{ starts in period } 2 \longrightarrow 72(0, 1/8, 7/12, 7/24) = (0, 9, 42, 21, 0, 0, \cdots).$$

Finally, 21 of the 168 starts in period 3 will emerge as output by time 3, 98 units will emerge as output by time 4, and 49 units will emerge by time 5, i.e.,

$$168 \text{ starts in period } 3 \longrightarrow 168(0, 0, 1/8, 7/12, 7/24) = (0, 0, 21, 98, 49, 0, 0, \cdots).$$

Summing the components of these three respective vectors, the output curve over the first 5 periods is

$$(3, 14, 7, 0, 0) + (0, 9, 42, 21, 0) + (0, 0, 21, 98, 49) = (3, 23, 70, 98, 49).$$

(The first three components of this output curve are recorded in Table 18.3 under scenario "FL.")

If the input-output transformation is as in scenario "BL," total output by time 3 will be only 64. If, on the other hand, one erroneously assumes the input-output transformation is instantaneous, total output by time 3 will be projected as 254, *a three-fold increase*. If the true scenario is "BL" but the (implicit) instantaneous model is assumed, then projected output will fall far short of actual output, leaving many customers unsatisfied (or extra unexpected costs to service the demand).

Table 18.3. Outputs in each period for each scenario.

	U			E			L		
	1	2	3	1	2	3	1	2	3
C	6	24	54	10	32	66	2	16	42
F	8	28	60	13	36	73	3	20	47
B	4	20	48	7	28	59	1	12	37

To summarize, when lead times are significant in relation to the period length, a nontrivial dynamic production function is necessary to model the production process. The class of distribution-based dynamic production functions is conceptually simple and can capture a wide variety of non-instantaneous input-output processes.

Remark 18.22. Left unanswered is how to obtain the lead time distribution $W(\cdot, \cdot)$. Practical approaches include using shop-floor statistics or simulation. Use of shop-floor statistics implicitly assumes the past production system will be reasonably representative of the future production system. Simulation implicitly assumes that the choice of input(s) over time is representative of future choices over time.

Table 18.4. Cumulative outputs at end of each period for each scenario.

	U			E			L		
	1	2	3	1	2	3	1	2	3
C	6	30	84	10	42	108	2	18	60
F	8	36	96	13	49	122	3	23	70
B	4	24	72	7	35	94	1	13	50

18.5 Exercises

18.1. Suppose $z(t) = 50$, $t \in [0, 3]$, and the constant lead time is $\ell = 2$. Determine $Z(\cdot)$ and the corresponding cumulative output curve $Y(\cdot)$.

18.2. Suppose

$$z(t) = 50 \cdot 1_{[0,3)}(t) + 75 \cdot 1_{[3,4)}(t) + 175 \cdot 1_{[4,9)}(t) + 50(t-9) \cdot 1_{[9,12)}(t) + 90(t-12)^2 \cdot 1_{[12,16]}(t).$$

(a) Suppose the constant lead time is $\ell = 2$. Determine $Z(\cdot)$ and the corresponding cumulative output curve $Y(\cdot)$.
(b) Suppose the constant lead time is $\ell = 2.2$. Determine values for $Z(12)$, $Y(12)$ and $Y(17)$.

18.3. Consider Example 18.2.2.

(a) Determine an expression for $Y(36)$ in terms of the $Z(\cdot)$ variables.
(b) Determine an expression for $y_{36} = \int_{35}^{36} y(\tau)d\tau$ in terms of the z_t.
(c) Suppose the values of z_t are now $z_t = 50t^2$, $t = 1, 2, 3, 4$. Determine the new output vector $y = (y_1, y_2, \ldots, y_{12})$ and cumulative output over the time interval $[0, 12]$.

18.4. Consider Example 18.2.3.

(a) Determine an expression for $Y(18)$ in terms of the $Z(\cdot)$ variables.
(b) Determine an expression for $y_{18} = \int_{17}^{18} y(\tau)d\tau$ in terms of the z_t.
(c) Suppose the values of z_t are now $z_t = 10t^3$, $t = 1, 2, 3, 4$. Determine the new output vector $y = (y_1, y_2, \ldots, y_{12})$ and cumulative output over the time interval $[0, 12]$. Explain the difference between the cumulative input and the cumulative output.

18.5. Consider Example 18.2.4.

(a) Derive the expression for $Y(8)$ and y_8.
(b) Suppose the values of z_t are now $z_1 = 80$, $z_2 = 180$, and $z_3 = 400$. Determine the new output vector $y = (y_1, y_2, \ldots, y_{12})$ and cumulative output over the time interval $[0, 12]$. Explain the difference between the cumulative input and the cumulative output.

18.6. Suppose there are three events with $w_1 = 0.15$, $w_2 = 0.60$, and $w_3 = 0.25$. Suppose $z(\cdot)$ is a rate-based flow, constant on each interval of time $(t-1, t]$, $t = 1, 2, \ldots$. Let z_t denote the constant rate in period t.

(a) Suppose $\ell_1 = 1$, $\ell_2 = 3$, and $\ell_3 = 4$. Determine the y_t in terms of the z_t.
(b) Suppose $\ell_1 = 1.2$, $\ell_2 = 3.35$, and $\ell_3 = 4.72$. Determine the y_t in terms of the z_t.
(c) Suppose $\ell_1 = 1.2$, $\ell_2 = 3.35$, and $\ell_3 = 4.72$, except that here $z(\cdot)$ is a rate-based flow, constant on each interval of time $1.5(t-1, t]$, $t = 1, 2, \ldots$. That is, the period lengths for the input curve are now 1.5 time units. Let z_t denote the constant rate in period t and determine the y_t in terms of the z_t.

18.7. Consider Example 18.3.2. Determine the value of $Y(2.7)$.

18.8. Consider Example 18.3.3. Determine the value of $Y(3.4)$.

18.9. Consider Example 18.3.3, except that here

$$\rho(\tau) = \begin{cases} \tau + \sqrt{\tau}, & 0 \le \tau \le 1, \\ \tau + \sqrt{\tau - 1}, & 1 \le \tau \le 2, \\ \tau + \sqrt{\tau - 2}, & 2 \le \tau \le 3, \\ \tau + \sqrt{\tau - 3}, & 3 \le \tau \le 4. \end{cases}$$

(a) Derive expressions for $\rho^{-1}_{current}(t)$ and $\rho^{-1}_{previous}(t)$.
(b) Determine the value for $Y(3.8)$.
(c) Determine the value for $Y(4)$.

18.10. Consider the probability mass functions provided in Table 18.2.

()a Show how to obtain the $\hat{\pi}$ values for scenario "FE."
(b) Determine the outputs in each of the first five periods for scenario "FE."
(c) Determine the cumulative output by time 4 for scenario "BE."

18.11. Consider the continuous lead time model in which $s(t) = 3(1-t)^2$, $t \in [0, 1]$, and $W(t) = 1 - (1 - t/2)^3$, $t \in [0, 2]$.

(a) What is the maximal lead time?
(b) Determine the probability mass function.
(c) Determine the outputs over the first five periods (when the inputs are $z_1 = 24$, $z_2 = 48$ and $z_3 = 96$, as in the motivating example).
(d) Determine the cumulative output by time 3.

18.12. How will your answers to parts (b) and (c) of Exercise 18.10 change if the initial conditions are $z_0 = 144$, $z_{-1} = 192$ and $z_{-2} = 72$?

18.6 Solutions to Exercises

18.1 The specific values are:

$$Z(t) = \begin{cases} 50t, & 0 \le t \le 3, \\ 150, & 3 \le t. \end{cases}$$

$$Y(t) = \begin{cases} 0, & 0 \le t \le 2, \\ 50(t-2), & 2 \le t \le 5, \\ 150, & 5 \le t. \end{cases}$$

18.2 (a) The specific values are:

$$Z(t) = \begin{cases} 50t, & 0 \le t \le 3, \\ 150 + 75(t-3), & 3 \le t \le 4, \\ 225 + 175(t-4), & 4 \le t \le 9, \\ 1100 + 25(t-9)^2, & 9 \le t \le 12, \\ 1325 + 30(t-12)^3, & 12 \le t \le 16, \\ 3245, & 16 \le t. \end{cases}$$

$$Y(t) = \begin{cases} 0, & 0 \le t \le 2, \\ 50(t-2), & 2 \le t \le 5, \\ 150 + 75(t-5), & 5 \le t \le 6, \\ 225 + 175(t-6), & 6 \le t \le 11, \\ 1100 + 25(t-11)^2, & 11 \le t \le 14, \\ 1325 + 30(t-14)^3, & 14 \le t \le 18, \\ 3245, & 18 \le t. \end{cases}$$

(b) $Z(12) = 1325$, $Y(12) = Z(9.8) = 225 + 25(9.8-9)^2 = 241$, and $Y(17) = Z(14.8) = 1325 + 30(14.8-12)^3 = 1983.56$.

18.3 (a) $Y(36) = 0.25Z(34) + 0.70Z(29) + 0.05Z(28)$.(b) $y_{36} = 0.25z_{34} + 0.70z_{29} + 0.05z_{28}$.

(c) Here $z^T = (50, 200, 450, 800)$ and

$$y^T = (0, 0, 12.5, 50, 112.5, 235, 140, 315, 562.50, 10, 22.5, 80).$$

With respect to the cumulative values, $Y(12) = 1500 = z_1 + z_2 + z_3 + z_4$.

18.4 (a) $Y(18) = 0.25Z(15.79) + 0.70Z(12.37) + 0.05Z(9.58)$.
(b)

$$y_{18} = 0.021z_9 + 0.029z_{10} + 0.441z_{12} + 0.259z_{13} + 0.1975z_{15} + 0.0525z_{16}.$$

(c) Here, $z^T = (10, 80, 270, 2560)$ and

$$y^T = (0, 0, 1.975, 16.325, 57.525, 522.365, 25.13, 105.21, 782.4, 1131.49, 9.51, 79.91).$$

With respect to the cumulative values,

$$Y(12) = 2866.24 = z_1 + z_2 + z_3 + z_4 - 53.76.$$

The gap is due to the fact that 5% of the starts over the interval $[3.58, 4]$, i.e., $0.05[0.42(2560)] = 53.76$, will not be completed by time 12.

18.5 (a) $Y(8) = 0.665z_1 + 0.9025z_2 + 1.0915z_3$ and $y_8 = 0.196z_2 + 0.504z_3$.
(b) Here,

$$y^T = (0, 0, 18.05, 70.75, 170.75, 291.47, 415.37, 652.25, 934.57, 1119.37, 1138.6, 1158.6).$$

The cumulative output is 1158.6, which equals the cumulative input of 1167 less 8.4. This gap is due to fact that 5% of the starts in the interval, i.e., $[3.58, 4] = 0.05[(4 - 3.58)400] = 8.4$, will not be completed by time 12.

18.6 (a) $y_t = 0.15z_{t-1} + 0.60z_{t-2} + 0.25z_{t-4}$.
(b) Here,

$$\begin{aligned} y_t &= 0.15 \int_{t-1}^{t} z(\tau - 12.)d\tau + 0.60 \int_{t-1}^{t} z(\tau - 3.35)d\tau \\ &\quad + 0.25 \int_{t-1}^{t} z(\tau - 4.72)d\tau \\ &= 0.15[0.2z_{t-2} + 0.8z_{t-1}] + 0.60[0.35z_{t-4} + 0.65z_{t-3}] \\ &\quad + 0.25[0.72z_{t-5} + 0.25z_{t-4}] \\ &= 0.18z_{t-5} + 0.28z_{t-4} + 0.39z_{t-3} + 0.03z_{t-2} + 0.12z_{t-1}. \end{aligned}$$

(c) We have

$$Y(t) = 0.15Z(t - 1.2) + 0.60Z(t - 3.35) + 0.25Z(t - 4.72).$$

Since the period lengths = 1.5, a generic point in time t corresponds to $t/1.5$ time periods and belongs to the time interval $[1.5(t/1.5)^-, 1.5(t/1.5)^+]$. For example, $t = 4.8$ corresponds to $4.8/1.5 = 3.2$ time periods and belongs to the time interval $[4.5, 6.0]$. Accordingly,

$$Z(t) = z_1 + z_2 + \cdots + z_{(t/1.5)^-} + [(t/1.5) - (t/1.5)^-]z_{(t/1.5)^+}\ .$$

Thus,

$$Y(8) = 0.15Z(6.8) + 0.60Z(4.65) + 0.25Z(3.28),$$

where

$$\begin{aligned} Z(6.8) &= z_1 + z_2 + z_3 + z_4 + (0.8/1.5)z_5, \\ Z(4.65) &= z_1 + z_2 + z_3 + (0.15/1.5)z_4, \\ Z(3.28) &= z_1 + z_2 + (0.28/1.5)z_3. \end{aligned}$$

Consolidating terms,

$$Y(8) = z_1 + z_2 + 0.79\bar{6}z_3 + 0.21z_4 + 0.08z_5.$$

Similarly,

$$Y(7) = 0.15Z(5.8) + 0.60Z(3.65) + 0.25Z(2.28),$$

where

$$\begin{aligned} Z(5.8) &= z_1 + z_2 + z_3 + (1.3/1.5)z_4, \\ Z(3.65) &= z_1 + z_2 + (0.65/1.5)z_3, \\ Z(2.28) &= z_1 + (0.78/1.5)z_2. \end{aligned}$$

Consolidating terms,

$$Y(8) = z_1 + 0.88z_2 + 041z_3 + 0.13z_4.$$

Finally,

$$y_8 = Y(8) - Y(7) = 0.12z_2 + 0.38\bar{6}z_3 + 0.08z_4 + 0.08z_5.$$

18.7 First we calculate $\rho^{-1}(2.7) = 2 + (0.7)^2 = 2.49$. Thus,

$$\begin{aligned} Y(2.7) = Z(2.49) = Z(2) + 0.49z_3 &= z_1 + z_2 + 0.49z_3 \\ &= 100 + 200 + (0.49)(300) = 447. \end{aligned}$$

18.8 First, we calculate the respective $\rho^{-1}(3.4)$'s as

$$\begin{aligned} \rho^{-1}_{current}(3.4) &= 3 + 0.5(\sqrt{4(3.4) - 11} - 1) = 3.3062, \\ \rho^{-1}_{previous}(3.4) &= 2 + 0.5(\sqrt{4(3.4) - 7} - 1) = 2.7845. \end{aligned}$$

Thus,

$$\begin{aligned} Y(3.4) &= Z(2.7845) + 0.3062z_3 \\ &= z_1 + z_2 + 0.7845z_3 + 0.3062z_4 \\ &= 100 + 200 + (0.7845)300 + (0.3062)400 = 657.83. \end{aligned}$$

18.9 (a) Pick a $t \in [0,1]$. The unit of output that emerges at time t must have been started at the time $\tau = \rho^{-1}_{current}(t)$ for which $\tau + \sqrt{\tau} = t$. (There is no starts prior to time 0, and so there is no $\rho^{-1}_{previous}(t)$ when $t \in [0,1]$.) The solution is obtained by defining $u := \sqrt{\tau}$, solving $u^2 + u = t$ for u and then setting $\tau = u^2$. We obtain:

$$\rho^{-1}_{current}(t) = (-0.5 + 0.5\sqrt{4t+1})^2 = 0.25(\sqrt{4t+1} - 1)^2,\ t \in [0,1].$$

For example, if $t = 0.5$, then $\rho^{-1}_{current}(0.5) = 0.134$. The corresponding lead time $\ell(0.134) = 0.5 - 0.134 = 0.366$. Now pick a $t \in [1,2]$. The flow of output that emerges at time t *as a result of starts within* $[1,2]$ must have been started at the time $\tau = \rho^{-1}_{current}(t)$ for which $\tau + \sqrt{\tau - 1} = t$. The solution is obtained by defining $u := \sqrt{\tau - 1}$, solving $(u^2 + 1) + u = t$ for u and then setting $\tau = 1 + u^2$. We obtain:

$$\rho^{-1}_{current}(t) = 1 + 0.25(\sqrt{4t-3} - 1)^2,\ t \in [1,2].$$

For example, if $t = 1.8$, then $\rho^{-1}_{current}(1.8) = 1.2753$. The corresponding lead time $\ell(1.2753) = 1.8 - 1.2753 = 0.5247$. The flow of output that emerges at time t *as a result of starts within* $[0,1]$, however, must have been started at the time $\tau = \rho^{-1}_{previous}(t)$ for which $\tau + \sqrt{\tau} = t$. This value is

$$\rho^{-1}_{previous}(t) = 0.25(\sqrt{4t+1} - 1)^2,\ t \in [1,2].$$

For example, if $t = 1.8$, then $\rho^{-1}_{previous}(1.8) = 0.8682$. The corresponding lead time for this start is $\ell(0.8682) = 1.8 - 0.8682 = 0.9318$. One may repeat the calculation to show that the general formulae are given by

$$\rho^{-1}_{current}(t) = (n-1) + 0.25\left(\sqrt{4t - (4n-5)} - 1\right)^2,\ t \in [n-1, n],\ n \geq 1,$$
$$\rho^{-1}_{previous}(t) = (n-2) + 0.25\left(\sqrt{4t - (4n-9)} - 1\right)^2,\ t \in [n-1, n],\ n \geq 2.$$

(b) When $t = 3.8$,

$$\rho^{-1}_{current}(3.8) = 3 + 0.25(\sqrt{4(3.8) - 11} - 1)^2 = 3.2753,$$
$$\rho^{-1}_{previous}(3.8) = 2 + 0.25(\sqrt{4(3.8) - 7} - 1)^2 = 2.8682,$$

and so

$$\begin{aligned} Y(3.8) &= Z(2.8682) + 0.2753z_4 \\ &= z_1 + z_2 + (0.8682)z_3 + 0.2753z_4 \\ &= 100 + +200 + (0.8682)300 + (0.2753)400 = 670.58. \end{aligned}$$

(c) When $t = 4$,

$$\rho^{-1}_{current}(4) = 3 + 0.25(\sqrt{4(4) - 11} - 1)^2 = 3.382,$$
$$\rho^{-1}_{previous}(4) = 2 + 0.25(\sqrt{4(4) - 7} - 1)^2 = 3,$$

and so

$$\begin{aligned} Y(4) &= Z(3) + 0.382z_4 \\ &= z_1 + z_2 + z_3 + (0.382)z_4 \\ &= 100 + 200 + 300 + (0.382)400 = 752.80. \end{aligned}$$

18.10 (a) We have:

$$\begin{aligned} \hat{\Pi}_1 &= \int_0^1 [2(1-\tau)] \left[1 - \left(1 - \frac{1-\tau}{2}\right)^2\right] d\tau \\ &= \int_0^1 [1.5 - 2.5\tau + 0.5\tau^2 + 0.5\tau^3] d\tau \\ &= 1.5\tau - 2.5\tau^2/2 + 0.5\tau^3/3 + 0.5\tau^4/4 \Big|_0^1 = 13/24. \\ \hat{\Pi}_2 &= \int_0^1 [2(1-\tau)] \left[1 - \left(1 - \frac{2-\tau}{2}\right)^2\right] d\tau \\ &= \int_0^1 [2 - 2\tau - 0.5\tau^2 + 0.5\tau^3] d\tau \\ &= 2\tau - \tau^2 - 0.5\tau^3/3 + 0.5\tau^4/4 \Big|_0^1 = 23/24. \end{aligned}$$

Thus, $\hat{\pi}_1 = \hat{\Pi}_1 = 13/24$, $\hat{\pi}_2 = \hat{\Pi}_2 - \hat{\Pi}_1 = 23/24 - 13/24 = 5/12$, and $\hat{\pi}_3 = \hat{\Pi}_3 - \hat{\Pi}_2 = 1 - 23/24 = 1/24$.

(b) Using the notation of the chapter,

$$\begin{aligned} 24(13/24, 5/12, 1/24) &\longrightarrow (13, 10, 1, 0, 0), \\ 48(13/24, 5/12, 1/24) &\longrightarrow (0, 26, 20, 2, 0), \\ 96(13/24, 5/12, 1/24) &\longrightarrow (0, 0, 52, 40, 4), \end{aligned}$$

and so $y^T = (13, 36, 73, 42, 4)$ and $Y^T = (13, 49, 122, 164, 168)$.

(c) Using the notation of the chapter,

$$\begin{aligned} 24(7/24, 7/12, 1/8) &\longrightarrow (7, 14, 3, 0, 0), \\ 48(7/24, 7/12, 1/8) &\longrightarrow (0, 14, 28, 6, 0), \\ 96(7/24, 7/12, 1/8) &\longrightarrow (0, 0, 28, 56, 12), \end{aligned}$$

and so $y^T = (7, 28, 59, 62, 12)$ and $Y^T = (7, 35, 94, 156, 168)$. The cumulative output by time 4 is 156.

18.11 (a) The maximal lead time is 2 time units.

(b) We have:

$$\begin{aligned}
\hat{\Pi}_1 &= \int_0^1 [3(1-\tau)^2]\left[1 - \left(1 - \frac{1-\tau}{2}\right)^3\right] d\tau \\
&= 3\int_0^1 [(7/8) - (17/8)\tau + (5/4)\tau^2 + \tau^3/4 - \tau^4/8 - \tau^5/8] d\tau \\
&= 3[(7/8)\tau - (17/16)\tau^2 + (5/12)\tau^3 + \tau^4/16 - \tau^5/40 - \tau^6/48 \Big|_0^1 \\
&= 59/80 = 0.7375. \\
\hat{\Pi}_2 &= \int_0^1 [3(1-\tau)^2]\left[1 - \left(1 - \frac{2-\tau}{2}\right)^3\right] d\tau \\
&= 3\int_0^1 [1 - 2\tau + \tau^2 + \tau^3/8 + \tau^4/4 - \tau^5/8] d\tau \\
&= 3[\tau - \tau^2 + \tau^3/3 - \tau^4/32 + \tau^5/20 - \tau^6/48 \Big|_0^1 \\
&= 477/480 = 0.99375.
\end{aligned}$$

Thus, $\hat{\pi}_1 = \hat{\Pi}_1 = 59/80$, $\hat{\pi}_2 = \hat{\Pi}_2 - \hat{\Pi}_1 = 477/480 - 59/80 = 41/160$, and $\hat{\pi}_3 = \hat{\Pi}_3 - \hat{\Pi}_2 = 1 - 477/480 = 1/160$.

(c) Using the notation of the chapter,

$$\begin{aligned}
24(59/80, 41/160, 1/160) &\longrightarrow (17.7, 6.15, 0.15, 0, 0), \\
48(59/80, 41/160, 1/160) &\longrightarrow (0, 35.4, 12.3, 0.3, 0), \\
96(59/80, 41/160, 1/160) &\longrightarrow (0, 0, 70, 8, 24.6, 0.6),
\end{aligned}$$

and so $y^T = (17.7, 41.55, 83.25, 24.9, 0.6)$. (d) We have
$Y^T = (17.7, 59.25, 142.5, 167.4, 168)$, and so the cumulative output by time 3 is 142.5.

18.12 The 144 starts in $[-1, 0]$ will result in outputs of 78, 60, and 6 in periods 0, 1, and 2. The 192 starts in $[-2, -1]$ will result in outputs of 104, 80, and 8 in periods -1, 0, and 1. The 72 starts in $[-3, -2]$ will result in outputs of 39, 30, and 3 in periods -2, -1, and 0. The total output realized as a result of starts before time 0 is therefore therefore 78 + 184 + 72 = 334, with an additional 68 units to emerge as output by time 1 and another 6 units to emerge by time 2. Thus, the new output vector is $y^T = (81, 42, 73, 42, 4)$. (The 334 units of output could have been used to service demand; if not, they would be added to the inventory counts.)

19

Dynamic Production Function Approximations

In this chapter, we describe several approximations to the true dynamic production function that yield representations amenable for computation. We begin with a description of processes with load-dependent lead times that arise in manufacturing systems. Next, we show how to approximate an ideal description by using just two "boundary" input-output points. An application of this two-point boundary approximation to project-oriented production systems, such as a naval shipyard, is provided. We briefly describe serial and parallel aggregation of detailed dynamic production functions, and show how to extend the steady-state activity analysis models to the dynamic setting.

19.1 Load-Dependent Processes

Shop-floor statistics can be gathered to estimate a lead time distribution. If a system is reasonably stable and if projected input is reasonably consistent with the past input, then a distribution-based dynamic production function will provide a useful and practical model of the output process.

It is both realistic and practical to allow the distribution of lead time to depend on the time at which an input enters the system, since input that enters the system during a normally congested time can be expected to take longer to complete than an input that enters the system at an off-peak time. For capacity-constrained systems, the time it takes to complete an input (or input batch) depends on factors such as availability of key resources, which are often "τ-dependent" (e.g., which day of the week, which shift), and the amount of work-in-process or *"load"* currently in the system. When the lead time is a function of the system load, the process time for a part is no longer independent of past inputs to the system. Below, we describe a simple model to incorporate "load-dependent lead times," and show how to approximate it to arrive at another example of a linear, distribution-based dynamic production function.

19.1.1 Formulation

Let $x(t)$ and $y(t)$ denote, respectively, the instantaneous input and output rates at time t. We assume the input function $x(\cdot)$ is continuous and deterministic. By definition, the difference between the cumulative input, $X(t)$, and the cumulative output, $Y(t)$, at time $t \geq 0$,

$$q(t) := X(t) - Y(t), \tag{19.1}$$

is the number of units still in the system at time t, which is the **work queue** at time t. We assume the instantaneous rate at which input is completed (i.e., leaves the system) at time t is *solely* a function of the work queue, and is given by $\pi(q(t), t)q(t)$ for some continuous, positive function $\pi : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_{++}$. One interprets $\pi(q(t), t)q(t)\Delta t$ as the proportion of the current queue that will be completed in the next Δt units of time. For example,

$$\pi(q(t), t) = e^{-\lambda(t)q(t)}$$

is a convenient and reasonable functional form. Under these modeling assumptions,

$$\overbrace{q(t + \Delta t) - q(t)}^{\text{change in queue}} \approx \overbrace{x(t)\Delta t}^{\text{flow in}} - \overbrace{\pi(q(t), t)q(t)\Delta t}^{\text{flow out}},$$

which, as $\Delta t \to 0$, reduces to the differential equation

$$\dot{q}(t) = \frac{dq}{dt} = x(t) - q(t)\pi(q(t), t). \tag{19.2}$$

When the interval $[t_{i-1}, t_i]$ is sufficiently small,

$$\pi(q(t), t) \approx \pi(q(t_{i-1}), t_{i-1}) := \pi_i$$

on the interval $[t_{i-1}, t_i]$. Setting $\pi(q(t), t)$ to the constant π_i (known at time t_{i-1}), the nonlinear differential equation (19.2) becomes an ordinary differential equation whose solution on $[t_{i-1}, t_i]$ is

$$q(t) = q(t_{i-1})e^{-\pi_i(t-t_{i-1})} + \int_{t_{i-1}}^{t} x(\tau)e^{-\pi_i(t-\tau)}d\tau, \quad t_{i-1} \leq t \leq t_i. \tag{19.3}$$

Construct a uniform time grid whose period lengths equal Δ. Approximate the input function $x(\cdot)$ with the function

$$\hat{x}(t) := x_i \text{ for } t \in [i-1, i],$$

where x_i is taken to be the average of the function $x(\cdot)$ on the interval $[i-1, i]$. (The maximum or minimum of $x(\cdot)$ on this interval will also be reasonable choices; since $x(\cdot)$ is continuous, all three estimates will be close when Δ is sufficiently small.) Let $q_0 := q(0)$ and let $\pi_0 := \pi(q_0, 0)$. For sufficiently small

Δ, (19.3) will be an excellent approximation to the solution for $q(\cdot)$ on the interval $[0, \Delta]$. In particular,

$$q_1 := q(\Delta) \approx e^{-\pi_0 \Delta} q_0 + x_1 \frac{1 - e^{-\pi_0 \Delta}}{\pi_0}.$$

We can sequentially continue the process and define

$$\pi_i := \pi(q_{i-1},\ (i-1)\Delta), \quad i = 1, 2, \ldots, \tag{19.4}$$

and calculate

$$q_{i+1} := q((i+1)\Delta) \ \approx \ q_i e^{-\pi_i \Delta} + x_{i+1} \frac{1 - e^{-\pi_i \Delta}}{\pi_i}, \quad i = 1, 2, \ldots. \tag{19.5}$$

19.1.2 Example

We take $\pi(q(t), t) = e^{-0.01 q(t)}$ and approximate the queue process via a standard time grid. The input vector for the first six periods is

$$x = (x_1, x_2, x_3, x_4, x_5, x_6) = (25, 30, 35, 40, 45, 45).$$

The initial queue size is $q(0) = 100$. Using (19.4) and (19.5), we have

$$\begin{aligned}
\pi_0 &= \pi(q(0), 0) = e^{-0.01(100)} = 0.3679, \\
q_1 &= e^{-\pi_0} q_0 + x_1 \frac{1 - e^{-\pi_0}}{\pi_0} = (0.6922)(100) + 25 \left[\frac{1 - 0.6922}{0.3679} \right] = 90.14, \\
\pi_1 &= e^{-0.01(90.14)} = 0.4060, \\
q_2 &= e^{-\pi_1} q_1 + x_2 \frac{1 - e^{-\pi_1}}{\pi_1} = (0.6663)(90.14) + 30 \left[\frac{1 - 0.6663}{0.4060} \right] = 84.72, \\
\pi_2 &= e^{-0.01(84.72)} = 0.4286, \\
q_3 &= e^{-\pi_2} q_2 + x_3 \frac{1 - e^{-\pi_2}}{\pi_2} = (0.6514)(84.72) + 35 \left[\frac{1 - 0.6514}{0.4286} \right] = 83.65, \\
\pi_3 &= e^{-0.01(83.65)} = 0.4332, \\
q_4 &= e^{-\pi_3} q_3 + x_4 \frac{1 - e^{-\pi_3}}{\pi_3} = (0.6484)(83.65) + 40 \left[\frac{1 - 0.6484}{0.4322} \right] = 86.78, \\
\pi_4 &= e^{-0.01(86.78)} = 0.4199, \\
q_5 &= e^{-\pi_4} q_4 + x_5 \frac{1 - e^{-\pi_4}}{\pi_4} = (0.6571)(86.78) + 45 \left[\frac{1 - 0.6571}{0.4199} \right] = 93.77, \\
\pi_5 &= e^{-0.01(93.77)} = 0.3915, \\
q_6 &= e^{-\pi_5} q_5 + x_6 \frac{1 - e^{-\pi_5}}{\pi_5} = (0.6760)(93.77) + 45 \left[\frac{1 - 0.6760}{0.3915} \right] = 100.63.
\end{aligned}$$

As before, let y_t denote the cumulative output that emerges in period t, i.e., over the interval $[t-1, t]$. Inventory balance states that the queue at

time t equals the starting queue at time $t-1$ plus the new input minus the output. The **inventory balance equations** are

$$q_t = q_{t-1} + x_t - y_t,$$

or equivalently,

$$y_t = q_{t-1} + x_t - q_t, \quad t = 1, 2, \ldots,$$

from which we obtain the y_t as $y_1 = 34.86$, $y_2 = 35.42$, $y_3 = 36.07$, $y_4 = 36.87$, $y_5 = 38.01$, and $y_6 = 38.14$. The average queue L is 91.38 and the average input rate λ is $36.\bar{6}$. Using **Little's Law** $L = \lambda W$, the average time in the system (cycle-time) W is approximately 2.5 periods.

Each π_i depends on $x_1, x_2, \ldots, x_{i-1}$ through q_{i-1}. If the π_i's are exogenously specified so that they are *independent* of the x_i's, then it follows from the recursive equations (19.4) and (19.5) that each q_i is a *linear* function of the x_i's and so are the output and cumulative output functions. This observation can be used to obtain a linear functional approximation to the output curve, as described in the next section.

19.1.3 Linear Approximation

Let $\bar{x}$ denote a baseline input function, and let the $\bar{\pi}$'s denote the solution obtained from (19.4) and (19.5) using $\bar{x}$. If each input function $x(\cdot)$ is in a neighborhood of $\bar{x}(\cdot)$, then the (constant) $\bar{\pi}$'s in (19.4) and (19.5) can be used to obtain the queue function $q(\cdot)$ for $x(\cdot)$ and thus the output curve.

We continue with data provide in the example of Section 19.1.2. Fix $\bar{x}$ to be the input vector x of this example. Let q_t^k denote the amount in queue at time t due to the starts in period $k = 0, 1, \ldots, t$. Here, q_t^0 is the number of units of the original queue at time 0 still in the system at time t. All units in the queue at time t must have been started sometime before time t, and so by definition

$$\sum_{k=0}^{t} q_t^k = q_t.$$

Let y_t^k denote the cumulative amount of output in period t completed from queue q_t^k. To determine the values of the y_t^k and q_t^k, it is necessary to describe the *queue discipline*. We examine two possibilities in the following examples.

Example 19.1. Assume that withdrawal from inventory follows the "First-In, First-Out (FIFO)" discipline. That is, output in any period is sequentially taken from the earliest queues of starts. First, we determine q_t and y_t. The calculations are easiest to explain using the example data. (We leave it to an exercise to algebraically express these derivations.) The results are displayed in Tables 19.1 and 19.2.

The queue q_0 at time 0 is 100 and the cumulative output in period 1 is $y_1 = 34.86$. Thus, the remaining initial queue at time 1 is $q_1^0 = 100 - 34.86 =$

Table 19.1. Queue matrix q_t^k for example 19.1.

	Period						
Queue due to:	0	1	2	3	4	5	6
Initial inventory		65.14	29.72	0	0	0	0
Starts in period 1		25.00	25.00	18.65	0	0	0
Starts in period 2		0	30.00	30.00	11.78	0	0
Starts in period 3		0	0	35.00	35.00	8.77	0.00
Starts in period 4		0	0	0	40.00	40.00	10.63
Starts in period 5		0	0	0	0	45.00	45.00
Starts in period 6		0	0	0	0	0	45.00
Queue at end of period:	100	90.14	84.72	83.65	86.78	93.77	100.63

Table 19.2. Output matrix y_t^k for example 19.1.

	Period					
Output due to:	1	2	3	4	5	6
Initial inventory	34.86	35.42	29.72	0	0	0
Starts in period 1	0	0	6.35	18.65	0	0
Starts in period 2	0	0	0	18.22	11.78	0
Starts in period 3	0	0	0	0	26.23	8.77
Starts in period 4	0	0	0	0	0	29.37
Starts in period 5	0	0	0	0	0	0
Starts in period 6	0	0	0	0	0	0
Output at end of period:	34.86	35.42	36.07	36.87	38.01	38.14
Cumulative output at end of period:	34.86	70.28	106.35	143.22	181.23	219.37

65.14. None of the 25 starts in period 1 emerges as output by time 1; they all enter a queue and so $q_1^1 = 25$. With respect to output, all 34.86 units are taken from the initial queue, which implies that $y_1^0 = 34.86$ and $y_1^1 = 0$.

The output in period 2 is $y_2 = 35.42$. All of this output is taken from the initial queue whose size at time 2 is now $q_2^0 = 65.14 - 35.42 = 29.72$. None of the 25 starts in period 1 or 30 starts in period 2 emerge as output by time 2; they all enter their respective queues and so $q_2^1 = 25$ and $q_2^2 = 30$. With respect to output, all 35.42 units are taken from the initial queue, which implies that $y_2^0 = 35.42$, $y_2^1 = y_2^2 = 0$.

The output in period 3 is $y_3 = 36.07$. This quantity is first drawn from the queue of initial starts whose size at the beginning of this period is 29.72. The remaining output of 36.07 - 29.42 = 6.35 is withdrawn from the queue of starts from period 1 whose size at the beginning of this period is 25. None of the 30 starts in period 2 or 35 starts in period 3 emerge as output by time 3; they all enter their respective queues and so $q_3^2 = 30$ and $q_3^3 = 35$. With

respect to output, we have $y_2^0 = 29.72$, $y_2^1 = 6.35$, and $y_3^2 = y_3^3 = 0$. The process continues.

Next, we convert the entries in Table 19.2 to percentages of the respective starts—see Table 19.3. For example, the entries in the third column are obtained as 29.72/100 and 6.35/25, respectively; the entries in the fourth column are obtained as 18.65/25 and 18.22/30, respectively, and so on. We can use these percentages to estimate the output in each period as a linear combination of the starts, as follows:

$$\begin{aligned} y_1 &= 0.349q_0, \\ y_2 &= 0.354q_0, \\ y_3 &= 0.297q_0 + 0.254x_1, \\ y_4 &= 0.746x_1 + 0.607x_2, \\ y_5 &= 0.393x_2 + 0.749x_3, \\ y_6 &= 0.251x_3 + 0.734x_4. \end{aligned}$$

These equations will be reasonably accurate as long as the input vector x is reasonably close to $\bar{x}$ and the initial queue size is about 100.

Table 19.3. Percentage of starts in period k emerging as output in period t for example 19.1.

	Period					
	1	2	3	4	5	6
Initial inventory	0.349	0.354	0.297	0	0	0
Starts in period 1	0	0	0.254	0.746	0	0
Starts in period 2	0	0	0	0.607	0.393	0
Starts in period 3	0	0	0	0	0.749	0.251
Starts in period 4	0	0	0	0	0	0.734
Starts in period 5	0	0	0	0	0	0
Starts in period 6	0	0	0	0	0	0

Example 19.2. In this example, we examine the case when inventory withdrawal is *simultaneously* taken from each queue in the same proportions. To ease notational burdens, define

$$\begin{aligned} \alpha_i &:= e^{-\pi_i}, \ i = 0, 1, \ldots, \\ \beta_i &:= 1 - \frac{1 - e^{-\pi_i}}{\pi_i}, \ i = 0, 1, \ldots \ . \end{aligned}$$

Using the notation of Example 19.1,

$$
\begin{aligned}
q_t^k &= \alpha_k q_{t-1}^k,\ 0 \le k \le t-1,\\
q_t^t &= q_t - \sum_{k=0}^{t-1} q_t^k = x_t - y_t^t,\\
y_t^k &= (1-\alpha_k) q_{t-1}^k = q_{t-1}^k - q_t^k,\ 0 \le k \le t-1,\\
y_t^t &= y_t - \sum_{k=0}^{t-1} y_t^k = \beta_t x_t\ .
\end{aligned}
$$

Table 19.4. Queue matrix q_t^k for example 19.2.

	Period						
Queue due to:	0	1	2	3	4	5	6
Initial inventory		69.22	46.12	30.05	19.48	12.80	8.65
Starts in period 1		20.92	13.94	9.08	5.89	3.87	2.62
Starts in period 2		0	24.66	16.07	10.42	6.85	4.63
Starts in period 3		0	0	28.45	18.45	12.12	8.19
Starts in period 4		0	0	0	32.56	21.40	14.47
Starts in period 5		0	0	0	0	36.76	24.85
Starts in period 6		0	0	0	0	0	37.26
Queue at end of period:	100	90.14	84.72	83.65	86.78	93.77	100.63

Table 19.5. Output matrix y_t^k for example 19.2.

	Period					
Output due to:	1	2	3	4	5	6
Initial inventory	30.78	23.10	16.07	10.57	6.68	4.15
Starts in period 1	4.08	6.98	4.86	3.19	2.02	1.25
Starts in period 2	0	5.34	8.59	5.65	3.57	2.22
Starts in period 3	0	0	6.55	10.00	6.33	3.93
Starts in period 4	0	0	0	7.44	11.16	6.93
Starts in period 5	0	0	0	0	8.24	11.91
Starts in period 6	0	0	0	0	0	7.74
Output at end of period:	34.86	35.42	36.07	36.87	38.01	38.14
Cumulative output at end of period:	34.86	70.28	106.35	143.22	181.23	219.37

These equations are best illustrated with the example data. The results are displayed in Tables 19.4 and 19.5. The α and β vectors are

$$
\begin{aligned}
\alpha &= (\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) = (0.692, 0.666, 0.651, 0.648, 0.657, 0.676),\\
\beta &= (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = (0.163, 0.178, 0.187, 0.186, 0.183, 0.172).
\end{aligned}
$$

The queue q_1 at time 1 is 90.14 of which

$$(0.6922)(100) = 69.22 = q_1^0$$

is due to the initial queue with the remaining queue

$$q_1^1 = 90.14 - 69.22 = 20.92$$

due to the starts in period 1. With respect to output,

$$y_1^0 = (0.3078)(100) = 100 - 69.22 = 30.78$$

of the initial queue of 100 emerge as output by time 1, and thus

$$y_1^1 = 34.86 - 30.78 = 4.08 = (0.163)(25)$$

units of output emerge as output by time 1 due to the starts in period 1. Moving to the next period, we have $\alpha_1 = 0.6663$ and $\beta_1 = 0.178$. The queue q_2 at time 2 is 84.72 of which

$$(0.6663)(69.22) = 46.12 = q_2^0$$

is due to the initial queue,

$$(0.6663)(20.92) = 13.94 = q_2^1$$

is due to the starts in period 1 with the remaining queue

$$q_2^2 = 84.72 - (46.12 + 13.94) = 24.66$$

due to the starts in period 2. With respect to output,

$$y_2^0 = (0.3337)(69.22) = 69.22 - 46.12 = 23.10$$

of the remaining queue of 69.22 (of the initial queue of 100) emerge as output by time 2,

$$y_2^1 = (0.3337)(20.92) = 20.92 - 13.94 = 6.98$$

of the remaining queue of 20.92 (of the 25 starts in period 1) emerge as output by time 2, and thus

$$y_2^2 = 35.42 - (23.10 + 6.98) = 5.34 = (0.178)(30)$$

units of output emerge as output by time 2 due to the starts in period 2. The remaining calculations proceed accordingly.

Next, we convert the entries in Table 19.5 to percentages of the respective starts—see Table 19.6. For example, the entries in the second column are obtained as 23.10/100, 6.98/25, and 5.34/30, respectively; the entries in the third column are obtained as 16.07/100, 4.86/25, 8.59/30, and 6.55/35,

respectively, and so on. We can use these percentages to estimate the output in each period as a linear combination of the starts, as follows:

$$
\begin{aligned}
y_1 &= 0.308q_0 + 0.163x_1,\\
y_2 &= 0.231q_0 + 0.279x_1 + 0.178x_2,\\
y_3 &= 0.161q_0 + 0.194x_1 + 0.286x_2 + 0.187x_3,\\
y_4 &= 0.106q_0 + 0.128x_1 + 0.188x_2 + 0.286x_3 + 0.186x_4,\\
y_5 &= 0.067q_0 + 0.081x_1 + 0.119x_2 + 0.181x_3 + 0.279x_4 + 0.183x_5,\\
y_6 &= 0.042q_0 + 0.050x_1 + 0.074x_2 + 0.112x_3 + 0.173x_4 + 0.265x_5 + 0.183x_6.
\end{aligned}
$$

Table 19.6. Percentage of starts in period k emerging as output in period t for example 19.2.

	Period					
	1	2	3	4	5	6
Initial inventory	0.308	0.231	0.161	0.106	0.067	0.042
Starts in period 1	0.163	0.279	0.194	0.128	0.081	0.050
Starts in period 2	0	0.178	0.286	0.188	0.119	0.074
Starts in period 3	0	0	0.187	0.286	0.181	0.112
Starts in period 4	0	0	0	0.186	0.279	0.173
Starts in period 5	0	0	0	0	0.183	0.265
Starts in period 6	0	0	0	0	0	0.183

The linear equations that translate the input vector x and initial queue q_0 into outputs y_t over time in Examples 19.1 and 19.2 will be reasonably accurate as long as the input vector x is sufficiently close to $\bar{x}$ and the initial queue size is about 100. (The initial queue size is a known value at time 0.) There are several, common ways to measure closeness.

Example 19.3. The input vector x is sufficiently close to the input vector $\bar{x}$ if $|x_t - \bar{x}_t| < \epsilon_t$ for all t. Here, the ϵ_t represent the allowable tolerances within each period.

Example 19.4. The input vector x is sufficiently close to the input vector $\bar{x}$ if $\sum_t \omega_t |x_t - \bar{x}_t| < \epsilon$. Here, the ω_t represent the relative importance of being sufficiently close in each time period and ϵ represents the degree of tolerance.

Example 19.5. The input vector x is sufficiently close to the input vector $\bar{x}$ if $\sum_t \omega_t (x_t - \bar{x}_t)^2 < \epsilon$. This is a nonlinear (convex) extension to the previous example. The use of the square ensures a more uniform distribution of error.

There can be several candidate input curves represented by $\bar{x}^c(\cdot)$, $c = 1, 2, \ldots, C$. The coefficients that translate each candidate input curve $\bar{x}^c$ and

initial queue q_0 into the output curve can be computed, as in Examples 19.1 and 19.2. If the input curve $x(\cdot)$ is in the neighborhood of $\bar{x}^c$, then that system of linear equations should be used to represent the input-output process.

19.1.4 Load-Dependent, Linear Approximation

We have described a single-input, single output technology in which the distribution of output depends on the system load or work queue $q(\tau)$ at time τ. An approximation that expresses the input-output transformation via a collection of linear equations has been developed. The coefficients that characterize this linear system *presuppose knowledge of the input curve*, either its exact shape or the neighborhood of which pre-specified candidate input curve it lies in. It is desirable and practical to allow the coefficients of the linear system to *depend* on the input curve. In this section, we extend the continuous lead time representation of Section 18.4 to represent this more general model of technology.

For a continuous lead time process, the cumulative output at time t is

$$Y(t) = \int_{-\infty}^{t} z(\tau) W(t-\tau, \tau) d\tau,$$

where $z(\cdot)$ is the index of starts and $W(\cdot, \cdot)$ is a lead time distribution. For a discrete-time approximation using a standard time grid, cumulative output is

$$Y_t = \sum_{\tau \le t} z_\tau \Pi_{\tau,t}. \tag{19.6}$$

See Section 18.4, p. 325. In (19.6), the input-output transformation is represented via a system of linear equations. The parameter $\Pi_{\tau,t}$ can be an *exogenous* function of τ-dependent information, but, as written, is not a function of the work queue

$$q_\tau = q_{\tau-1} + z_\tau - y_\tau \tag{19.7}$$

at time τ. Conceptually, the $W(\cdot, \tau)$ in (19.6) can be replaced with $W(\cdot, q(\tau), \tau)$ so that

$$Y_t = \sum_{\tau \le t} z_\tau \Pi_{q(\tau),\tau,t}. \tag{19.8}$$

In (19.8), the lead time distribution is now permitted to be a function of the $q(\tau)$ and other exogenous state information at time τ. Expression (19.8) appears innocuous. Unfortunately, without further approximation, this more general input-output representation is no longer characterized by a linear system; consequently, it is not directly amenable for computation via optimization software. The difficulty lies in an inherent *recursion*. If the subsequent

distribution of output due to starts at time τ depends on the work queue $q(\tau)$, then it follows from (19.7) that $q(\tau)$ *itself* depends on the *prior* history of input $\{z(s) : s \le \tau\}$ up to time τ. (Revisit Examples 19.1 and 19.2.)

Here is a practical way to represent the lead time distribution as a function of the work queue. Let $y_{\tau,t}$ denote the output in period t due to the input in period $\tau \le t$, and let y_t denote the output in period t as a result of all past input including the initial queue, q_0. Let

$$0 := q^0 < q^1 < q^2 < \cdots < q^L$$

denote pre-specified *queue levels*. (The last level, q^L, is sufficiently high to bound above any realized queue size.) For each τ, t such that $\tau \le t$, the new model for output is

$$y_{\tau,t} := z_\tau \Pi^\ell_{\tau,t}, \quad \text{if } q(\tau) \in [q^{\ell-1}, q^\ell), \tag{19.9}$$

$$y_t := \sum_{\tau \le t} y_{\tau,t} \; + \; q_0 \Pi_{0,t}. \tag{19.10}$$

The parameters $\Pi_{0,t}$ define the initial lead time distribution, which is a function of the initial queue q_0. In (19.6), $y_{\tau,t} = z_\tau \Pi_{\tau,t}$, independent of $q(\tau)$; in (19.9), $\Pi_{\tau,t}$ is replaced with $\Pi^\ell_{\tau,t}$ if $q(\tau) \in [q^{\ell-1}, q^\ell)$.

It remains to implement the logical expression (19.9). This can be achieved by using *binary* variables and adding two sets of linear constraints, as follows.

- *Output constraints.* For each τ, t such that $\tau \le t$, and each $\ell = 1, 2, \ldots, L$, add these two constraints:

$$y_{\tau,t} \le z_\tau \Pi^\ell_{\tau,t} \; + \; M(1 - \xi_{\tau,\ell}), \tag{19.11}$$

$$y_{\tau,t} \ge z_\tau \Pi^\ell_{\tau,t} \; - \; M(1 - \xi_{\tau,\ell}), \tag{19.12}$$

  where M is a sufficiently large number, and

$$\xi_{\tau,\ell} := \begin{cases} 1, & \text{if } q(\tau) \in [q^{\ell-1}, q^\ell), \\ 0, & \text{otherwise.} \end{cases} \tag{19.13}$$

  If $\xi_{\tau,\ell} = 0$, then (19.11) and (19.12) constrain $y_{\tau,t}$ to lie in the interval $(-M, M)$, which, to all intents and purposes, is equivalent to $(-\infty, \infty)$. In this case, the constraints (19.11) and (19.11) are automatically satisfied. On the other hand, if $\xi_{\tau,\ell} = 1$, then (19.11) and (19.12) constrain $y_{\tau,t}$ to lie both above and below $z_\tau \Pi^\ell_{\tau,t}$. Obviously, this can only happen if $y_{\tau,t} = z_\tau \Pi^\ell_{\tau,t}$.
- *Queue constraints.* The logical variables $\xi_{\tau,\ell}$ must be linked to the queue values, the q_τ, so as to be consistent with their intended purpose. In conjunction with the logical constraints

$$\sum_{\ell=1}^{L} \xi_{\tau,\ell} = 1, \quad \text{for each } \tau, \tag{19.14}$$

this is accomplished by adding two new constraints for each time period τ:

$$q_\tau \leq \sum_{\ell=1}^{L} \xi_{\tau,\ell} q^\ell, \tag{19.15}$$

$$q_\tau \geq \sum_{\ell=1}^{L} \xi_{\tau,\ell} q^{\ell-1}. \tag{19.16}$$

Suppose $q_\tau \in (q^{\hat{\ell}-1}, q^{\hat{\ell}})$. Since the $\xi_{\tau,\ell}$ are binary variables, for each τ, (19.7) implies there will be exactly one index $\ell(\tau)$ such that $\xi_{\tau,\ell(\tau)} = 1$. To satisfy constraints (19.15) and (19.16), the only choice for $\ell(\tau)$ is $\hat{\ell}$.

In sum, the proposed load-dependent, continuous lead time model uses constraints (19.7) and (19.11)-(19.16). An extension to the multiple product case is described in Section 21.6.6, p. 414.

Remark 19.6. Presumably, the lead time distribution is a continuous function of the queue size. When there are a finite number of queue levels, the lead time distribution does not change between the queue levels. Ostensibly, this presents a problem when $q_\tau = q^{\hat{\ell}}$ for some $\hat{\ell}$, since either $\xi_{\tau,\hat{\ell}} = 1$ or $\xi_{\tau,\hat{\ell}+1} = 1$ is consistent with (19.14)-(19.16). In practice, the queue levels q^ℓ are perturbed by a rational ϵ so that no realized queue value will precisely equal a queue level.

Remark 19.7. The number of binary variables is $L \cdot T$. At present, to be computationally tractable, this number should be no more than several hundred. For typical values of T, the number of different lead time distributions will be on the order of 10 or so. (The number of additional constraints is order $L \cdot T$, which is modest for reasonable values of L and T.)

19.2 Two-Point Boundary Approximation

In this section, the cumulative index function $Z(\cdot)$ is bounded above by $Z^U(\cdot)$ and bounded below by $Z^L(\cdot)$, so that the domain D of the index of the dynamic production function is

$$\mathsf{D} := \{Z(\cdot) : \ Z^L(t) \leq Z(t) \leq Z^U(t)\}. \tag{19.17}$$

Assume the corresponding cumulative output functions $Y^U(\cdot) := [F(Z^U)](\cdot)$ and $Y^L(\cdot) := [F(Z^L)](\cdot)$ are known. The goal is to extend $F[\cdot]$ to all of D. We call such an extension a **two-point boundary approximation** and denote the approximation as $F^a[\cdot]$. In what follows all functions of time are assumed differentiable. We let T^{in}, T^{out} denote, respectively, the finite points in time beyond which there is no further input nor output; of course, $0 \leq T^{in} \leq T^{out}$.

For starters, the approximation $F^a[\cdot]$ should satisfy the following two basic properties:

$$[F^a(Z^L)](\cdot) = Y^L(\cdot), \quad [F^a(Z^U)](\cdot) = Y^U(\cdot), \tag{19.18}$$

$$Y^L(\cdot) \le [F^a(Z)](\cdot) \le Y^U(\cdot) \text{ for each } Z(\cdot) \in \mathsf{D}. \tag{19.19}$$

It is also reasonable to insist $F^a[\cdot]$ is continuous, namely, the "distance" between $[F^a(Z)](\cdot)$ and $[F^a(Z')](\cdot)$ should be small when the "distance" between $Z(\cdot)$ and $Z'(\cdot)$ is sufficiently small. Assuming continuity is not sufficient to pin down an exact choice for $F^a[\cdot]$. We shall define a distance measure that will uniquely determine $F^a[\cdot]$.

19.2.1 Relative Area Ratio

Consider a cumulative input curve $Z(\cdot) \in \mathsf{D}$. One simple way to measure how $Z(\cdot)$ lies between its boundary curves, $Z^U(\cdot)$ and $Z^L(\cdot)$, on the interval $[0, t]$ is the **relative area ratio**

$$r[Z; Z^U; Z^L](t) := \frac{\int_0^t [Z(\tau) - Z^L(\tau)]d\tau}{\int_0^t [Z^U(\tau) - Z^L(\tau)]d\tau} . \tag{19.20}$$

The value $r[Z; Z^U; Z^L](t)$ is between 0 and 1. If this function of time is close to 1, then it indicates that $Z(\cdot)$ is close to $Z^U(\cdot)$, and if this function of time is close to 0, then it indicates that $Z(\cdot)$ is close to $Z^L(\cdot)$. Analogously, a measure of how $[F^a(Z)](\cdot)$ lies between its boundary curves, $[F(Z^U)](\cdot)$ and $[F(Z^L)](\cdot)$, on the interval $[0, t]$ is the relative area ratio

$$r[F^a(Z); Y^U; Y^L](t) := \frac{\int_0^t ([F^a(Z)](\tau) - Y^L(\tau))d\tau}{\int_0^t (Y^U(\tau) - Y^L(\tau))d\tau} . \tag{19.21}$$

This value is between 0 and 1. If the function of time is close to 1, then $[F^a(Z)](\cdot)$ is close to $Y^U(\cdot)$, and if this function of time is close to 0, then $[F^a(Z)](\cdot)$ is close to $Y^L(\cdot)$. The respective relative area ratio functions of time uniquely determine the input function $Z(\cdot)$ or the output function $F^a(Z)(\cdot)$.

Output can be an intermediate good required by a follow-on activity or can be a final good. Let $X(\cdot)$ denote the cumulative input function obtained from all consumers of this output. For each $Z(\cdot) \in \mathsf{D}$, the set of feasible cumulative input functions is

$$\Omega(Z) := \{X(\cdot) : X(t) \le [F^a(Z)](t),\ t \ge 0\}.$$

The relative area function $r[F^a(Z); Y^U; Y^L]$ provides a useful way to measure the "size" of $\Omega(Z)$, namely, the degree to which $[F^a(Z)](\cdot)$ constrains the possible choices for $X(\cdot)$. From this perspective, it seems reasonable to insist that the degree to which $[F^a(Z)](\cdot)$ constrains the possible choices for $X(\cdot)$ over time, as measured by $r[F^a(Z); Y^U; Y^L](\cdot)$, should be closely in line

with how $Z(\cdot)$ fits between its boundary curves over time, as measured by $r[Z; Z^U; Z^L](\cdot)$. In fact, $[F^a(Z)](\cdot)$ is uniquely determined if we insist these two measures are identical, i.e.,

$$r[F^a(Z); Y^U; Y^L](t) = r[Z; Z^U; Z^L](\rho(t)), \tag{19.22}$$

where $\rho(\cdot)$ is a differentiable, strictly increasing function of time such that

(i) $\rho(0) = 0$,
(ii) $\rho(t) \leq t$ for all $t \geq 0$, and
(iii) $\rho(T^{out}) = T^{in}$.

Since production may not be instantaneous, there is a need to incorporate the function $\rho(\cdot)$ in (19.22) to model the possibility that output realizations over the interval $[0, t]$ are a result of input over the interval $[0, \rho(t)]$, where $\rho(t)$ can be less than t.

19.2.2 Linear Approximation

Identity (19.22) can be expressed as

$$\int_0^t ([F^a(Z)](\tau) - Y^L(\tau))d\tau = \mathcal{R}_\rho(t) \int_0^{\rho(t)} [Z(\tau) - Z^L(\tau)]d\tau, \tag{19.23}$$

where

$$\mathcal{R}_\rho(t) := \frac{\int_0^t (Y^U(\tau) - Y^L(\tau))d\tau}{\int_0^{\rho(t)} [Z^U(\tau) - Z^L(\tau)]d\tau}. \tag{19.24}$$

Differentiating both sides of (19.23) with respect to t,

$$[F^a(Z)](t) - Y^L(t) = \mathcal{R}_\rho(t)\rho'(t)[Z(\rho(t)) - Z^L(\rho(t))] + \mathcal{R}'_\rho(t) \int_0^{\rho(t)} [Z(\tau) - Z^L(\tau)]d\tau. \tag{19.25}$$

To *uniquely* determine $\rho(\cdot)$ and $F^a(Z)(\cdot)$, we assume the cumulative output up to time t is a function of the *quantity* of cumulative input up to time $\rho(t)$ but does not depend on the structure of the input *before* time $\rho(t)$. For this assumption to hold, the integral on the right-hand-side of (19.25) must vanish, which implies the derivative $\mathcal{R}'_\rho(t)$ must always be zero. Thus, $\mathcal{R}_\rho(t)$ in (19.24) is a constant, say equal to R. Any one value for $\rho(t)$ will determine the constant R from which the identity (19.24) implicitly determines $\rho(t)$ for all t. In particular, since $\rho(T^{in}) = T^{out}$, it follows that

$$R = \frac{\int_0^{T^{out}} [Y^U(\tau) - Y^L(\tau)]d\tau}{\int_0^{T^{in}} [Z^U(\tau) - Z^L(\tau)]d\tau}. \tag{19.26}$$

Consequently, $\rho(\cdot)$ is implicitly defined via the identity

$$\frac{\int_0^t [Y^U(\tau) - Y^L(\tau)]d\tau}{\int_0^{T^{out}} [Y^U(\tau) - Y^L(\tau)]d\tau} = \frac{\int_0^{\rho(t)} [Z^U(\tau) - Z^L(\tau)]d\tau}{\int_0^{T^{in}} [Z^U(\tau) - Z^L(\tau)]d\tau} \tag{19.27}$$

for all $0 \leq t \leq T^{out}$.

Using $\mathcal{R}_\rho^{'}(t) = 0$ and $\mathcal{R}_\rho(\cdot) = R$ in (19.25), for each $Z(\cdot) \in \mathsf{D}$ it follows that

$$[F^a(Z)](t) - Y^L(t) = R\rho^{'}(t)[Z(\rho(t)) - Z^L(\rho(t))]. \tag{19.28}$$

Since (19.28) holds for $Z(\cdot) = Z^U(\cdot)$, it is also true that

$$[Y^U(t) - Y^L(t)] = R\rho^{'}(t)[Z(\rho(t)) - Z^L(\rho(t))]. \tag{19.29}$$

Dividing each side of (19.28) by the corresponding sides of (19.29) yields

$$\frac{[F^a(Z)](t) - [F^a(Z^L)](t)}{Y^U(t) - Y^L(t)} = \frac{Z(\rho(t)) - Z^L(\rho(t))}{Z^U(\rho(t)) - Z^L(\rho(t))}\,. \tag{19.30}$$

The approximate dynamic production function uniquely determined from (19.30) is

$$[F^a(Z)](t) = Y^L(t) + \left(\frac{Y^U(t) - Y^L(t)}{Z^U(\rho(t)) - Z^L(\rho(t))}\right)\,[Z(\rho(t)) - Z^L(\rho(t))], \tag{19.31}$$

where $\rho(\cdot)$ is (implicitly) defined by (19.27).

The function $\rho(\cdot)$ is a *nonlinear* function of time; however, it is easily precomputable. The function $[F(Z)](\cdot)$ is *always* a linear function of the underling $z(\cdot)$ variables when $Z(\cdot)$ is piecewise linear, thus facilitating computational analyses. We illustrate how this works with an example in the next section.

19.2.3 Example

The boundary curves for the example are as follows:

$$Z^U(t) = \begin{cases} 3t, & 0 \leq t \leq 2, \\ 2t + 2, & 2 \leq t \leq 6, \\ 14, & 6 \leq t \leq 8. \end{cases}$$

$$Z^L(t) = \begin{cases} 0, & 0 \leq t \leq 2, \\ t - 2, & 2 \leq t \leq 5, \\ 3.\bar{6}t - 15.\bar{3}, & 5 \leq t \leq 8. \end{cases}$$

$$Y^U(t) = \begin{cases} 0, & 0 \leq t \leq 3, \\ 2.5t - 7.5, & 3 \leq t \leq 7, \\ 2t - 4, & 7 \leq t \leq 12, \\ 20, & 12 \leq t \leq 15. \end{cases}$$

$$Y^L(t) = \begin{cases} 0, & 0 \le t \le 5, \\ 0.5t - 2.5, & 5 \le t \le 9, \\ 3t - 25, & 9 \le t \le 15. \end{cases}$$

For this example, $T^{in} = 8$ and $T^{out} = 15$. A standard time grid will be used. In discrete-time the z_t variables represent the constant rates in each period t. Given these boundary curves, we will show how to express $[F^a(Z)](\cdot)$ defined in (19.31) as a linear function of the z_t variables.

To compute the $\rho(\cdot)$ function, it is necessary to compute the cumulative area between the boundary curves as a function of time. These two curves, denoted respectively by $\mathcal{A}_1(\cdot)$ and $\mathcal{A}_2(\cdot)$, are as follows:

$$\mathcal{A}_1(t) = \begin{cases} 1.5t^2, & 0 \le t \le 2, \\ 0.5t^2 + 4t - 4, & 2 \le t \le 5, \\ -0.8\bar{3}t^2 + 17.\bar{3}t - 65.8\bar{3}, & 5 \le t \le 6, \\ -1.8\bar{3}t^2 + 29.\bar{3}t - 110, & 6 \le t \le 8. \end{cases}$$

$$\mathcal{A}_2(t) = \begin{cases} 0, & 0 \le t \le 3, \\ 1.25t^2 - 7.5t + 11.25, & 3 \le t \le 5, \\ t^2 - 5t + 5, & 5 \le t \le 7, \\ 0.75t^2 - 1.5t - 7.25, & 7 \le t \le 9, \\ -0.5t^2 + 21t - 108.5, & 9 \le t \le 12, \\ -1.5t^2 + 45t - 252.5, & 12 \le t \le 15. \end{cases}$$

For future reference,

$$\mathcal{A}_1(2) = 6,\ \mathcal{A}_1(5) = 28.5,\ \mathcal{A}_1(6) = 36.\bar{6},\ \mathcal{A}_1(8) = 44,$$

$$\mathcal{A}_2(2) = 6,\ \mathcal{A}_2(5) = 5,\ \mathcal{A}_2(7) = 19,\ \mathcal{A}_2(9) = 20,\ \mathcal{A}_2(12) = 71.5,\ \mathcal{A}_2(15) = 85.$$

Example 19.8. The derivation for the functional form for $\mathcal{A}_2(\cdot)$ on the interval [3, 5] is obtained by integrating the difference in the boundary curves

$$\int_3^t (2.5\tau - 7.5)d\tau = 1.25\tau^2 - 7.5\tau \Big|_3^t = 1.25t^2 - 7.5t + 11.25.$$

The functional form for $\mathcal{A}_2(\cdot)$ on the interval [5, 7] is obtained by first integrating the difference in the boundary curves to obtain the *incremental* area

$$\int_5^t [(2.5\tau - 7.5) - (0.5\tau - 2.5)]d\tau = \int_5^t [(2\tau - 5)d\tau = \tau^2 - 5\tau \Big|_5^t = t^2 - 5t,$$

and then adding $\mathcal{A}(5) = 5$ to this expression to obtain the cumulative area. The remaining functional forms are derived in the same fashion.

We illustrate the calculations for three different values of t, $t = 4$, $t = 5$, and $t = 8$. Fix $t = 4$. The total area between the boundary curves $Z^U(\cdot)$ and

$Z^L(\cdot)$ is 44; the total area between the boundary curves $Y^U(\cdot)$ and $Y^L(\cdot)$ is 85. Since $\mathcal{A}_2(4) = 1$, the relative area ratio at time t is 1/85, which implies that $\rho(t)$ must satisfy $\mathcal{A}_1(\rho(t)) = (1/85)44 = 0.5176$. Since $\mathcal{A}_1(2) = 6$, it follows that

$$1.5(\rho(t))^2 = 0.5176 \Rightarrow \rho(t) = 0.5875.$$

Now $Z^U(0.5875) = 1.7625$, $Z^L(0.5875) = 0$, $Y^U(4) = 2.5$, and $Y^L(4) = 0$. Furthermore, $Z(0.5875) = 0.5875z_1$. Substituting these quantities into (19.31),

$$[F^a(Z)](4) = \frac{2.5}{1.7625}\{0.5875z_1\} = 0.8\bar{3}z_1.$$

Fix $t = 5$. Since $\mathcal{A}_2(5) = 5$, the relative area ratio at time t is 5/85, which implies that $\rho(t)$ must satisfy $\mathcal{A}_1(\rho(t)) = (5/85)44 = 2.5882$. Since $\mathcal{A}_1(2) = 6$, it follows that

$$1.5(\rho(t))^2 = 2.5882 \Rightarrow \rho(t) = 1.3136.$$

Now $Z^U(1.3136) = 3.9407$, $Z^L(1.3136) = 0$, $Y^U(5) = 10$, and $Y^L(5) = 0$. Furthermore, $Z(1.3136) = z_1 + 0.3136z_2$. Substituting these quantities into (19.31),

$$[F^a(Z)](5) = \frac{10}{3.9407}\{z_1 + 0.3136z_2\} = 2.5376z_1 + 0.7958z_2.$$

Fix $t = 8$. Since $\mathcal{A}_2(8) = 28.75$, the relative area ratio at time t is 28.75/85, which implies that $\rho(t)$ must satisfy $\mathcal{A}_1(\rho(t)) = (28.75/85)44 = 14.8823$. Since $\mathcal{A}_1(2) = 6 < 14.8823 < \mathcal{A}_1(5)$, it follows that

$$\begin{aligned} 0.5(\rho(t))^2 + 4\rho(t) - 4 = 14.8823 &\Rightarrow \rho(t) \\ &= -4 + \sqrt{16 + 4(18.8823)(0.5)} = 3.3324. \end{aligned}$$

Now $Z^U(3.3324) = 8.6648$, $Z^L(3.3324) = 1.3324$, $Y^U(8) = 12$, and $Y^L(8) = 1.5$. Furthermore, $Z(3.3324) = z_1 + z_2 + z_3 + 0.3324z_4$. Substituting these quantities into (19.31),

$$\begin{aligned} [F^a(Z)](8) &= 1.5 + \left(\frac{12 - 1.5}{8.6648 - 1.3324}\right)\{z_1 + z_2 + z_3 + 0.3324z_4 - 1.3324\} \\ &= 1.4320z_1 + 1.4320z_2 + 1.4320z_3 + 0.4760z_4 - 0.4080. \end{aligned}$$

19.2.4 Extensions

Generically, consider two boundary curves $B^L(\cdot)$ and $B^U(\cdot)$ such that $0 \leq B^L(\cdot) \leq B^U(\cdot)$, and let

$$\mathcal{B} := \{B(\cdot) : \ B^L(\cdot) \leq B(\cdot) \leq B^U(\cdot)\}.$$

For each $B(\cdot) \in \mathcal{B}$ let $\mathcal{M}[B; B^L, B^U](t)$ denote a nonnegative scalar measure of how $B(\cdot)$ "fits" between the two boundary curves on the interval $[0, t]$. It is natural to insist the measure is constant at each boundary curve, and we assume $\mathcal{M}$ has been normalized so that

$$\mathcal{M}[B^U; B^L, B^U](\cdot) = 1,\ \mathcal{M}[B^L; B^L, B^U](\cdot) = 0.$$

With this normalization, it is then natural to insist that $\mathcal{M}$ be nondecreasing in $B(\cdot)$, namely,

$$\mathcal{M}[B; B^L, B^U] \leq \mathcal{M}[B'; B^L, B^U] \text{ if } B(\cdot) \leq B'(\cdot).$$

Given $\rho(\cdot)$ and a measure $\mathcal{M}[\cdot]$, the abstraction of (19.22) is to insist the measure of how the cumulative output curve $[F^a(Z)](\cdot)$ fits between its two boundary curves, $[F^a(Z^U)](\cdot)$, $[F^a(Z^L)](\cdot)$, on the interval $[0, t]$ should be identical to the measure of how the cumulative input curve $Z(\cdot)$ fits between its two boundary curves, $Z^U(\cdot)$, $Z^L(\cdot)$, on the interval $[0, \rho(t)]$. That is, the approximate dynamic production function $F^a[\cdot]$ is implicitly defined by the following identity

$$\mathcal{M}[F^a(Z); Y^L, Y^U](t) = \mathcal{M}[Z; Z^L, Z^U](\rho(t)). \tag{19.32}$$

Different approximations are obtained depending on the choices for $\rho(\cdot)$ and $\mathcal{M}[\cdot]$.

There is a simple way to generate a valid measure $\mathcal{M}[\cdot]$. Let $\delta[f, g](t)$ denote a measure of distance between two functions of time $f(\cdot)$ and $g(\cdot)$ on the interval $[0, t]$ such that

(i) $\delta[f, g] \geq 0$,
(ii) $\delta[f', g](t) \geq \delta[f, g](t)$ whenever $f'(\cdot) \geq f(\cdot) \geq g(\cdot)$, and
(iii) the function of time $\delta[f, g](\cdot)$ is differentiable.

Each $\delta[f, g](\cdot)$ induces a valid measure $\mathcal{M}[\cdot]$ simply by forming the ratio

$$\mathcal{M}[B; B^L, B^U](t) := \frac{\delta[B, B^L](t)}{\delta(B^U, B^L)(t)}. \tag{19.33}$$

(Assume the denominator is never zero.) A common measure of distance for integrable $f(\cdot)$ and $g(\cdot)$ is the **L^p-norm**, $p \geq 1$, defined by

$$\delta[f, g](t) := \left(\int_0^t |f(\tau) - g(\tau)|^p \, w(\tau) \, d\tau \right)^{1/p}, \tag{19.34}$$

where $w(\cdot)$ can be interpreted as some appropriate weighting function of time. With this choice for $\delta[f, g]$, identity (19.32) becomes

$$\frac{\int_0^t [F(Z)(\tau) - Y^L(\tau)]^p \, w(\tau) d\tau}{\int_0^t [Y^U(\tau) - Y^L(\tau)]^p \, w(\tau) d\tau} = \frac{\int_0^{\rho(t)} [Z(\tau) - Z^L(\tau)]^p \, w(\tau) d\tau}{\int_0^{\rho(t)} [Z^U(\tau) - Z^L(\tau)]^p \, w(\tau) d\tau}. \tag{19.35}$$

(In our previous development, the parameter $p = 1$ and $w(\cdot) \equiv 1$.) The identity (19.35) can be expressed as

$$\int_0^t ([F^a(Z)](\tau) - Y^L(\tau))w(\tau)d\tau = \mathcal{R}_\rho(t) \int_0^{\rho(t)} [Z(\tau) - Z^L(\tau)]w(\tau)d\tau, \tag{19.36}$$

where

$$\mathcal{R}_\rho(t) := \frac{\int_0^t (Y^U(\tau) - Y^L(\tau))w(\tau)d\tau}{\int_0^{\rho(t)} [Z^U(\tau) - Z^L(\tau)]w(\tau)d\tau}. \tag{19.37}$$

Differentiating both sides of (19.36) with respect to t, we have

$$\begin{aligned}[F^a(Z)(t) - Y^L(t)]^p \;\; w(t) &= \mathcal{R}_\rho(t)\rho^{'}(t)[Z(\rho(t)) - Z^L(\rho(t))]^p \; w(\rho(t)) \\ &+ \;\; \mathcal{R}_\rho^{'}(t) \int_0^{\rho(t)} [Z(\tau) - Z^L(\tau)]^p \; w(\tau)d\tau. \end{aligned} \tag{19.38}$$

Once again, to *uniquely* determine $\rho(\cdot)$ and $[F^a(Z)](\cdot)$, we assume the cumulative output up to time t is a function of the *quantity* of cumulative input up to time $\rho(t)$ but does not depend on the structure of the input *before* time $\rho(t)$. As before, for this assumption to hold the integral on the right-hand-side of (19.38) must vanish, which implies the derivative $\mathcal{R}_\rho^{'}(t)$ must always be zero. Thus, $\mathcal{R}_\rho(t)$ in (19.37) is a constant, say equal to R, and

$$\int_0^t [Y^U(\tau) - Y^L(\tau)]^p \; w(\tau)d\tau = R \int_0^{\rho(t)} [Z^U(\tau) - Z^L(\tau)]^p \; w(\tau)d\tau. \tag{19.39}$$

Since $\rho(T^{input}) = T^{output}$, it follows that

$$R = \frac{\int_0^{T^{output}} [Y^U(\tau) - Y^L(\tau)]^p \; w(\tau)d\tau}{\int_0^{T^{input}} [Z^U(\tau) - Z^L(\tau)]^p \; w(\tau)d\tau}, \tag{19.40}$$

from which we conclude that $\rho(\cdot)$ is implicitly defined via the identity

$$\frac{\int_0^t [Y^U(\tau) - Y^L(\tau)]^p \; w(\tau)d\tau}{\int_0^{T^{output}} [Y^U(\tau) - Y^L(\tau)]^p \; w(\tau)d\tau} = \frac{\int_0^{\rho(t)} [Z^U(\tau) - Z^L(\tau)]^p \; w(\tau)d\tau}{\int_0^{T^{input}} [Z^U(\tau) - Z^L(\tau)]^p \; w(\tau)d\tau} \tag{19.41}$$

for all $0 \le t \le T^{output}$.

Using the fact that $\mathcal{R}_\rho^{'}(t) = 0$ and $\mathcal{R}_\rho(\cdot) = R$ in (19.38), for each $Z(\cdot) \in \mathsf{D}$ it follows that

$$[F^a(Z)(t) - Y^L(t)]^p \; w(t) = R\rho^{'}(t)[Z(\rho(t)) - Z^L(\rho(t))]^p \; w(\rho(t)). \tag{19.42}$$

Since (19.42) holds for $Z(\cdot) = Z^U(\cdot)$, it is also true that

$$[Y^U(t) - Y^L(t)]^p \; w(t) = R\rho^{'}(t)[Z(\rho(t)) - Z^L(\rho(t))]^p \; w(\rho(t)). \tag{19.43}$$

Dividing each side of (19.42) by the corresponding sides of (19.43) yields the identity

$$\frac{[F^a(Z)](t) - Y^L(t)}{Y^U(t) - Y^L(t)} = \frac{Z(\rho(t)) - Z^L(\rho(t))}{Z^U(\rho(t)) - Z^L(\rho(t))} \ . \tag{19.44}$$

The approximate dynamic production function uniquely determined from (19.44) is

$$[F^a(Z)](t) = Y^L(t) + \left(\frac{Y^U(t) - Y^L(t)}{Z^U(\rho(t)) - Z^L(\rho(t))} \right) \ [Z(\rho(t)) - Z^L(\rho(t))], \tag{19.45}$$

where $\rho(\cdot)$ is (implicitly) defined by (19.41).

Remark 19.9. The parameter p and weighting function $w(\cdot)$ do *not* appear in (19.44), which is *identical* to (19.30); however, $F^a[\cdot]$ is very much dependent on them, as they determine the function $\rho(\cdot)$.

Once again, if $Z(\cdot)$ is piecewise linear, then $F^a(Z)(\cdot)$ will be a linear function of the underling $z(\cdot)$ variables, thus facilitating computational analyses.

19.3 Application to Project-Oriented Production Systems

19.3.1 Description

In a **project-oriented production system**, several large concurrent projects are simultaneously carried out subject to inflexible capacities for resources such as skilled labor and equipment. In a naval shipyard, for example, as many as ten ships may in overhaul at the same time, and each ship undergoes thousands of activities over a period of up to many months. Effective management is essential since labor costs can be staggering—potentially thousands of workers with specialized skills are employed. In such organizations, highest levels of management do not schedule individual projects; rather, they are responsible for securing new business, negotiating prices and due dates, and planning project milestones. In a naval shipyard, examples of important milestones are when to dock and undock the ships, power-up nuclear systems, light-off boilers, or push steam through the turbines. The planning of project milestones is no easy task. On the one hand, milestone dates must not overload shop labor; otherwise, customer commitments (due dates) are not met. On the other hand, milestone dates should avoid under-utilizing shop labor; otherwise, productivity is reduced, risking budget overruns. The process of setting milestone dates proceeds iteratively. An initial set of milestone dates is proposed, an analysis of which milestones cannot be met and which resources are under-utilized or exceed capacity is undertaken, and new milestones are then proposed. This process is also used to suggest where to

economically expand capacity to increase productivity. Depending on problem size, resource-constrained scheduling algorithms/software can be used for management of project-oriented production systems. In this section, we show how to represent project execution via a continuous-time model that can be approximated via a set of linear equalities. This data structure can be manipulated easily and quickly by methods of linear programming to perform the required analyses.

In industrial project networks, typically there are many groups of similar *activities in parallel*—collections of activities that represent the same kind of work and can be scheduled at the same time. Parallel activities frequently use the same type of resources. In the shipyard context, there are groups of rip-out, repair, and re-installation activities. There is **strict precedence** among these activities: no re-installation work can begin before *all* of the repair work has been completed, and no repair work can begin until all rip-out has been completed. For data reduction purposes, groups of detailed rip-out, repair, and re-installation activities can be combined into corresponding rip-out, repair and re-installation aggregate activities. At the aggregate level, production at the rip-out, repair and re-installation aggregate activities can (and typically do) *overlap* in time. Therefore, it is inaccurate at the aggregate level to maintain strict precedence between these aggregate activities. At the aggregate level, it is necessary to develop a general work flow model that permits, but reasonably constrains, simultaneous resource use by consecutive activities.

Conceptually, an aggregate activity produces intermediate output used by a follow-on aggregate activity. For example, the repair aggregate produces repaired equipment needed as input by the re-installation aggregate. Naturally, the rate at which the re-installation aggregate can use repaired equipment depends on the rate of supply of repaired equipment produced by the repair aggregate. From this perspective, the rates of resource consumption at successive aggregate activities are linked by a material balance constraint. The output of the repair aggregate is represented, not surprisingly, by a dynamic production function using a two-point boundary approximation.

19.3.2 Detailed Activities

For each detailed activity l, let ES_l denote its earliest start-time, LS_l denote its latest start-time, d_l denote its fixed duration, and a_l^k denote the total amount of resource k it consumes. (The earliest and latest start-times are computed given a set of key milestone dates.) At the detailed level, it is assumed that an activity consumes its resources at a constant rate between its start-time S_l and finish-time $S_l + d_l$. Consequently, the input vector for activity l is

$$x_l(\tau) = (a_l^1, a_l^2, \ldots, a_l^n) z_l(\tau), \tag{19.46}$$

where the index function is of the form

$$z_l(\tau) = (1/d_l) \cdot 1_{[S_l, S_l+d_l]}(\tau). \tag{19.47}$$

The index $z_l(\tau)$ is called the *operating intensity* for activity l. The cumulative intensity $Z_l(t) = \int_0^t z_l(\tau)d\tau$ measures the fraction of the resources required by activity l that has been consumed by time t. By construction, $Z(t)$ eventually reaches its maximum value of one.

With respect to the dynamic production function, each activity l produces a *unique* product labeled (l, m) for each activity that immediately follows it. Let

$$y_l^m(\tau) = [f_l^{(l,m)}(z_l)](\tau)$$

denote the output of 'product' (l, m) produced by activity l at time τ. Since activity m uses only one unit of product (l, m) at rate $z_m(\cdot)$, a material balance constraint

$$\int_0^t y_i^m(\tau)d\tau \;\geq\; \int_0^t z_m(\tau)d\tau \tag{19.48}$$

should ensure that the start-time S_m of activity m cannot be earlier than the finish-time of activity l; that is,

$$S_l + d_l \leq S_m. \tag{19.49}$$

One obvious choice for the dynamic production function is

$$y_l^m(\tau) = [f_l^{(l,m)}(z_l)](\tau) = \begin{cases} 1, & \text{if } \tau = S_l + d_l, \\ 0, & \text{otherwise.} \end{cases} \tag{19.50}$$

Substituting (19.50) and (19.47) into (19.48) yields constraint (19.49). For this choice of dynamic production function, the outputs produced by activity l are *event-based*.

Instead of using the obvious event-based dynamic production function, the following *rate-based* dynamic production function will prove a superior choice:

$$y_l^m(\tau) = [f_l^{(l,m)}(z_l)](\tau) = (1/d_m) \cdot 1_{[S_l+d_l, S_l+d_l+d_m]}(\tau). \tag{19.51}$$

Substituting (19.51) and (19.47) into (19.48) also yields constraint (19.49). With this choice of dynamic production function, however, the cumulative output curve is piecewise linear and represents the *earliest* operating intensity of activity m *consistent* with the finish-time of activity l.

19.3.3 Aggregate Activities

At the aggregate level, each activity A_i represents an aggregation of a number of parallel detailed activities l. Each detailed activity l is assigned to exactly one aggregate, indicated by $l \in A_i$. Activity A_i produces an intermediate product for use by activity A_j if there exists an $l \in A_i$ and $m \in A_j$ such

that detailed activity l is an immediate predecessor of activity m. Aggregate activities are formed only if the detailed activities within the aggregate use the same *mix* of resources. That is, the ratios

$$\frac{a_l^1}{\sum_{l\in A_i} a_l^1},\ \frac{a_l^1}{\sum_{l\in A_i} a_l^1},\ \cdots,\ \frac{a_l^n}{\sum_{l\in A_i} a_l^n} := \alpha_l^i$$

are independent of resource $k = 1, 2, \ldots, n$. For example, suppose there are three detailed activities within aggregate A_i. If $\alpha_1^i = 0.20$, $\alpha_2^i = 0.30$, and $\alpha_3^i = 0.50$, then detailed activities 1, 2, and 3 consume, respectively, 20%, 30%, and 50% of the total amount of *each* resource used by all three activities. By construction, $\sum_{l\in A_i} \alpha_l^i = 1$. For many industrial project networks, this requirement is not restrictive, as there are many parallel activities using identical or near-identical mixes of resources.

This resource-use requirement implies that the input of resource k by aggregate activity A_i is

$$\begin{aligned} x_i^k(\tau) &= \sum_{l\in A_i} x_l^k \\ &= \sum_{l\in A_i} a_l^k z_l(\tau) \\ &= \left(\sum_{l\in A_i} a_l^k\right)\left[\sum_{l\in A_i} \frac{a_l^k}{\sum_{l\in A_i} a_l^k} z_l(\tau)\right] \\ &= \left(\sum_{l\in A_i} a_l^k\right)\sum_{l\in A_i} \alpha_l^i z_l(\tau) \\ &:= a_i^k z_i(\tau). \end{aligned} \tag{19.52}$$

In (19.52), a_i^k represents the total amount of resource k used by all detailed activities within A_i, and $z_i(\cdot)$ is the *aggregate operating intensity* of A_i. Since $z_i(\cdot)$ is a convex combination of detailed operating intensities, it follows that $Z_i(t)$ eventually reaches one, too. Moreover, $Z_i(t)$ represents the fraction of the total resources required to complete all detailed activities within A_i consumed by time t.

For each detailed activity l, let

$$\begin{aligned} z_l^L(\cdot) &:= (1/d_l)\cdot 1_{[LS_l, LS_l+d_l]}(\cdot), \\ z_l^E(\cdot) &:= (1/d_l)\cdot 1_{[ES_l, ES_l+d_l]}(\cdot) \end{aligned}$$

denote, respectively, the operating intensities associated with the late- and early-start schedules, and let

$$\begin{aligned} z_i^L(\cdot) &:= \sum_{l\in A_i} \alpha_l^i z_l^L(\cdot), \\ z_i^E(\cdot) &:= \sum_{l\in A_i} \alpha_l^i z_l^E(\cdot) \end{aligned}$$

denote, respectively, the aggregate operating intensities of A_i associated with the late- and early-start schedules of its detailed activities. For each feasible schedule of detailed activities,

$$Z_i^L(t) \leq Z_i(t) \leq Z_i^E(t). \tag{19.53}$$

Thus, the $Z_i^L(\cdot)$ and $Z_i^E(\cdot)$ define *boundary curves* for the feasible domain of the cumulative aggregate operating intensity $Z_i(\cdot)$.

19.3.4 Aggregate Dynamic Production Function

Let A_j be an aggregate activity that immediately follows A_i. At the aggregate level, the material balance constraint is of the form

$$[F_i^{(i,j)}(Z_i)](t) \geq Z_j(t). \tag{19.54}$$

The output of 'product' (i, j) of A_i is defined to be the *earliest* aggregate operating intensity for A_j consistent with the start-times for the detailed activities within A_i, formally,

$$[F_i^{(i,j)}(Z_i)](\cdot) := \sum_{m \in A_j} \alpha_m^j [F_l^{(l,m)}(z_l)](\cdot). \tag{19.55}$$

Substituting (19.55) into (19.54) yields a necessary constraint: The cumulative aggregate operating intensity of A_j cannot be "earlier" than the cumulative aggregate operating intensity for A_j obtained by setting each of the detailed operating intensities in A_j to their earliest start-times consistent with the finish-times of their predecessor detailed activities in A_i. It represents the ideal production function for this setting. Note that

$$[F_i^{(i,j)}(Z_i^L)](\cdot) = Z_j^L(\cdot) \text{ and } [F_i^{(i,j)}(Z_i^E)](\cdot) = Z_j^E(\cdot); \tag{19.56}$$

that is, this production function maps the boundary curves of A_i onto the boundary curves of A_j.

There is a fundamental problem with the definition (19.55) of the aggregate production function: it incorporates knowledge of the schedules (i.e. start-times) for the detailed activities within A_i. A model for the aggregate production function must be *independent* of such knowledge, so as to not defeat the point of aggregation. At this point, it is necessary to approximate the ideal production function. A reasonable starting point is to represent the feasible domain D_i for A_i as the collection of all nondecreasing, nonnegative $Z_i(\cdot)$ that satisfy (19.53). In light of (19.56), a two-point boundary approximation may can be used.

Remark 19.10. In application, the decision variables are the $Z_i(\cdot)$. These functions are constrained to be nondecreasing and piecewise linear, and the material balance constraints are applied at discrete points in time. Consequently, the material balance constraints are *linear* in the decision variables.

19.4 Aggregation of Dynamic Production Functions

19.4.1 Serial Aggregation

Dynamic production functions can be obtained via *composition* of detailed dynamic production functions when modeling output flow of a network of activities in series. We briefly outline this type of aggregation.

In manufacturing systems, it is not uncommon for activities to be arranged in series, commonly referred to as a **flow line**. In a flow line, output of each activity in the series (except for the last one) is used immediately as input by its successor activity, etc. There are no buffers of inventory between the successive activities.[1] Consider N serial activities. The input domain of each activity is index-based, and the unique, cumulative output curve of activity $i = 1, 2, \ldots, N$ is represented by a dynamic production function $[F_i(Z_i)](\cdot)$. Starting from the input curve $Z_1(\cdot)$ to the first activity, the flow of output across the flow line can be conceptually represented as

$$\begin{aligned} Z_1 \longrightarrow F_1(Z_1) \longrightarrow Z_2 = F_2\big(F_1(Z_1)\big) \longrightarrow Z_3 = F_3\Big(F_2\big(F_1(Z_1)\big)\Big) \\ \longrightarrow \quad \cdots \quad \longrightarrow F_N\left(F_{N-1}\Big(F_{N-2}\big(\cdots\big)\Big)\right). \end{aligned}$$

Letting

$$\mathsf{F} := F_N \circ F_{N-1} \circ \cdots \circ F_1$$

denote the compositions of the functions $F_1, F_2, \ldots, F_N$, the dynamic production function for the flow line is

$$[\mathsf{F}(Z_1)](\cdot) := [(F_N \circ F_{N-1} \circ \cdots \circ F_1)(Z_1)](\cdot).$$

Example 19.11. In a simple setting, suppose the input of resource $k = 1, 2, \ldots, n$ and output of each activity i at time τ are, respectively,

$$\begin{aligned} x_i^k(\tau) &= a_i^k z_i(\tau), \\ y_i(\tau) &= [f_i(z_i)](\cdot) = z_i(\tau - \ell_i). \end{aligned}$$

Let $x^k(\tau)$ denote the input of resource k consumed by the flow line at time τ. Then,

$$\begin{aligned} [\mathsf{F}(Z_1)](\tau) &= Z_1\Big(\tau - \sum_{j=1}^{N} \ell_j\Big), \\ x^k(\tau) &= \sum_{i=1}^{N} a_i^k z_1\Big(\tau - \sum_{j=1}^{i} \ell_j\Big). \end{aligned}$$

[1] In some flow lines, buffers of inventory between activities do exist; if so, the description below does not apply.

This type of serial aggregation has been successfully applied in the semiconductor industry. See Leachman et. al. [1996] and Leachman [2002] for a detailed description.

19.4.2 Parallel Aggregation

It is also possible to aggregate activities in parallel, as follows. We adopt the setup of the previous subsection. Let $\lambda_i(\cdot)$, $i = 1, 2, \ldots, N$, be weighting functions such that for each i

(i) $\lambda_i(\cdot) \geq 0$ for all $t \geq 0$, and
(ii) $\sum_{i=1}^{N} \lambda_i(t) = 1$ for all $t \geq 0$.

An example of a dynamic production function obtained via parallel aggregation is

$$[\mathsf{F}(Z_1, Z_2, \ldots, Z_N)](\cdot) := \sum_{i=1}^{N} \lambda_i(t)[F_i(Z_i)](\cdot).$$

Here, the aggregate output is a *time-varying, convex combination* of the outputs of the detailed activities.

19.5 Estimation via Dynamic Activity Analysis

Application of activity analysis yields yet another, practical way to estimate an output curve via a linear production functional. Let $x_1(\cdot), x_2(\cdot), \ldots, x_N(\cdot)$ denote a representative sample of input functions, and let $y_1(\cdot), y_2(\cdot), \ldots, y_N(\cdot)$ denote the corresponding observed output functions.

19.5.1 Basic Model

If the input function $x(\cdot)$ is restricted to be a linear combination of the $x_i(\cdot)$, then it must be of the form

$$x(\cdot) = \sum_{i=1}^{N} \lambda_i x_i(\cdot), \ \lambda_i \geq 0, \ i = 1, 2, \ldots, N. \tag{19.57}$$

If the dynamic production function is linear on this domain, it follows that

$$y(\cdot) = f\left[\sum_{i=1}^{N} \lambda_i x_i\right](\cdot) = \sum_{i=1}^{N} \lambda_i f[x_i](\cdot) = \sum_{i=1}^{N} \lambda_i y_i(\cdot). \tag{19.58}$$

In this approximation, a great deal of information about the input-output process via $\{(x_i(\cdot), y_i(\cdot))\}$ is used to approximate the dynamic production function. For example, this approach permits a very detailed simulation-based approach to arrive at each $y_i(\cdot)$. The cost is that the possible shapes of the

input function are restricted to be linear combinations of a representative sample of input functions. Arguably, this may not be too egregious.

The input-output model characterized by (19.57) and (19.58) exhibits constant returns-to-scale. Variable returns can be modeled by adding the condition that $\sum_i \lambda_i = 1$.

19.5.2 Extensions

The first extension to the basic dynamic activity analysis models described above is to let the vector λ be a function of time, namely,

$$\lambda(\cdot) = (\lambda_1(\cdot), \lambda_2(\cdot), \ldots, \lambda_n(\cdot)) \geq 0,$$

in which case, for each $t \geq 0$,

$$x(t) = \sum_{i=1}^{N} \lambda_i(t) x_i(t), \quad y(t) = \sum_{i=1}^{N} \lambda_i(t) y_i(t) \tag{19.59}$$

in the constant returns-to-scale model. The variable returns-to-scale model adds the condition $\sum_i \lambda_i(t) = 1$ for all $t \geq 0$.

The second extension recognizes that inputs x_i and outputs y_i in (19.59) can each be a vector of functions to model a multi-input, multi-output system, as in

$$x_i(\cdot) = (x_i^1(\cdot), x_i^2(\cdot), \ldots, x_i^n(\cdot)), \quad y_i(\cdot) = (y_i^1(\cdot), y_i^2(\cdot), \ldots, y_i^m(\cdot)). \tag{19.60}$$

A discrete-time approximation yields a sequence of independent CRS-DEA or VRS-DEA models described in Chapter 4. For example, consider a standard time grid $\mathcal{G}$. For each $t = 1, 2, \ldots$, let (x_i^t, y_i^t) denote the i^{th} pair of input and output vectors in period t, and define

$$x^t(\lambda^t) := \sum_{i=1}^{N} \lambda_i^t x_i^t, \quad y^t(\lambda^t) := \sum_{i=1}^{N} \lambda_i^t y_i^t. \tag{19.61}$$

Suppressing the time index t, these expressions are identical to (4.12) on p. 63. The technology is then characterized by the collection

$$\left\{ (x^1(\lambda^1), y^1(\lambda^1)),\ (x^2(\lambda^2), y^2(\lambda^2)), \ldots, (x^t(\lambda^t), y^t(\lambda^t)), \ldots \right\}.$$

No restrictions on the λ^t (other than non-negativity) yield a constant returns-to-scale model. A variable returns-to-scale model adds the conditions $\sum_i \lambda_i^t = 1$ for each t.

19.6 Exercises

19.1. Assume $q_0 := q_0^0$ and the y_t are given. Derive algebraic expressions for the q_t^k and the y_t^k displayed in Tables 19.1 and 19.2. Show the computations using your formulae for $k = 1$ and $k = 2$ (i.e., the first two periods).

19.2. Consider the load-dependent, lead time model in which $\pi(q(t), t) = e^{-0.02q(t)}$, $q(0) = 40$, and $x = (x_1, x_2, x_3, x_4) = (20, 30, 25, 10)$. (A standard time grid is used.) Use the approximation discussed in Section 19.1.2 to answer the following questions.

(a) Determine the q_t and y_t for the first four periods $t = 1, 2, 3, 4$.
(b) Assuming the FIFO discipline, replicate Tables 19.1, 19.2, and 19.3 for this example; that is, determine the q_t^k and y_t^k and show how to express the y_t as linear functions of the x_t.
(c) Assuming the simultaneous withdrawal, replicate Tables 19.4, 19.5, and 19.6 for this example; that is, determine the q_t^k and y_t^k and show how to express the y_t as linear functions of the x_t.

19.3. Consider the following boundary curves:

$$Z^U(t) = \begin{cases} 5t, & 0 \le t \le 2, \\ 2t + 6, & 2 \le t \le 5. \end{cases}$$

$$Z^L(t) = \begin{cases} t, & 0 \le t \le 4, \\ 12t - 44, & 4 \le t \le 5. \end{cases}$$

$$Y^U(t) = \begin{cases} 0, & 0 \le t \le 3, \\ 4t - 8, & 2 \le t \le 5, \\ t + 7, & 5 \le t \le 9. \end{cases}$$

$$Y^L(t) = \begin{cases} 0, & 0 \le t \le 2, \\ 2t - 4, & 2 \le t \le 7, \\ 3t - 11, & 7 \le t \le 9. \end{cases}$$

For this example, $T^{in} = 5$ and $T^{out} = 9$. A standard time grid will be used.

(a) Determine $\mathcal{A}_1(\cdot)$ and $\mathcal{A}_2(\cdot)$.
(b) Determine $\rho(t)$ for $t = 3, 4, 5, 6, 7, 8$.
(c) Express $Y(t) = [F^a(Z)](t)$ defined in (19.31) as a linear function of the z_t variables for $t = 3, 4, 5, 6, 7, 8$.

19.7 Bibliographical Notes

The idea for the relative area ratio in the context of shipyard planning originates with Boysen's [1982] dissertation and refined in Leachman and Boysen [1985]. Consult the latter paper for detailed explanations of the shipyard planning problem and the development of a linear programming model to solve it. The formal development of the two-point boundary approximation refines and extends the presentation found in Hackman and Leachman [1989].

Descriptions of and extensions to the dynamic activity analysis models presented in Section 19.5.2 can be found in Fare et. al. [1996]. De Mateo et. al. [2006] embed the model of Section 19.5.2 into a dynamic DEA model that explicitly accounts for the cost of adjustment on capacities over time, budget constraints, etc.

Riano [2002] describes an iterative scheme to obtain a load-dependent lead time distribution that contains explicit approximations to the underlying queue process. See also Riano et. al. [2007].

19.8 Solutions to Exercises

19.1 There are a number of ways to express the respective queues and outputs. To simplify the subsequent notation, define cumulative input and output values, respectively, as $X_t := \sum_{k=1}^{t} x_k$ and $Y_t := \sum_{k=1}^{t} y_k$, $t = 1, 2, \ldots$. For the respective queue values, we define:

$$q_t^k := \max\Big\{ \min\Big(q_0 + X_k - Y_t,\ x_k\Big),\ 0\Big\},\ k \le t,\ t = 1, 2, \ldots \ .$$

The calculations for $k = 1$ and $k = 2$ are:

$$\begin{aligned}
q_1^1 &= \max\{\min(125 - 34.86, 25),\ 0\} = 25,\\
q_2^1 &= \max\{\min(125 - 70.28, 25),\ 0\} = 25,\\
q_3^1 &= \max\{\min(125 - 106.35, 25),\ 0\} = 18.65,\\
q_4^1 &= \max\{\min(125 - 143.22, 25),\ 0\} = 0,\\
q_5^1 &= \max\{\min(125 - 181.23, 25),\ 0\} = 0,\\
q_6^1 &= \max\{\min(125 - 219.37, 25),\ 0\} = 0,\\
q_1^2 &= 0,\\
q_2^2 &= \max\{\min(155 - 70.28, 30),\ 0\} = 30,\\
q_3^2 &= \max\{\min(155 - 106.35, 30),\ 0\} = 30,\\
q_4^2 &= \max\{\min(155 - 143.22, 30),\ 0\} = 11.78,\\
q_5^2 &= \max\{\min(155 - 181.23, 30),\ 0\} = 0,\\
q_6^2 &= \max\{\min(155 - 219.37, 30),\ 0\} = 0.
\end{aligned}$$

For the respective output values, we define:

$$y_t^k := \min\Big\{ \max\Big(y_t - \sum_{j=0}^{k-1} q_{t-1}^j,\ 0\Big),\ q_{t-1}^k\Big\},\ k < t,\ t = 1, 2, \ldots$$

$$y_t^t := \min\Big\{ \max\Big(y_t - \sum_{j=0}^{k-1} q_{t-1}^j,\ 0\Big),\ x_t\Big\},\ t = 1, 2, \ldots \ .$$

The calculations for $k = 1$ and $k = 2$ are:

$$\begin{aligned}
y_1^1 &= \min\{\max(34.86 - 100, 0), 25\} = 0,\\
y_2^1 &= \min\{\max(35.42 - 65.14, 0), 25\} = 0,\\
y_3^1 &= \min\{\max(36.07 - 29.72, 0), 25\} = 6.35,\\
y_4^1 &= \min\{\max(36.87 - 0, 0), 18.65\} = 18.65,\\
y_5^1 &= \min\{\max(38.01 - 0, 0), 0\} = 0,\\
y_6^1 &= \min\{\max(38.14 - 0, 0), 0\} = 0,
\end{aligned}$$

$$
\begin{aligned}
y_1^2 &= 0,\\
y_2^2 &= \min\{\max(35.42 - 90.14, 0), 30\} = 0,\\
y_3^2 &= \min\{\max(36.07 - 54.72, 0), 30\} = 0,\\
y_4^2 &= \min\{\max(36.87 - 18.65, 0), 30\} = 18.22,\\
y_5^2 &= \min\{\max(38.01 - 0, 0), 11.78\} = 11.78,\\
y_6^2 &= \min\{\max(38.14 - 0, 0), 0\} = 0.
\end{aligned}
$$

19.2 (a) Using (19.4) and (19.5), we have:

$$
\begin{aligned}
\pi_0 &= \pi(q(0), 0) = e^{-0.02(40)} = 0.4493,\\
q_1 &= e^{-\pi_0} q_0 + x_1 \frac{1 - e^{-\pi_0}}{\pi_0} = (0.6381)(40) + 20\left[\frac{1 - 0.6381}{0.4493}\right] = 41.63,\\
\pi_1 &= e^{-0.02(41.63)} = 0.4349,\\
q_2 &= e^{-\pi_1} q_1 + x_2 \frac{1 - e^{-\pi_1}}{\pi_1} = (0.6473)(41.63) + 30\left[\frac{1 - 0.6473}{0.4349}\right] = 51.27,\\
\pi_2 &= e^{-0.02(51.27)} = 0.3586,\\
q_3 &= e^{-\pi_2} q_2 + x_3 \frac{1 - e^{-\pi_2}}{\pi_2} = (0.6986)(51.27) + 25\left[\frac{1 - 0.6986}{0.3586}\right] = 56.83,\\
\pi_3 &= e^{-0.02(56.83)} = 0.3209,\\
q_4 &= e^{-\pi_3} q_3 + x_4 \frac{1 - e^{-\pi_3}}{\pi_3} = (0.7255)(56.83) + 10\left[\frac{1 - 0.7255}{0.3209}\right] = 49.78.
\end{aligned}
$$

Using the inventory balance equation

$$y_t = q_{t-1} + x_t - q_t,\ t = 1, 2, \ldots,$$

we obtain the y_t as $y_1 = 18.37$, $y_2 = 20.36$, $y_3 = 19.44$, and $y_4 = 17.05$.

(b) The tables are shown below. Accordingly, the expressions for the y_t in terms of the x_t are:

$$
\begin{aligned}
y_1 &= 0.459 q_0,\\
y_2 &= 0.509 q_0,\\
y_3 &= 0.032 q_0 + 0.909 x_1,\\
y_4 &= 0.091 x_1 + 0.507 x_2.
\end{aligned}
$$

(c) The α and β vectors for this problem are:

$$
\begin{aligned}
\alpha &= (\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (0.6381, 0.6473, 0.6986, 0.7255),\\
\beta &= (\beta_1, \beta_2, \beta_3, \beta_4) = (0.1945, 0.189, 0.1595, 0.1446).
\end{aligned}
$$

Table 19.7. Queue matrix q_t^k for Exercise 19.2(b).

	Period				
Queue due to:	0	1	2	3	4
Initial inventory		21.63	1.27	0	0
Starts in period 1		20.00	20.00	1.83	0
Starts in period 2		0	30.00	30.00	14.78
Starts in period 3		0	0	25.00	25.00
Starts in period 4		0	0	0	10.00
Queue at end of period:	40	41.63	51.27	56.83	48.78

Table 19.8. Output matrix y_t^k for Exercise 19.2(b).

	Period			
Output due to:	1	2	3	4
Initial inventory	18.37	20.36	1.27	0
Starts in period 1	0	0	18.17	1.83
Starts in period 2	0	0	0	15.22
Starts in period 3	0	0	0	0
Starts in period 4	0	0	0	0
Output at end of period:	18.37	20.36	19.44	17.05
Cumulative output at end of period:	18.37	38.73	58.17	75.22

Table 19.9. Percentage of starts in period k emerging as output in period t for Exercise 19.2(b).

	Period			
	1	2	3	4
Initial inventory	0.459	0.509	0.032	0
Starts in period 1	0	0	0.909	0.091
Starts in period 2	0	0	0	0.507
Starts in period 3	0	0	0	0
Starts in period 4	0	0	0	0

The tables are shown below. Accordingly, the expressions for the y_t in terms of the x_t are:

$$
\begin{aligned}
y_1 &= 0.362q_0 + 0.195x_1, \\
y_2 &= 0.225q_0 + 0.284x_1 + 0.189x_2, \\
y_3 &= 0.125q_0 + 0.157x_1 + 0.244x_2 + 0.160x_3, \\
y_4 &= 0.079q_0 + 0.100x_1 + 0.155x_2 + 0.231x_3 + 0.145x_4.
\end{aligned}
$$

Table 19.10. Queue matrix q_t^k for Exercise 19.2(c).

	Period				
Queue due to:	0	1	2	3	4
Initial inventory		25.52	16.52	11.54	8.37
Starts in period 1		16.11	10.43	7.29	5.29
Starts in period 2		0	24.32	16.99	12.33
Starts in period 3		0	0	21.01	15.24
Starts in period 4		0	0	0	8.55
Queue at end of period:	40	41.63	51.27	56.83	48.78

Table 19.11. Output matrix y_t^k for Exercise 19.2(c).

	Period			
Output due to:	1	2	3	4
Initial inventory	14.48	9.00	4.98	3.17
Starts in period 1	3.89	5.68	3.14	2.00
Starts in period 2	0	5.68	7.33	4.66
Starts in period 3	0	0	3.99	5.77
Starts in period 4	0	0	0	1.45
Output at end of period:	18.37	20.36	19.44	17.05
Cumulative output at end of period:	18.37	38.73	58.17	75.22

Table 19.12. Percentage of starts in period k emerging as output in period t for Exercise 19.2(c).

	Period			
	1	2	3	4
Initial inventory	0.362	0.225	0.125	0.079
Starts in period 1	0.195	0.284	0.157	0.100
Starts in period 2	0	0.189	0.244	0.155
Starts in period 3	0	0	0.160	0.231
Starts in period 4	0	0	0	0.145

19.3 (a) The two curves $\mathcal{A}_1(\cdot)$ and $\mathcal{A}_2(\cdot)$ are:

$$\mathcal{A}_1(t) = \begin{cases} 2t^2, & 0 \le t \le 2, \\ 0.5t^2 + 6t - 6, & 2 \le t \le 4, \\ -5t^2 + 50t - 94, & 4 \le t \le 5. \end{cases}$$

$$\mathcal{A}_2(t) = \begin{cases} 0, & 0 \le t \le 2, \\ t^2 - 4t + 4, & 2 \le t \le 5, \\ -0.5t^2 - 11t - 33.5, & 5 \le t \le 7, \\ -t^2 + 18t - 58, & 7 \le t \le 9. \end{cases}$$

The critical values for the respective areas are $\mathcal{A}_1(2) = 8$, $\mathcal{A}_1(4) = 26$, $\mathcal{A}_1(5) = 31$, and $\mathcal{A}_2(5) = 9$, $\mathcal{A}_2(7) = 19$, $\mathcal{A}_2(9) = 23$.

(b)

$$\begin{aligned}
\mathcal{A}_2(3) &= 1 \Rightarrow \mathcal{A}_1(\rho(3)) = (1/23)31 = 1.3478 \\
&\Rightarrow 2(\rho(3))^2 = 1.3478 \Rightarrow \rho(3) = 0.8209. \\
\mathcal{A}_2(4) &= 4 \Rightarrow \mathcal{A}_1(\rho(4)) = (4/23)31 = 5.3913 \\
&\Rightarrow 2(\rho(4))^2 = 5.3913 \Rightarrow \rho(4) = 1.6418. \\
\mathcal{A}_2(5) &= 9 \Rightarrow \mathcal{A}_1(\rho(5)) = (9/23)31 = 12.1304 \\
&\Rightarrow 0.5(\rho(5))^2 + 6(\rho(5)) - 6 = 12.1304 \\
&\Rightarrow \rho(5) = -6 + \sqrt{36 + 4(18.1304)(0.5)} \Rightarrow \rho(5) = 2.501. \\
\mathcal{A}_2(6) &= 14.5 \Rightarrow \mathcal{A}_1(\rho(6)) = (14.5/23)31 = 19.5435 \\
&\Rightarrow 0.5(\rho(6))^2 + 6(\rho(6)) - 6 = 19.5435 \\
&\Rightarrow \rho(6) = -6 + \sqrt{36 + 4(25.5435)(0.5)} \Rightarrow \rho(6) = 3.3320. \\
\mathcal{A}_2(7) &= 19 \Rightarrow \mathcal{A}_1(\rho(7)) = (19/23)31 = 25.6087 \\
&\Rightarrow 0.5(\rho(7))^2 + 6(\rho(7)) - 6 = 25.6087 \\
&\Rightarrow \rho(6) = -6 + \sqrt{36 + 4(25.6087)(0.5)} \Rightarrow \rho(6) = 3.9608. \\
\mathcal{A}_2(8) &= 22 \Rightarrow \mathcal{A}_1(\rho(8)) = (22/23)31 = 29.6522 \\
&\Rightarrow -5(\rho(8))^2 + 50(\rho(8)) - 94 = 29.6522 \\
&\Rightarrow \rho(8) = 50 - \sqrt{2500 - 4(123.6522)(5)} \Rightarrow \rho(8) = 4.4808.
\end{aligned}$$

(c) The cumulative outputs are:

$$\begin{aligned}
Y(3) &= 2 + \frac{4-2}{4.1045 - 0.8209}[Z(.8209) - 0.8209] \\
&= 0.6091Z(0.8209) + 1.5 \\
&= 0.5z_1 + 1.5. \\
Y(4) &= 4 + \frac{8-4}{8.209 - 1.6418}[Z(1.6418) - 1.6418] \\
&= 0.6091Z(1.6418) + 3 \\
&= 0.6091[z_1 + 0.6418z_2] + 3 \\
&= 0.6091z_1 + 0.3909z_2 + 3. \\
Y(5) &= 6 + \frac{12-6}{11.002 - 2.501}[Z(2.501) - 2.501] \\
&= 0.7058Z(2.501) + 4.2348 \\
&= 0.7058[z_1 + z_2 + 0.501z_3] + 4.2348 \\
&= 0.7058z_1 + 0.7058z_2 + 0.3536z_3 + 4.2348. \\
Y(6) &= 8 + \frac{13-8}{12.664 - 3.332}[Z(3.332) - 3.332] \\
&= 0.5358Z(3.332) + 6.2147
\end{aligned}$$

$$
\begin{aligned}
&= 0.5358[z_1 + z_2 + z_3 + 0.332z_4] + 6.2147 \\
&= 0.5358z_1 + 0.5358z_2 + 0.5358z_3 + 0.1779z_4 + 6.2147. \\
Y(7) &= 10 + \frac{14 - 10}{13.9216 - 3.9608}[Z(3.9608) - 3.9608] \\
&= 0.4016Z(3.9608) + 8.4094 \\
&= 0.4016[z_1 + z_2 + z_3 + 0.9608z_4] + 8.4094 \\
&= 0.4016z_1 + 0.4016z_2 + 0.4016z_3 + 0.3859z_4 + 8.4094. \\
Y(8) &= 13 + \frac{15 - 13}{14.9616 - 9.7696}[Z(4.4808) - 9.7696] \\
&= 0.3852Z(4.4808) + 9.2368 \\
&= 0.3852[z_1 + z_2 + z_3 + z_4 + 0.4808z_5] + 9.2368 \\
&= 0.3852z_1 + 0.3852z_2 + 0.3852z_3 + 0.3852z_4 + 0.1852z_5 + 9.2368.
\end{aligned}
$$

20

A Stochastic Input-Output Model

In this chapter, we develop a stochastic input-output model in which the formula for $Y(t)$ in (18.3) is the *expected cumulative output* by time t. Using this model, we provide "confidence interval curves" to bound the expected cumulative output curve. Analogous development holds for the cumulative input $X(t)$. To simplify the notation, we assume there is no input prior to time 0. We use the convention that all expectations below are assumed to exist. Consult Appendix I for a review of basic definitions and properties from probability required for the modeling development that follows.

20.1 Input-Output Model with Single Inputs

Instead of viewing the input-output process as a deterministic process, we now view it as a stochastic process described as follows. Input events of the system are governed by a Poisson process $\{N(t),\ t \geq 0\}$ whose intensity function is denoted quite naturally by $z(t)$. That is,

$$Z(t) = E[N(t)] = \int_0^t z(\tau)d\tau. \tag{20.1}$$

We assume a single input eventually results in a single output. (The next section relaxes this assumption.) When input i enters the system at time T_i, its process or lead time is a random variable W_i with cumulative distribution

$$P(W_i \leq t) = W(t, T_i),$$

where $W(\cdot)$ is a cumulative lead time distribution as defined on p. 310. We further assume the W_i are independent and independent of the times T_i. We may view this process as a collection of pairs $\{(T_i,\ W_i)\}_{i=1}^{\infty}$ for which $T_1 < T_2 < \dots$.

Fix a point in time $t > 0$, and let $Y(t)$ denote the cumulative output by time t, which is now random. Under the present setup, an input i will be counted as output by time t if and only if $T_i + W_i \leq t$. Thus,

$$Y(t) = \sum_{i=1}^{N(t)} 1(T_i + W_i \le t), \tag{20.2}$$

where $1(A)$ denotes the indicator function of the event A, i.e., $1(A)$ equals one if A it true, 0 if A is false.

We will now derive an expression for the expected cumulative output $E[Y(t)]$ by time t. It will be useful to imagine a system observer who faithfully records on separate index cards, one for each input, the input time on one side and the process time on the opposite side, but not the event index. (The system observer keeps track of the event index.) Now suppose we are informed that $N(t) = n$. Our system observer hands us a sealed envelope containing an index card randomly chosen from the set of n index cards. What is the probability this input will have been completed by time t? Let $\tilde{T}$ and $\tilde{W}$ denote, respectively, this input's arrival and process times listed on the index card. Now suppose we are permitted to open the sealed envelope and observe the arrival time *but not the process time.* If we learn that $\tilde{T} = \tau$, then the conditional probability this input will have been completed by time t is $W(t - \tau, \tau)$. Since we are unable to open the sealed envelope, to determine the *total* probability of this input having been completed by time t, we need to weight this conditional probability by the probability the arrival time for this input is τ and sum over all τ. Therefore,

$$P(\tilde{T} + \tilde{W} \le t \mid N(t) = n) = \int_0^t P(\tilde{T} + \tilde{W} \le t \mid N(t) = n,\ \tilde{T} = \tau) g(\tau) d\tau, \tag{20.3}$$

where $g(\cdot)$ denotes the conditional probability density function of $\tilde{T}$ given $N(t) = n$. Using an important property of Poisson processes, $g(\cdot)$ is the derivative of the cumulative distribution function given in (I.3), p. 508, where $Z(\cdot)$ replaces $\Lambda(\cdot)$. Consequently,

$$g(\tau) = \frac{z(\tau)}{Z(t)}, \quad \tau \in [0, t]. \tag{20.4}$$

Note that the constant n does not appear on the right-hand side of (20.4).

Substituting (20.4) into (20.3), we conclude that

$$\begin{aligned} P(\tilde{T} + \tilde{W} \le t \mid N(t) = n) &= \int_0^t W(t - \tau, \tau) \frac{z(\tau)}{Z(t)}\, d\tau \\ &:= p_t. \end{aligned} \tag{20.5}$$

The probability p_t is constant, independent of n and which index card we were handed. Consequently, identifying the index cards with the n inputs, we see that each of the n inputs will have the *same* probability p_t of having been completed by time t. Identifying "success" with "output achieved by time t," and thinking of p_t as the probability of success, we have shown, *conditioned on*

$N(t) = n$, the random variable $Y(t)$ is *binomially distributed with parameters* n *and* p_t. That is,

$$P(Y(t) = k \mid N(t) = n) = \binom{n}{k} p_t^k (1 - p_t)^{n-k},\ 0 \le k \le n. \tag{20.6}$$

Since the mean of this binomial distribution is np_t, it follows that

$$E[Y(t) \mid N(t)] = N(t)p_t, \tag{20.7}$$

which is the expected cumulative output by time t conditioned on $N(t)$.

Conditioning on $N(t)$, and using (I.6) and (20.7), the expected cumulative output by time t is

$$E[Y(t)] = E(E[Y(t) \mid N(t)]) = E(N(t)p_t).$$

Therefore, by (20.1), (20.5), and the fact that p_t is a constant,

$$E[Y(t)] = Z(t)p_t = \int_0^t z(\tau)W(t - \tau, \tau)d\tau. \tag{20.8}$$

Remark 20.1. When $N(t)$ and the T_i are deterministic, it directly follows from (20.2) that

$$E[Y(t)] = \sum_{i=1}^{N(t)} W(t - T_i, T_i).$$

In summary, we have provided a stochastic model in which the $Y(t)$ in the formula (18.3) is the *expected output* $E[Y(t)]$ when $\Phi(z(t), t) = z(t)$. A similar interpretation holds for the input process $X(t)$ given in (18.4).

20.2 Input-Output Model with Batch Input

In this section we show that the $Y(t)$ in (18.3) is the *expected output* for general $\Phi(z(t), t)$. As in the previous section, the input process is a Poisson process, but now each input event corresponds to an **input batch** consisting of individual parts.[1] Let B_i denote the size of batch i corresponding to the occurrence time T_i. This batch size can be random and depend on T_i; if so, we assume this random variable has finite mean and variance, and is independent of the process time(s) and any other input batch. For example, the randomness could reflect the yield associated with a deterministic input batch.

We used the lead time distribution $W(\cdot, \tau)$ to determine the probability of the departure time of a *single* input that arrived at time τ. An input event now corresponds to an input *batch* of individual parts. We shall consider two possibilities.

[1] When the arrival times of the input events are governed by a Poisson process, but the events correspond to batch input (as described above), the overall input process is called a **compound Poisson process**.

20.2.1 Simultaneous Batch Case

For the **simultaneous batch** case, all parts in the batch are completed at the same time, and so $W(\cdot, \tau)$ denotes the cumulative distribution of the departure time of the entire batch. For this case, the cumulative output by time t is

$$Y(t) = \sum_{i=1}^{N(t)} B_i 1(T_i + W_i \leq t). \tag{20.9}$$

We now derive an expression for the expected cumulative output $E[Y(t)]$ up to time t. For a fixed $t > 0$, suppose $N(t) = n$, that is, n input batches have occurred by time t. Let $\tilde{Y}_1, \tilde{Y}_2, \ldots, \tilde{Y}_n$ denote the outputs obtained from the n input batches sequentially chosen at random. Note that $Y(t) = \sum_{i=1}^{N(t)} \tilde{Y}_i$. Recalling our system observer, the $\tilde{Y}_i$'s are independent, identically distributed random variables whose distribution coincides with the distribution of the output $\tilde{Y}$ of a randomly chosen input batch. Let $\tilde{T}$ denote the occurrence time with density (20.4), and let $\tilde{B}$ and $\tilde{W}$ denote its batch size and lead time, respectively. In the simultaneous case,

$$\tilde{Y} = \tilde{B}\ 1(\tilde{T} + \tilde{W} \leq t),$$

and

$$\begin{aligned} E[\tilde{Y} \mid \tilde{T}] &= E[\tilde{B}\ 1(\tilde{T} + \tilde{W} \leq t) \mid \tilde{T}], \\ &= E[\tilde{B} \mid \tilde{T}]\ P(\tilde{W} \leq t - \tilde{T} \mid \tilde{T}), \\ &= E[\tilde{B} \mid \tilde{T}] W(t - \tilde{T}, \tilde{T}). \end{aligned} \tag{20.10}$$

(The second line uses the assumption of independence of the input batch.) It then follows that

$$E[\tilde{Y}] = E\left(E[\tilde{Y} \mid \tilde{T}]\right) = \int_0^t E[\tilde{B} \mid \tilde{T} = \tau] W(t - \tau, \tau) \frac{z(\tau)}{Z(t)}\, d\tau\ . \tag{20.11}$$

Thus, the expected output up to time t is

$$E[Y(t)] = E\left(E[Y(t) \mid N(t)]\right) \tag{20.12}$$

$$= E[N(t)\tilde{Y}] \tag{20.13}$$

$$= Z(t) E[\tilde{Y}] \tag{20.14}$$

$$= \int_0^t E[\tilde{B} \mid \tilde{T} = \tau] W(t - \tau, \tau) z(\tau) d\tau\ . \tag{20.15}$$

Note that (20.15) is consistent with (20.8) when the batch size is identically 1, as it should.

20.2.2 Independent Batch Case

For the **independent batch** case, the lead times for the parts in the batch are assumed to be independent, identically distributed random variables with cumulative distribution $W(\cdot, \tau)$. For this case, the output is

$$Y(t) = \sum_{i=1}^{N(t)} \left\{ \sum_{j=1}^{B_i} 1(T_i + W_{ij} \le t) \right\}, \tag{20.16}$$

where W_{ij} denotes the process time of the j^{th} part in the i^{th} batch.

We now derive an expression for the expected cumulative output $E[Y(t)]$ up to time t for this case. For a fixed $t > 0$, suppose $N(t) = n$, that is, n input batches have occurred by time t. Adopting the notation of the previous section, in the independent batch case

$$\tilde{Y} = \sum_{j=1}^{\tilde{B}} 1(\tilde{T} + \tilde{W}_j),$$

where $\tilde{W}_j$ denotes the process time for the j^{th} part in the batch, and

$$\begin{aligned} E[\tilde{Y} \mid \tilde{T}] &= E\left(E[\tilde{Y} \mid \tilde{T},\ \tilde{B}]\right) \\ &= E\left(\tilde{B}\ W(t - \tilde{T}, \tilde{T}) \mid \tilde{T}\right) \\ &= E[\tilde{B} \mid \tilde{T}] W(t - \tilde{T}, \tilde{T}). \end{aligned}$$

The expression (20.17) is identical to (20.10) for the simultaneous batch case. Thus, identities (20.11)-(20.15) apply, from which we conclude that the expected outputs in these two cases are *identical*.

Remark 20.2. Suppose at time τ an input batch of size 100 enters the system and $W(t - \tau, \tau) = 0.50$ for some $t > \tau$. In the simultaneous batch case, the observed output at time t due to this input batch is either 100 or 0. Its expected value is, of course, 50, its variance is $(100)^2(0.5) - (50)^2 = 2500$, and thus its standard deviation is 50. In the independent batch case, the observed cumulative output could be any number between 0 and 100 and is governed by a binomial distribution with parameters 100 and 0.5. Its expected value is 50, too, but its variance is $100(0.5)(1 - 0.5) = 25$, and thus its standard deviation is only 5. While the expected cumulative output is the *same* regardless of the input batch model (simultaneous or independent), the *variabilities* of these output processes are significantly different. The first case has a higher variance—it's "all-or-nothing" after all.

20.3 Confidence intervals

20.3.1 Without Batch Input

We begin by examining the case when $\Phi(z(\tau), \tau) = z(\tau)$ and the expected output for the single input case is given by (20.8). We shall determine the distribution of $Y(t)$ by computing its moment generating function $\phi_{Y(t)}(\cdot)$.

We have established that conditioned on $N(t) = n$ the output $Y(t)$ is binomially distributed with parameters $N(t)$ and p_t—see (20.6). Applying (I.4), p. 509,

$$\begin{aligned} E[e^{uY(t)} \mid N(t)] &= [1 + p(e^u - 1)]^{N(t)} \\ &= e^{\hat{u}N(t)}, \end{aligned}$$

where

$$\hat{u} := \ln\,[1 + p(e^u - 1)].$$

Consequently,

$$\begin{aligned} \phi_{Y(t)}(u) &= E[e^{uY(t)}] \\ &= E\left(E[e^{uY(t)} \mid N(t)]\right) \\ &= E[e^{\hat{u}N(t)}] \\ &= \phi_{N(t)}(\hat{u}). \end{aligned}$$

Since the input $N(t)$ is a Poisson random variable with mean $Z(t)$, we can apply (I.5), p. 509, to obtain that

$$\begin{aligned} \phi_{N(t)}(\hat{u}) &= e^{Z(t)(e^{\hat{u}}-1)} \\ &= e^{Z(t)[p_t(e^u-1)]} \\ &= e^{(Z(t)p_t)(e^u-1)}. \end{aligned}$$

These chain of equalities yield the identity

$$\phi_{Y(t)}(u) = e^{(Z(t)p_t)(e^u-1)}. \tag{20.17}$$

We have shown that $Z(t)p_t = E[Y(t)]$—see (20.8). Comparing (20.17) to (I.5), and using the fact that the moment generating function uniquely characterizes the distribution, we conclude that $Y(t)$ *is a Poisson random variable with mean given in (20.8).*

Since the *distribution* for $Y(t)$ is known, for a fixed $\alpha \in (0, 1)$, it is possible to determine the smallest value $Y_\alpha^U(t)$ for which

$$P(Y(t) > Y_\alpha^U(t)) \geq 1 - \alpha. \tag{20.18}$$

Similarly, let $Y_\alpha^L(t)$ be the largest value for which

$$P(Y(t) < Y_{\alpha}^{L}(t)) \leq \alpha. \tag{20.19}$$

In lieu of the expected output curve $E[Y(\cdot)]$, one may use either $Y_{\alpha}^{U}(\cdot)$, a "liberal" estimate, or $Y_{\alpha}^{L}(\cdot)$, a "conservative" estimate. The liberal or conservative choices provide reasonable *bounds* to the true output curve.

Example 20.3. We examine $Var[Y(3)]$ for the motivating example describe in Section 17.1. When the batch size is identically one, $Var[Y(3)] = E[Y(3)]$; the mean values are provided in Table 18.4. In the constant loading, uniform distribution 'CU' case, the $Var[Y(3)] = 84$. When α is set to 0.025 in (20.18) and (20.19), the exact values for $Y_{0.0.25}^{U}(3)$ and $Y_{0.0.25}^{L}(3)$ are 102 and 66, respectively.

20.3.2 With Batch Input

When inputs arrive in batches, it is difficult to determine the entire distribution of $Y(t)$. As a practical alternative, we shall use the variance to provide confidence intervals on $Y(t)$, as follows. Conditioned on $N(t) = n$, we know that $Y(t)$ has the same distribution as the sum of independent, identically distributed random variables with finite mean and variance. If there are a large number of input batches, then the distribution of $Y(t)$ conditioned on $N(t) = n$ is approximately normal. The unconditioned distribution of $Y(t)$ will be approximately normal, too. Assuming $Y(t)$ is normally distributed, a confidence interval takes the form

$$\left(E[Y(t)] - \kappa(t)\sigma_{Y(t)},\ E[Y(t)] - \kappa(t)\sigma_{Y(t)}\right),$$

where $\sigma_{Y(t)}^2$ denotes the variance of $Y(t)$ and $\kappa(t)$ is the number of standard deviations selected to ensure the requisite coverage at time t. For example, $\kappa(t)$ would be set to 1.96 to ensure a 95% coverage, which corresponds to the choice of $\alpha = 0.025$ in (20.18) and (20.19).

Example 20.4. Instead of using the exact Poisson distribution in Example 20.3, a simple approximation is to use the standard deviation of $\sqrt{84} \approx 9$ to approximate the 2-sigma coverage interval corresponding to $\alpha = 0.025$. In this example, the 2-sigma coverage interval is calculated as $[84-2(9),\ 84+2(9)] = [66, 102]$, which is *exact* for this case.

Recalling that $\tilde{Y}$ denotes the output obtained from a randomly chosen input batch conditioned on $N(t)$, the variance of cumulative output is

$$\begin{aligned} Var[Y(t)] &= E[Var(Y(t) \mid N(t))] + Var(E[Y(t) \mid N(t)]) \\ &= E[N(t)Var(\tilde{Y})] + Var(N(t)E[\tilde{Y}]) \\ &= E[N(t)](E[\tilde{Y}^2] - (E[\tilde{Y}])^2) + (E[\tilde{Y}])^2 Var(N(t)) \\ &= Z(t)E[\tilde{Y}^2]. \end{aligned} \tag{20.20}$$

The first line above applies the conditional variance formula (I.7), p. 510, the second line uses the fact that $Y(t)$, conditioned on $N(t)$, is the sum of $N(t)$ independent, identically distributed random variables each with distribution identical to $\tilde{Y}$, the third line uses the independence of $N(t)$ and $\tilde{Y}$, and the fourth line uses the fact that $E[N(t)] = Var[N(t)] = Z(t)$.

It remains to compute $E[\tilde{Y}^2]$. In the simultaneous batch case, where all parts in the batch are completed at the same time, $\tilde{Y} = \tilde{B}\ 1(\tilde{T} + \tilde{W} \leq t)$, and so

$$\begin{aligned} E[\tilde{Y}^2 \mid \tilde{T}] &= E[\tilde{B}^2\ 1(\tilde{T} + \tilde{W} \leq t) \mid \tilde{T}] \\ &= E[\tilde{B}^2 \mid \tilde{T}] W(t - \tilde{T}, \tilde{T}). \end{aligned}$$

It follows that

$$\begin{aligned} E[\tilde{Y}^2] &= E\left(E[\tilde{Y}^2 \mid \tilde{T}]\right) \\ &= \int_0^t E[\tilde{Y}^2 \mid \tilde{T} = \tau]\ \frac{z(\tau)}{Z(t)} d\tau \\ &= \int_0^t E[\tilde{B}^2 \mid \tilde{T} = \tau] W(t - \tau, \tau) \frac{z(\tau)}{Z(t)} d\tau\ . \end{aligned}$$

Using this expression in (20.20), we conclude that

$$Var[Y(t)] = \int_0^t E[\tilde{B}^2 \mid \tilde{T} = \tau] W(t - \tau, \tau) z(\tau) d\tau\ . \tag{20.21}$$

Remark 20.5. Since $Y(t)$ is a Poisson random variable, $Var[Y(t)] = E[Y(t)]$. When the batch size is identically one, (20.21) reduces to (20.8), as it should.

In the independent batch case, where the departure times for the parts in the batch are independent and identically distributed, the conditional distribution of $\tilde{Y}$ given *both* $\tilde{T}$ and $\tilde{B}$ is binomial with parameters $\tilde{B}$ and $W(t - \tilde{T}, \tilde{T})$. Applying (I.1) and (I.2), p. 507,

$$\begin{aligned} E[\tilde{Y} \mid \tilde{T},\ \tilde{B}] &= \tilde{B} W(t - \tilde{T}, \tilde{T}), \\ Var[\tilde{Y} \mid \tilde{T},\ \tilde{B}] &= \tilde{B} W(t - \tilde{T}, \tilde{T})(1 - W(t - \tilde{T}, \tilde{T})). \end{aligned}$$

The definition of variance implies that

$$E[\tilde{Y}^2 \mid \tilde{T},\ \tilde{B}] = Var[\tilde{Y} \mid \tilde{T},\ \tilde{B}] + \left(E[\tilde{Y} \mid \tilde{T},\ \tilde{B}]\right)^2 .$$

Putting the last three identities together, we conclude that

$$E[\tilde{Y}^2 \mid \tilde{T},\ \tilde{B}] = [\tilde{B} W(t - \tilde{T}, \tilde{T})(1 - W(t - \tilde{T}, \tilde{T}))] + [\tilde{B} W(t - \tilde{T}, \tilde{T})]^2,$$

and so

$$E[\tilde{Y}^2 \mid \tilde{T}] = E\left(E[\tilde{Y}^2 \mid \tilde{T},\ \tilde{B}]\right)$$
$$= E[\tilde{B} \mid \tilde{T}]W(t-\tilde{T},\tilde{T})(1-W(t-\tilde{T},\tilde{T}))$$
$$+E[\tilde{B}^2 \mid \tilde{T}]W(t-\tilde{T},\tilde{T})^2. \qquad (20.22)$$

Since

$$E[\tilde{Y}^2] = \int_0^t E[\tilde{Y}^2 \mid \tilde{T}=\tau]\ \frac{z(\tau)}{Z(t)}d\tau\ ,$$

it follows from (20.20) and (20.22) that

$$Var[Y(t)] = \int_0^t E[\tilde{B} \mid \tilde{T}=\tau]W(t-\tau,\tau)(1-W(t-\tau,\tau))z(\tau)d\tau$$
$$+\int_0^t E[\tilde{B}^2 \mid \tilde{T}=\tau]W(t-\tau,\tau)^2 z(\tau)d\tau\ . \qquad (20.23)$$

When the batch size is identically one, (20.23) reduces to (20.8), as it should—see Remark 20.5.

Remark 20.6. It will be left to an exercise to verify that the variance (20.23) in the independent batch case is *always* less than the variance (20.21) in the simultaneous batch case.

Example 20.7. We calculate $Var[Y(3)]$ for the motivating example described in Section 17.1. In this example we consider the simultaneous batch case for which the batch size, B, is constant and time-invariant. We analyze the constant loading, uniform distribution 'CU' case. For this case $z(t) = z_i$ if $t \in [i-1, i)$, and the lead time distribution is defined as $W(t,\tau) = W(t) = t/2$ for $t \in [0,2]$. (The lead time distribution is independent of the arrival time τ.) Accordingly, if an input batch arrives at time $\tau \in [0,3]$, the probability of it completing by time 3 is

$$W(3-\tau) = \begin{cases} 1, & 0 \le \tau \le 1, \\ (3-\tau)/2, & 1 \le \tau \le 3. \end{cases}$$

From (20.21),

$$Var[Y(3)] = B^2 \int_0^3 W(3-\tau)z(\tau)d\tau$$
$$= B^2\left[z_1\int_0^1 1\ d\tau + z_2\int_1^2\left(\frac{3-\tau}{2}\right)d\tau + z_3\int_2^3\left(\frac{3-\tau}{2}\right)d\tau\right]$$
$$= B^2\Big\{z_1 + (3/4)z_2 + (1/4)z_3\Big\}. \qquad (20.24)$$

As a check, when $B = 1$ and $(z_1, z_2, z_3) = (24, 48, 96)$, $Var(Y(3)) = E[Y(3)] = 84$, as it should. When $B = 10$, the expected output $E[Y(3)]$ is 840 and its standard deviation is $\sqrt{8400} \approx 92$, and so the 2-sigma coverage interval is $[656, 1024]$.

Remark 20.8. The expression in braces in (20.24) is the expected output by time 3 if the batch size is one. In the notation of Chapter 18, the expected output here is $\Pi_3 z_1 + \Pi_2 z_2 + \Pi_1 z_3$.

Example 20.9. We continue with Example 20.7, except that here we consider the independent batch case. From (20.23),

$$Var[Y(3)] = B\int_0^3 W(3-\tau)(1-W(3-\tau))z(\tau)d\tau + B^2\int_0^3 W(3-\tau)^2 z(\tau)d\tau. \tag{20.25}$$

The first integral on the right-hand side of (20.25) is

$$B\left[z_1\int_0^1 0d\tau \;+\; z_2\int_1^2 \left(\frac{3-\tau}{2}\right)\left(1-\frac{3-\tau}{2}\right)d\tau \right.$$
$$\left. +\; z_3\int_2^3 \left(\frac{3-\tau}{2}\right)\left(1-\frac{3-\tau}{2}\right)d\tau\right],$$

and the second integral on the right-hand side of (20.25) is

$$B^2\left[z_1\int_0^1 1\;d\tau + z_2\int_1^2 \left(\frac{3-\tau}{2}\right)^2 d\tau + z_3\int_2^3 \left(\frac{3-\tau}{2}\right)^2 d\tau\right],$$

which after integration yields

$$Var[Y(3)] = B\Big\{(1/6)z_2 + (1/6)z_3\Big\} + B^2\Big\{z_1 + (7/12)z_2 + (1/12)z_3\Big\}.$$

As a check, when $B = 1$ and $(z_1, z_2, z_3) = (24, 48, 96)$, $Var(Y(3)) = E[Y(3)] = 84$, as it should. When $B = 10$, the expected output $E[Y(3)]$ is 840 and its standard deviation is $\sqrt{6240} = 79$, and so the 2-sigma coverage interval is $[682, 998]$.

Remark 20.10. As promised, the variance in the simultaneous batch case is larger than the variance in the independent batch case. Consequently, the corresponding 2-sigma coverage interval is larger, too.

20.3.3 Linear Approximation

In the discrete-time setting with a standard time grid, the variances (20.21) and (20.23) can be represented in the form

$$Var[Y(t)] = \sum_{i=1}^{t} z_i \int_{i-1}^{i} \xi_t(\tau)s_i(\tau)d\tau := \sum_{i=1}^{t} z_i \Upsilon_{it},$$

and so

$$\sigma_{Y(t)} := \sqrt{\sum_{i=1}^{t} z_i \Upsilon_{it}}\;.$$

The standard deviation is not linear, but can be approximated by taking a first-order Taylor series expansion about the standard deviation derived from some average trajectory $\bar{z}$ to obtain

$$\sigma_{Y(t)} \approx \sqrt{\sum_{i=1}^{t} \bar{z}_i \Upsilon_{it}} + \frac{\sum_{i=1}^{t} (z_i - \bar{z}_i)\Upsilon_{it}}{2\sqrt{\sum_{i=1}^{t} \bar{z}_i \Upsilon_{it}}},$$

$$= \frac{\sum_{i-1}^{t} z_i \Upsilon_{it}}{2\bar{\sigma}_{Y(t)}} + \bar{\sigma}_{Y(t)}/2, \tag{20.26}$$

where $\bar{\sigma}_{Y(t)}$ denotes the standard deviation of $Y(t)$ when $z = \bar{z}$.

This approximation is linear in the z_i. Since $E[Y(t)]$ is linear in the z_i, a projected output of the form

$$E[Y(t)] + \kappa(t)\sigma_{Y(t)}$$

will be linear in the z_i, too.

20.4 Exercises

20.1. Let X_1 denote a binomial random variable with parameters $n = 50$ and $p = 0.04$. Let X_2 denote a Poisson random variable with parameter $\lambda = 50(0.04) = 2$. Compare the probabilities $P(X_1 = k)$ and $P(X_2 = k)$ for $k = 0, 1, 2$.

20.2. Let $N(\cdot)$ denote a time-homogeneous Poisson process with intensity function $\lambda(t) = \lambda = 0.5$.

(a) What is the expected number of arrivals in the time interval $[0, 10]$?
(b) What is the probability that there will be no arrivals in this time interval?
(c) What is the probability that there will be at least 3 arrivals in this time interval?
(d) Answer parts (a)-(c) when the time interval is $[100, 110]$.

20.3. Let $N(\cdot)$ denote a Poisson process with intensity function $\lambda(t) = \sqrt{t}$.

(a) What is the expected number of arrivals in the time interval $[0, 4]$?
(b) What is the probability that there will be at least 2 arrivals in this time interval?
(c) Suppose it is known that there were 20 arrivals in this time interval. What is the probability that the arrival time of a randomly chosen arrival occurred in the first half of this time interval?
(d) Answer parts (a)-(c) when the time interval is $[4, 8]$.

20.4. The moment generating function of a discrete random variable X is $\phi_X(u) = [1 + 0.2(e^u - 1)]^{10}$.

(a) What is the mean and variance of X?
(b) What is the probability that $X \leq 2$?

20.5. The moment generating function of a discrete random variable X is $\phi_X(u) = e^{0.5(e^u-1)}$.

(a) What is the mean and variance of X?
(b) What is the probability that $X \leq 2$?

20.6. Let X_i, $i = 1, 2, \ldots$, be independent, identically distributed binomial random variables with parameters $n = 10$ and $p = 0.4$, and let N be a Poisson random variable with mean $\lambda = 2$. Consider the *random* sum $X = \sum_{i=1}^{N} X_i$. Determine the variance of X.

20.7. Suppose system inputs in the time interval $[0, 1]$ follow a time-homogeneous Poisson process with rate $\lambda = 10$.

(a) Suppose an input that enters the system at time $\tau \in [0, 1]$ will emerge as output by time 1 with probability $1 - \tau$.
  (i) What is the expected number of inputs in the time interval $[0, 1]$?
  (ii) What is the expected output by time 1?
(b) Answer part (a) when an input that enters the system at time $\tau \in [0, 1]$ will emerge as output by time 1 with probability $(1 - \tau)^2$.

20.8. Answer Exercise 20.7 when the inputs arrive *deterministically* at times $0.1(i - 1)$, $i = 1, 2, \ldots, 10$.

20.9. Suppose system inputs in the time interval $[0, 1]$ follow a Poisson process with intensity function $z(\tau) = 20\tau$.

(a) Suppose an input that enters the system at time $\tau \in [0, 1]$ will emerge as output by time 1 with probability $1 - \tau$.
  (i) What is the expected number of inputs in the time interval $[0, 0.5]$?
  (ii) What is the expected output by time 0.5?
  (iii) What is the expected number of inputs in the time interval $[0, 1]$?
  (iv) What is the expected output by time 1?
(b) Answer part (a) when an input that enters the system at time $\tau \in [0, 1]$ will emerge as output by time 1 with probability $(1 - \tau)^2$.

20.10. Assume a batch size of identically one. For the constant loading, uniform distribution 'CU' case described in the motivating example of Section 17.1, determine a 2-sigma coverage interval for the output $Y(2)$.

20.11. Consider the front loading, late distribution 'FL' case described in the motivating example of Section 17.1 with a constant batch size B.

(a) Assume the simultaneous batch case.
  (i) Express $Var[Y(3)]$ in terms of the z_i.
  (ii) Verify that $Var[Y(3)] = E[Y(3)]$ when $B = 1$.

(ii) Approximate $\sigma_{Y(3)}$ as a linear function of the z_i. Use $\bar{z} = (24, 48, 96)$.

(b) Assume the independent batch case.

(i) Express $Var[Y(3)]$ in terms of the z_i.

(ii) How does the variance here compare to the variance calculated in the simultaneous batch case?

(iii) Verify that $Var[Y(3)] = E[Y(3)]$ when $B = 1$.

(iv) Approximate $\sigma_{Y(3)}$ as a linear function of the z_i. Use $\bar{z} = (24, 48, 96)$.

20.12. Show that the variance (20.23) in the independent batch case is *always* less than the variance (20.21) in the simultaneous batch case.

20.5 Bibliographical Notes

Expression (20.8) can be viewed as a type of *transient* Little's Law. See Riano et. al. [2007] for a proof under more general conditions.

20.6 Solutions to Exercises

20.1 For the random variable X_1, we have $P(X_1 = 0) = (0.96)^{50} = 0.12989$, $P(X_1 = 1) = 50(0.04)(0.96)^{49} = 0.27060$, and $P(X_1 = 2) = (1225)(0.04)^2 \cdot (0.96)^{48} = 0.27623$. For the random variable X_2, we have $P(X_2 = 0) = e^{-2} = 0.13534$, $P(X_2 = 1) = 2e^{-2} = 0.27067$, and $P(X_2 = 2) = 2e^{-2} = 0.27067$. These respective values are quite close.

20.2 (a) $N(10)$ is a Poisson random variable with mean $0.5(10) = 5$.

(b) $P(N(10) = 0) = e^{-5} = 0.00674$.

(c)

$$\begin{aligned} P(N(10) \geq 3) &= 1 - P(N(10) < 3) \\ &= 1 - [P(N(10) = 0) + P(N(10) = 1) + P(N(10) = 1)] \\ &= 1 - [e^{-5} + 5e^{-5} + 5^2 e^{-5}/2] = 0.87535. \end{aligned}$$

(d) $N(110) - N(100)$ is a Poisson random variable with mean $0.5(10) = 5$. Hence, all of the answers remain the same.

20.3 (a) $\Lambda(t) = \int_0^t \lambda(\tau) d\tau = (2/3)t^{3/2}$. Since $\Lambda(4) = 5.\bar{3}$, $N(4)$ is a Poisson random variable with mean $5.\bar{3}$.

(b)

$$\begin{aligned} P(N(4) \geq 2) &= 1 - P(N(9) < 2) \\ &= 1 - [P(N(9) = 0) + P(N(9) = 1)] \\ &= 1 - [e^{-5.\bar{3}} + 5.\bar{3}e^{-5.\bar{3}}] = 0.96942. \end{aligned}$$

(c) This probability equals $\Lambda(2)/\Lambda(4) = 1.88562/5.\bar{3} = 0.35356$.

(d) Since $\Lambda(8) = 15.08494$ and $\Lambda(4) = 5.\bar{3}$, $N(8) - N(4)$ is a Poisson random variable with mean 9.75161. $P(N(8) - N(4) \geq 2) = 1 - [e^{-9.75161} + 9.75161e^{-9.75161}] = 0.99937$. The probability that the arrival time of a randomly chosen arrival occurred in the time interval $[4, 6]$ is $[\Lambda(6) - \Lambda(4)]/[\Lambda(8) - \Lambda(4)] = 0.45783$.

Remark 20.11. As time progresses, the square root function flattens out and approximates a linear function, and so this probability will converge to 0.5.

20.4 (a) The form of this moment generating function matches the one for the binomial distribution with parameters $n = 10$ and $p = 0.2$. Thus, the mean is $(10)(0.2) = 2$ and the variance is $(10)(0.2)(1 - 0.2) = 1.6$.

(b) $P(X \leq 2) = 0.8^{10} + 10(0.2)(0.8)^9 + 45(0.2)^2(0.8)^8 = 0.67780$.

20.5 (a) The form of this moment generating function matches the one for the Poisson distribution with parameter $\lambda = 0.5$. Thus, both the mean and variance equal 0.5.

(b) $P(X \leq 2) = e^{-0.5} + 0.5e^{-0.5} + 0.25e^{-0.5}/2 = 0.98561$.

20.6 We apply the conditional variance formula (I.7), p. 510, with $Y = N$. We have $E[X_i] = 10(0.4) = 4$, $Var[X_i] = 10(0.4)(1 - 0.4) = 1.6$, and $E[N] = Var[N] = 2$. Since $Var[X \mid N] = NVar[X_i]$ and $E[X \mid N] = NE[X_i]$, we have

$$\begin{aligned} Var[X] &= E(Var[X \mid N]) + Var(E[X \mid N]) \\ &= E[1.6N] + Var[4N] \\ &= 1.6(2) + 16(2) = 35.2. \end{aligned}$$

20.7 (a) System input in the time interval $[0, 1]$ is a Poisson random variable with mean 10. The expected output is $10 \int_0^1 (1 - \tau) d\tau = 5$.

(b) System input does not change. The expected output in this case is $10 \int_0^1 (1 - \tau)^2 d\tau = 10/3$.

20.8 Ten inputs arrive deterministically. In the first case, expected output is $1 + 0.9 + 0.8 + \cdots + 0.1 = 5.5$. In the second case, expected output is $1 + 0.9^2 + 0.8^2 + \cdots + 0.1^2 = 3.85$.

Remark 20.12. If N inputs arrive deterministically at times $t_i = (i - 1)/N$, $i = 1, 2, \ldots, N$, each with weight equal to $10/N$ (so that the total input equals 10), then the sum of the output as computed above will converge to the answers for Exercise 20.7, 5 and 10/3, respectively, as $N \to \infty$.

20.9 (a) (i) $Z(0.5) = \int_0^{0.5} 20\tau d\tau = 2.5$. (ii) Expected output is

$$\int_0^{0.5} 20\tau(1 - \tau) d\tau = 20[\tau^2/2 - \tau^3/3 \Big|_0^{0.5} = 5/3.$$

(iii) $Z(1) = \int_0^1 20\tau d\tau = 10$. (iv) Expected output is

$$\int_0^1 20\tau(1 - \tau) d\tau = 20[\tau^2/2 - \tau^3/3 \Big|_0^1 = 10/3.$$

(b) The input process does not change. As for the expected output,

$$E[Y(0.5)] = \int_0^{0.5} 20\tau(1 - \tau)^2 d\tau = 20[\tau^2/2 - 2\tau^3/3 + \tau^4/4 \Big|_0^{0.5} = 55/48.$$

$$E[Y(1)] = \int_0^1 20\tau(1 - \tau)^2 d\tau = 20[\tau^2/2 - 2\tau^3/3 + \tau^4/4 \Big|_0^1 = 5/3.$$

20.10 For the 'CU' case with a batch size of one, $Var[Y(2)] = E[Y(2)]$, which equals 30 from Table 18.4. When α is set to 0.025 in (20.18) and (20.19), the exact values for $Y^U_{0.0.25}(3)$ and $Y^L_{0.0.25}(3)$ are 41 and 19, respectively.

Remark 20.13. The standard deviation is $\sqrt{30} = 5.4772$. An approximate 2-sigma coverage interval is therefore $[30 - 2(5.4772), 30 + 2(5.4772)] = [19, 41]$, which is identical to the exact coverage interval.

20.11 For the 'FL' scenario, $s(t) = 2(1-t)$, which means that $z(\tau) = z_1[2(1-\tau)]$ if $\tau \in [0,1]$, $z(\tau) = z_2[2(2-\tau)]$ if $\tau \in [1,2]$, and $z(\tau) = z_3[2(3-\tau)]$ if $\tau \in [2,3]$. The lead time distribution is $W(t) = t^2/4$ for $t \in [0,2]$.

(a) (i) In the simultaneous batch case,

$$\begin{aligned} Var[Y(3)] &= B^2 \int_0^3 W(3-\tau)z(\tau)d\tau \\ &= B^2\Big[z_1 \int_0^1 2(1-\tau)\cdot 1\, d\tau + z_2 \int_1^2 [(3-\tau)^2/4][2(2-\tau)]d\tau \\ &\qquad + z_3 \int_2^3 [(3-\tau)^2/4][2(3-\tau)]d\tau\Big] \\ &= B^2\{\hat{\Pi}_3 z_1 + \hat{\Pi}_2 z_2 + \hat{\Pi}_1 z_3\} \\ &= B^2\{z_1 + (17/24)z_2 + (1/8)z_3\}. \end{aligned}$$

The next-to-last line follows the observation in Remark 20.8, p. 382, and the last line uses the $\hat{\pi}_i$ numbers in Table 18.4. (ii) When $B = 1$ and $z = (z_1, z_2, z_3) = (24, 48, 96)$, $Var[Y(3)] = E[Y(3)] = 70$, as it should. (iii) Using (20.26), $\sigma_{Y(3)}$ approximately equals

$$\frac{B^2 z_1 + B^2(17/24)z_2 + B^2(1/8)z_3}{2B\sqrt{70}} + B\sqrt{70}/2.$$

(b) (i) In the independent batch case,

$$\begin{aligned} Var[Y(t)] &= B\int_0^t W(t-\tau)(1-W(t-\tau))z(\tau)d\tau \\ &\quad + B^2 \int_0^t W(t-\tau)^2 z(\tau)d\tau \\ &= B\int_0^3 W(3-\tau)z(\tau)d\tau \\ &\quad + (B^2 - B)\int_0^3 W(3-\tau)^2 z(\tau)d\tau. \end{aligned} \tag{20.27}$$

Using the observation in Remark 20.8, the first integral on the right-hand side of (20.27) evaluates to

$$B[\hat{\Pi}_3 z_1 + \hat{\Pi}_2 z_2 + \hat{\Pi}_1 z_3] = B[z_1 + (17/24)z_2 + (1/8)z_3].$$

The second integral on the right-hand side of (20.27) is

$$(B^2 - B)\Big[z_1 \int_0^1 2(1-\tau)d\tau + z_2 \int_1^2 2(2-\tau)[(3-\tau)^2/4]^2 d\tau \\ + z_3 \int_1^2 2(3-\tau)[(3-\tau)^2/4]^2 d\tau\Big],$$

which evaluates to

$$(B^2 - B)[z_1 + (43/80)z_2 + (1/48)z_3].$$

Thus,

$$Var[Y(3)] = B^2 z_1 + [(43/80)B^2 + (41/240)B]z_2 + [(1/48)B^2 + (5/48)B]z_3. \tag{20.28}$$

(ii) When $B = 1$ and $z = (z_1, z_2, z_3) = (24, 48, 96)$, $Var[Y(3)] = E[Y(3)] = 70$, as it should. (iii) Let $\bar{\sigma}_{Y(3)}$ denote the square root of the expression (20.28) for $Var[Y(3)]$ evaluated at $z = (24, 48, 96)$. Using (20.26), $\sigma_{Y(3)}$ approximately equals

$$\frac{B^2 z_1 + [(43/80)B^2 + (41/240)B]z_2 + [(1/48)B^2 + (5/48)B]z_3}{\bar{\sigma}_{Y(3)}} \; + \; \bar{\sigma}_{Y(3)}/2.$$

20.12 In the independent batch case,

$$\begin{aligned} Var[Y(t)] &= \int_0^t E[\tilde{B} \mid \tilde{T} = \tau] W(t-\tau, \tau)(1 - W(t-\tau, \tau)) z(\tau) d\tau \\ &\qquad + \int_0^t E[\tilde{B}^2 \mid \tilde{T} = \tau] W(t-\tau, \tau)^2 z(\tau) d\tau \\ &= \int_0^t E[\tilde{B} \mid \tilde{T} = \tau] W(t-\tau, \tau) z(\tau) d\tau + \\ &\qquad \int_0^t \Big(E[\tilde{B}^2 \mid \tilde{T} = \tau] - E[\tilde{B} \mid \tilde{T} = \tau] \Big) W(t-\tau, \tau)^2 z(\tau) d\tau. \end{aligned}$$

The value

$$E[\tilde{B}^2 \mid \tilde{T} = \tau] - E[\tilde{B} \mid \tilde{T} = \tau]$$

is nonnegative since $\tilde{B}^2 \geq \tilde{B} \geq 1$. Since $W(t-\tau, \tau)^2 \leq W(t-\tau, \tau) \leq 1$ and $z(\tau) \geq 0$, it follows that

$$\begin{aligned} Var[Y(t)] &\leq \int_0^t E[\tilde{B} \mid \tilde{T} = \tau] W(t-\tau, \tau) z(\tau) d\tau \\ &\qquad + \int_0^t \Big(E[\tilde{B}^2 \mid \tilde{T} = \tau] - E[\tilde{B} \mid \tilde{T} = \tau] \Big) W(t-\tau, \tau) z(\tau) d\tau \\ &= \int_0^t E[\tilde{B}^2 \mid \tilde{T} = \tau] W(t-\tau, \tau) z(\tau) d\tau \\ &= Var[Y(t)] \text{ in the simultaneous batch case.} \end{aligned}$$

21

Multi-Stage, Dynamic Models of Technology

A production process typically begin with raw materials, parts, subassemblies, and transforms them via several intermediate stages to produce final outputs sold to end users. At the "molecular level" a production process is a network of *activities* or *stages*. Each activity's input-output process is characterized by a dynamic production function. The multi-stage models developed in this chapter are most useful for short-term production planning.

Storable goods are all raw materials, purchased parts or subassemblies, and intermediate or final products produced by activities. Storable goods used by an activity will either be acquired from outside the system (i.e., exogenously supplied) or obtained via *intermediate product transfers* from other activities within the system. Material balance constraints are required to ensure that the requisite storable inputs are available at the time they are used in the production process. Service capacity equations are required to ensure that the rates of the aggregate machine and labor services available are sufficient to meet internal aggregate demand.

We begin by describing a basic, continuous-time model of dynamic production involving a network of interrelated activities. Next, we develop specific models by substituting the instantaneous, constant lead time, multi-event lead time, and distribution-based dynamic production functions into the fundamental equations. Practical considerations, such as how to handle initial conditions, are also discussed. We show how to translate these continuous-time models into their discrete-time counterparts suitable for computation. Two examples from manufacturing and assembly with rework are described in detail. We describe several practical extensions to the basic model. We close this chapter with a discussion of how to connect the models described herein to efficiency and productivity analysis.

21.1 Basic Model

21.1.1 Primitives

A production system consists of N producing activities, labeled $1, 2, \ldots, N$, and two non-producing activities, labeled 0 and $N + 1$, respectively.[1]

- *Producing activities.* Each producing activity i supplies a unique storable product also labeled i. (No services are produced internally.) Each producing activity's technology is governed by a dynamic production function. The vector
$$x_i = (x_i^1(\cdot), x_i^2(\cdot), \ldots, x_i^n(\cdot))$$
denotes the inputs used by activity i, and
$$y_i(\cdot) = [f_i(x_i)](\cdot)$$
denotes the output produced by activity i from input x_i. Each component of x_i and the output y_i are functions of time.
- *Non-producing activities.* Activity 0, the *source* activity, supplies the exogenous services, raw materials, and purchased parts or subassemblies from outside vendors to the producing activities. The source activity does not provide any product made by a producing activity. Activity $N+1$, the *sink* activity, receives outputs from the producing activities. (An intermediate product will not be sent to the sink activity if it is not "sold" to an end user.)

Let $v_{i,j}(t)$ denote the transfer of product i by activity i *sent* to activity j at time t, and let $v_{0,j}^m(t)$ denote the transfer of material m by activity 0 *sent* to activity j at time t. Transfers of product may not be instantaneous. Let
$$\hat{v}_{i,j}(t) = t_{i,j}[v_{i,j}](t)$$
denote the transfer of product i by activity i *received* by activity j at time t, and let
$$\hat{v}_{0,j}^m(t) = t_{0,j}^m[v_{0,j}^m](t)$$
denote the transfer of material m by activity 0 *received* by activity j at time t.[2] Conceptually, each $t_{i,j}[\,\cdot\,]$ is a dynamic production function.[3]

[1] These labels are replaced with more descriptive names in actual applications.

[2] The transformation functions $\hat{v}_{i,j}(\cdot) = [t_{i,j}(v_{i,j}](\cdot)$ implicitly assume there are no "joint" constraints on the $v_{i,j}(\cdot)$, to achieve, for example, economies of transportation. The detailed modeling of such transformations is beyond the scope of this chapter.

[3] Activities labeled 'ij' could be introduced and the symbol t could be replaced with an f. Such notation is cumbersome, excessively so when extensions of the basic model are described. For notational simplicity, we do not identify these special transfer functions with separate activities.

Below, we define sets of constraints that define the technology $\mathcal{T}$ for the basic model. First, some notation. The vector $\tilde{x} = (\tilde{x}^1, \tilde{x}^2, \ldots, \tilde{x}^n)$ denotes an exogenous *system* input vector, and the vector $\tilde{y} = (\tilde{y}^1, \tilde{y}^2, \ldots, \tilde{y}^N)$ denotes a feasible vector of outputs obtained from the system. The letter 's' will denote a generic service input, the letter m will denote a generic material input, and the letters i, j will denote generic products produced by the system. All flows are rate-based. For a generic rate-based flow $h(\cdot)$, its corresponding cumulative flow will be denoted by the upper case letter H, i.e., $H(t) = \int_{-\infty}^{t} h(\tau)d\tau$. (All integrals are assumed to exist.)

For the basic model, $(\tilde{x}, \tilde{y}) \in \mathcal{T}$ if there exists producing activity input vectors x_i, transfers sent $v_{i,j}(\cdot)$, and transfers received $\hat{v}_{i,j}(\cdot)$ such that the material balance and service capacity constraints (21.2)-(21.6) defined in the next section are satisfied.

21.1.2 Material Balance and Service Capacity Constraints

For two points in time s, t, $s < t$, material balance can be conceptually represented as

$$I(t) = I(s) + FlowIn[s,t] - FlowOut[s,t]. \tag{21.1}$$

In (21.1), $I(t)$ denotes the inventory (amount stored) of a generic good at time t, FlowIn$[s,t]$ denotes the cumulative amount of this good "flowing in" during the time interval $[s,t]$, and FlowOut$[s,t]$ denotes the cumulative amount of this good "flowing out" during the time interval $[s,t]$. We adopt standard convention and assign time 0 as the point in time when new production takes place. All flows prior to time 0 are assumed known, as well as the initial inventory $I(0)$ at time 0.

Remark 21.1. When production processes are not instantaneous, $FlowIn[0,t]$ will be the result of decisions made *prior* to time 0.

Let $I_i^m(t)$ and $I_i^j(t)$ denote, respectively, the inventory of material m and product j stored at activity i at time t. The material balance constraints are as follows:

- For each producing activity i and for all $t \geq 0$:

$$I_i^i(t) = I_i^i(0) + \int_0^t y_i(\tau)d\tau - \sum_{j=1}^{N+1} \int_0^t v_{ij}(\tau)d\tau \geq 0. \tag{21.2}$$

  These constraints ensure that the production of product i is sufficient to meet all system-wide demands.
- For each producing activity i, (internally produced) product $j \neq i$, and for all $t \geq 0$:

$$I_i^j(t) = I_i^j(0) + \int_0^t \hat{v}_{j,i}(\tau)d\tau - \int_0^t x_i^j(\tau)d\tau \geq 0. \tag{21.3}$$

These constraints ensure that there is sufficient supply of product j at activity i to support its usage in activity i's production process.

- For each producing activity i, (exogenously supplied) material m, and for all $t \geq 0$:

$$I_i^m(t) = I_i^m(0) + \int_0^t \hat{v}_{0,i}^m(\tau)d\tau - \int_0^t x_i^m(\tau)d\tau \geq 0. \tag{21.4}$$

These constraints ensure that there is sufficient supply of material m at activity i to support its usage in activity i's production process.

- For each product i and for all $t \geq 0$:

$$I_{N+1}^i(t) = I_{N+1}^i(0) + \int_0^t \hat{v}_{i,N+1}(\tau)d\tau - \int_0^t \tilde{y}^i(\tau)d\tau \geq 0. \tag{21.5}$$

These constraints ensure that final demands, as represented by the output vector $\tilde{y}$, are met.

The service capacity constraints are as follows:

- For each service s and for all $\tau \geq 0$:

$$\sum_{i=1}^{N} x_i^s(\tau) \leq \tilde{x}^s(\tau). \tag{21.6}$$

21.2 Index-Based Models

21.2.1 Instantaneous Processes

In this model, production is instantaneous and the inputs used by each activity are in constant proportions. (See the description of the *fixed proportions dynamic model* on p. 300.) Since each activity produces a unique output and production is instantaneous, the index $z_i(\cdot)$ will be measured in units of output. Consequently, the input vector is

$$x_i(\tau) = (a_i^1, a_i^2, \ldots, a_i^n)z_i(\tau), \tag{21.7}$$

and the output is

$$y_i(\tau) = [f_i(x_i)](\tau) = z_i(\tau). \tag{21.8}$$

Transfers of product and raw materials are instantaneous. Inventory arrives just-in-time for its use. There are no other constraints on transfers. Consequently,

$$\hat{v}_{i,j}(\tau) = v_{i,j}(\tau) = a_j^i z_j(\tau), \quad 1 \leq i, j \leq N, \tag{21.9}$$

$$\hat{v}_{0,i}^m(\tau) = v_{0,i}^m(\tau) = a_i^m z_i(\tau), \ 1 \leq i \leq N, \text{ all } m, \tag{21.10}$$

$$\hat{v}_{i,N+1}(\tau) = v_{i,N+1}(\tau) = \tilde{y}^i(\tau), \quad 1 \leq i \leq N. \tag{21.11}$$

Given (21.7)-(21.11), the material balance constraints (21.3)-(21.5) are automatically satisfied and therefore can be ignored. Material balance constraint (21.2) and service capacity constraint (21.6) become, respectively,

- For each producing activity i and for all $t \geq 0$:

$$I_i^i(t) = I_i^i(0) + \int_0^t z_i(\tau)d\tau - \sum_{j=1}^{N} \int_0^t a_j^i z_j(\tau)d\tau - \int_0^t \tilde{y}^i(\tau)d\tau \geq 0. \quad (21.12)$$

- For each service s and for all $\tau \geq 0$:

$$\sum_{i=1}^{N} a_i^s z_i(\tau) \leq \tilde{x}^s(\tau). \quad (21.13)$$

21.2.2 Constant Lead Time Processes

In some production processes, a product cannot be released to inventory for subsequent use for a period of time after it has been produced (the services have been applied). For example, a fixed amount of time may be required for a part to dry after a painting operation or to inspect and grade output. We assume that no scarce resources are consumed by these operations; otherwise, such resource consumption must be included in the model. In this model, inputs used by each activity are still in constant proportions, but output by activity i emerges a constant ℓ_i units of time after application of resources. In this setting, $z_i(\cdot)$ indexes "starts," i.e., when application of service resources and withdrawal of intermediate products and materials begins. The input vector x_i is still represented as

$$x_i(\tau) = (a_i^1, a_i^2, \ldots, a_i^n)z_i(\tau), \quad (21.14)$$

but the output is now

$$y_i(\tau) = [f_i(x_i)](\tau) = z_i(\tau - \ell_i). \quad (21.15)$$

In some production processes, a period of time elapses after a product has been withdrawn from inventory until it is used as input at a follow-on activity. Examples include transportation time or time to complete inspection. Once again, no scarce resources are consumed by these operations. Transfers of product sent from producing activity i to producing activity j are received a constant $\ell_{i,j}$ units of time later. Transfers of material m sent from the source activity 0 to producing activity i are received a constant $\ell_{0,i}^m$ units of time later. There are no other constraints on transfers. Consequently,

$$\hat{v}_{i,j}(\tau) = v_{i,j}(\tau - \ell_{i,j}), \ \ 1 \leq \ i, j \ \leq N, \quad (21.16)$$

$$\hat{v}_{0,i}^m(\tau) = v_{0,i}^m(\tau - \ell_{0,i}^m), \ 1 \leq i \leq N, \text{ all } m, \quad (21.17)$$

$$\hat{v}_{i,N+1}(\tau) = v_{i,N+1}(\tau - \ell_{i,N+1}), \ \ 1 \leq i \leq N. \quad (21.18)$$

Given (21.14)-(21.18), the material balance constraints (21.2)-(21.5) and service capacity constraint (21.6) become, respectively,

- For each producing activity i and for all $t \geq 0$:

$$I_i^i(t) = I_i^i(0) + \int_0^t z_i(\tau - \ell_i)d\tau - \sum_{j=1}^{N+1} \int_0^t v_{ij}(\tau)d\tau \geq 0. \quad (21.19)$$

- For each producing activity i, (internally produced) product $j \neq i$ and for all $t \geq 0$:

$$I_i^j(t) = I_i^j(0) + \int_0^t v_{j,i}(\tau - \ell_{j,i})d\tau - \int_0^t a_i^j z_i(\tau)d\tau \geq 0. \quad (21.20)$$

- For each producing activity i, (exogenously supplied) material m, and for all $t \geq 0$:

$$I_i^m(t) = I_i^m(0) + \int_0^t v_{0,i}^m(\tau - \ell_{0,i}^m)d\tau - \int_0^t a_i^m z_i(\tau)d\tau \geq 0. \quad (21.21)$$

- For each product i and for all $t \geq 0$:

$$I_{N+1}^i(t) = I_{N+1}^i(0) + \int_0^t \hat{v}_{i,N+1}(\tau - \ell_{i,N+1})d\tau - \int_0^t \tilde{y}^i(\tau)d\tau \geq 0. \quad (21.22)$$

- For each service s and for all $\tau \geq 0$:

$$\sum_{i=1}^{N} a_i^s z_i(\tau) \leq \tilde{x}^s(\tau). \quad (21.23)$$

21.2.3 Multi-Event, Constant Lead Time Processes

Following the description of multi-event, constant lead time processes in Section 18.2.1, p. 313, in this section the output rate is

$$y_i(\tau) = \sum_r w_{i,r}^y z_i(\tau - \ell_{i,r}^y), \quad (21.24)$$

and the input vector is

$$x_i(\tau) = (a_i^1, a_i^2, \ldots, a_i^n) \left(\sum_r w_{i,r}^x z_i(\tau - \ell_{i,r}^x) \right). \quad (21.25)$$

In (21.24) and (21.25), the weights $w_{i,r}^y$ and $w_{i,r}^x$ are positive constants that respectively sum to one, and the lead times $\ell_{i,r}^y$ and $\ell_{i,r}^x$ are positive. Similarly, for activities $i \in \{1, 2, \ldots, N\}$, $j \in \{1, 2, \ldots, N+1\}$, $i \neq j$,

$$\hat{v}_{i,j}(\tau) = \sum_r w_{i,j,r} v_{i,j}(\tau - \ell_{i,j,r}), \quad (21.26)$$

and for all materials m

$$\hat{v}^m_{0,i}(\tau) = \sum_r w_{0,i,r} v^m_{0,i}(\tau - \ell_{0,i,r}). \tag{21.27}$$

In (21.26) and (21.27), the respective weights some to one and the lead times are all positive.

It is straightforward to substitute identities (21.24)-(21.27) into (21.2)-(21.6) to obtain the material balance and system capacity constraints. For example, the expressions on the left-hand side below are replaced by the expressions on the right-hand side

$$\int_0^t z_i(\tau - \ell_i)d\tau \quad \longleftarrow \quad \int_0^t \sum_r w^y_{i,r} z_i(\tau - \ell^y_{i,r})d\tau, \tag{21.28}$$

$$\int_0^t v_{j,i}(\tau - \ell_{j,i})d\tau \quad \longleftarrow \quad \int_0^t \sum_r w_{j,i,r} v_{j,i}(\tau - \ell_{j,i,r})d\tau, \tag{21.29}$$

$$\sum_{i=1}^N a^s_i z_i(\tau) \quad \longleftarrow \quad \sum_{i=1}^N a^s_i \left(\sum_r w^x_{i,r} z_i(\tau - \ell^x_{i,r}) \right) \tag{21.30}$$

in equations (21.19), (21.20), and (21.23), respectively.

21.2.4 Continuous Lead Time Based Processes

Following the description of continuous lead time processes in Section 18.4, p. 324, in this section the *cumulative* output is

$$Y_i(t) = \int_{-\infty}^t z_i(\tau) W^y_i(t - \tau, \tau)d\tau, \tag{21.31}$$

and the input vector is

$$x_i(\tau) = (a^1_i, a^2_i, \ldots, a^n_i) \left(\int_{-\infty}^t z_i(\tau) w^x_i(t - \tau, \tau)d\tau \right). \tag{21.32}$$

In (21.31) and (21.32), the $W^y_i(\cdot,\cdot)$ and $w^x_i(\cdot,\cdot)$ are cumulative lead time distributions and lead time densities, respectively. Similarly, for activities $i \in \{1, 2, \ldots, N\}$, $j \in \{1, 2, \ldots, N+1\}$, $i \neq j$, the *cumulative* product transfers are

$$\hat{V}_{i,j}(\tau) = \int_{-\infty}^t v_{i,j}(\tau) W_{i,j}(t - \tau, \tau)d\tau, \tag{21.33}$$

and for all materials m

$$\hat{V}^m_{0,i}(\tau) = \int_{-\infty}^t v^m_{0,i}(\tau) W^m_{i,j}(t - \tau, \tau)d\tau. \tag{21.34}$$

In (21.33) and (21.34), the $W_{i,j}(\cdot,\cdot)$ and $W^m_{i,j}(\cdot,\cdot)$ are cumulative lead time distributions.

It is a bit less straightforward to substitute identities (21.31)-(21.34) into (21.2)-(21.6) to obtain the material balance and system capacity constraints. Since the expressions for output and transfers above are defined in terms of their cumulative values, it is necessary to subtract the cumulative amount up to time 0 to ensure that only flow in the interval $[0, t]$ is counted. For example, the expressions on the left-hand side of (21.28)-(21.30) are replaced respectively by the expressions

$$\begin{aligned} Y_i(t) - Y_i(0) &= \int_{-\infty}^{t} z_i(\tau) W_i^y(t-\tau,\tau)d\tau \;-\; \int_{-\infty}^{0} z_i(\tau) W_i^y(-\tau,\tau)d\tau, \\ \hat{V}_{j,i}(t) - \hat{V}_{j,i}(0) &= \int_{-\infty}^{t} v_{j,i}(\tau) W_{j,i}(t-\tau,\tau)d\tau \;-\; \int_{-\infty}^{0} v_{j,i}(\tau) W_{j,i}(-\tau,\tau)d\tau, \\ & \qquad a_i^s \left(\int_{-\infty}^{t} z_i(\tau) w_i^x(t-\tau,\tau)d\tau \right) \end{aligned}$$

above in equations (21.19), (21.20) and (21.23), respectively.

21.2.5 Initial Conditions

The current time 0 marks the point in time when new calculations/decisions can be made. When all production processes are instantaneous, all information about the past decisions is conveniently encapsulated in the (initial) inventories at time 0. However, when production processes are not instantaneous, outputs and transfers could be realized after the current time 0 as a result of prior decisions before the current time 0. The past is no longer summarized by the initial inventories; it is necessary to project, in some way, the flow after the current time 0 as a result of decisions made prior to time 0.

We describe two approaches for making the requisite projections.

Approach I: Apply model's projections to prior decisions

In this approach, all values of variables associated with time $\tau < 0$ are considered part of the past and are *pre-specified*. For example, in the constraints (21.19)-(21.23) associated with the constant lead time model, values of index $z_i(\cdot)$ are pre-specified for $-\ell_i \le \tau \le 0$, values for intermediate product transfers $v_{i,j}(\cdot)$, $1 \le\ i, j,\ \le N$, are pre-specified for $-\ell_{i,j} \le \tau \le 0$, and values for raw material transfers $v^m_{0,N+1}(\cdot)$ are pre-specified for $-\ell^m_{0,N+1} \le \tau \le 0$.

Remark 21.2. In the constant lead time model, it is possible to go one step further. As previously mentioned, intermediate product or raw material transfers were sent prior to the current time 0 ostensibly to support *planned* production after the current time 0 when the model was previously run. Suppose

that all previously planned production that required transfers prior to the current time 0 is *frozen*, i.e., not permitted to change, and suppose all inventory arrives just-in-time. In addition to constraints (21.16)-(21.18), these assumptions imply that

$$\tilde{v}_{i,j}(\tau) = a_j^i z_j(\tau),\ \tilde{v}_{0,i}^m(\tau) = a_i^m z_i(\tau),\ \tilde{v}_{i,N+1}(\tau) = \tilde{y}^i(\tau).$$

It then follows that the values of index $z_i(\cdot)$ must now be pre-specified for $-\ell_i \leq \tau \leq \max_{j,m}\{\ell_{j,i}, \ell_{0,i}^m\}$. Constraints (21.20)-(21.22) are now automatically satisfied, and constraint (21.19) becomes: for each producing activity i and for all $t \geq 0$,

$$I_i^i(t) = I_i^i(0) + \int_0^t z_i(\tau - \ell_i)d\tau - \sum_{j=1}^N \int_0^t a_j^i z_j(\tau + \ell_{i,j})d\tau - \int_0^t \tilde{y}^i(\tau + \ell_{i,N+1})d\tau. \tag{21.35}$$

This constraint can also be expressed as

$$I_i^i(t) = I_i^i(0) + \int_{-\ell_i}^{t-\ell_i} z_i(\tau)d\tau - \sum_{j=1}^N \int_{-\ell_{i,j}}^{t-\ell_{i,j}} a_j^i z_j(\tau)d\tau - \int_{-\ell_{i,N+1}}^{t-\ell_{i,N+1}} \tilde{y}^i(\tau)d\tau. \tag{21.36}$$

In this setting, there is no inventory of product i at activity $j \neq i$; however, there are *pipeline inventories* in the amounts of

$$\int_{t-\ell_{i,j}}^t a_i^j z_i(\tau)d\tau \text{ and } \int_{t-\ell_{0,i}^m}^t a_i^m z_i(\tau)d\tau$$

throughout the system.

With new information available at time 0, it can be economically beneficial to *change* previously planned production. When *new* calculations are permitted after time 0, requiring that the index $z_i(\cdot)$ be frozen *after* time 0 is unnecessarily restrictive. The drawback of the first approach is that it applies the model of the input-output transformation to what occurred prior to the current time 0. With today's information systems, it can be possible to know how the flows prior to the current time 0 have *in actuality* progressed through the system. This information is not being used in the first approach. For example, suppose a 100 units were started in the interval $(-3, -2)$. Suppose that as a result of this input—*according to the model*—there should be a queue of 60 units at time 0, 40 of which should emerge as output in the interval $(0, 1)$. Suppose we know that, in actuality, the queue only contains 30 units, half as much. It would be more accurate—*ceteris paribus*—to project only 20 units of output emerging in the interval of $(0, 1)$. It could be the case that the other 30 units had already emerged as output ahead of schedule. If this is the case, then this output is already counted as completed inventory (as of time 0) and should not be double-counted. It could be the case that the other 30 units had been removed from the system due to poor quality, which means such input will never emerge as output.

Approach II: Use shop-floor information to make independent projections

In this approach, shop-floor information is used to obtain the status of all stages of the work-in-process, including the current account of completed output. Given this additional information, one has to model how this work-in-process will emerge as output, which could entail *different* transformations than the ones used in the model. While more accurate, it requires additional "bean-counting."

To adopt this approach in the prior model formulations, all values of the variables prior to the current time 0 are set to zero. Flow functions are added to the equations to project future flow as a function of past flow. For example, in the constant lead time model, cumulative output on the interval $[0, t]$ becomes

$$\int_0^t y_i(\tau)d\tau = \tilde{F}_i(t) + \int_0^{t-\ell_i} z_i(\tau)d\tau. \tag{21.37}$$

Here, $\tilde{F}_i(\cdot)$ represents the projected schedule of output after the current time 0 due to prior decisions before the current time 0. Similarly, in the multi-event, constant lead time model, cumulative output on the interval $[0, t]$ becomes

$$\int_0^t y_i(\tau)d\tau = \tilde{F}_i(t) + \sum_r \int_0^{t-\ell_{i,r}} w_r^y z_i(\tau)d\tau. \tag{21.38}$$

For the continuous lead time model,

$$Y_i(t) - Y_i(0) = \tilde{F}_i(t) + \int_0^t z_i(\tau)W_i^y(t-\tau,\tau)d\tau. \tag{21.39}$$

With respect to the constraints involving the transfers received, the projected flow functions added to these constraints are represented as either $\tilde{V}_{j,i}(t)$ and $\tilde{V}_{0,i}^m(t)$.

Remark 21.3. Adopting Approach I to project cumulative output for the continuous lead time model,

$$\begin{aligned} Y_i(t) - Y_i(0) &= \int_{-\infty}^t z_i(\tau)W_i^y(t-\tau,\tau)d\tau \; - \; \int_{-\infty}^0 z_i(\tau)W_i^y(-\tau,\tau)d\tau \\ &= \int_{-\infty}^0 z_i(\tau)\left[W_i^y(t-\tau,\tau) - W_i^y(-\tau,\tau)\right]d\tau + \int_0^t z_i(\tau)W_i^y(t-\tau,\tau)d\tau. \end{aligned}$$

If the model is perfectly accurate, then the appropriate projected flow function to use for Approach II is

$$\tilde{F}_i(t) := \int_{-\infty}^0 z_i(\tau)\left[W_i^y(t-\tau,\tau)\right] - W_i^y(-\tau,\tau)\right]d\tau.$$

21.3 Computational Models

The continuous-time models developed in the previous section are impractical from a computational perspective. A discrete-time approximation is used to generate a computational model.

An appropriate time grid $\mathcal{G}$ is selected. The $z_i(\cdot)$, the transfers $v_{i,j}(\cdot)$, $v^m_{0,i}(\cdot)$, the exogenous inputs $\tilde{x}$, and the final outputs $\tilde{y}^i(\cdot)$ are each constrained to be constant on each period $[t_{k-1}, t_k)$. Consequently, their corresponding cumulative functions, $Z_i(\cdot)$, $V_{i,j}(\cdot)$, $V^m_{0,i}(\cdot)$, $\tilde{Y}^i(\cdot)$ are each piecewise linear with constant slope on each period.

It is standard practice to maintain material balance only at the time grid points t_k. The following example illustrates this practice could lead to a production plan that appears feasible, but is, in fact, *not* achievable.

Example 21.4. A standard time grid is assumed. All flows prior to time 0 are zero. There are no initial inventories. Two units of product i are required per unit of product j, i.e., $a^i_j = 2$. Activity i's production plan calls for a constant rate of 100 starts in period 1 and activity j's production plan calls for a constant rate of 40 starts in period 4, i.e.,

$$z_i(\tau) = 250 \cdot 1_{[0,1)}(\cdot) \text{ and } z_j(\tau) = 40 \cdot 1_{[3,4)}(\cdot).$$

If production processes are instantaneous, this production plan is (easily) feasible: by the end of period 1 there will be 250 units of product 1 in inventory; this level will remain constant until time 3 at which point the inventory of product 1 will be drawn down at a constant rate of 80 to 170 by the end of period 4. At the time grid points, the projected inventories of product i are $I^i_i(1) = 250$, $I^i_i(2) = I^i_i(3) = 250$, $I^i_i(4) = 170$, respectively.

Suppose processes are not instantaneous. Output of product i emerges after a constant lead time of 1.7 time units, and it takes a constant 1.8 time units for activity j to inspect product i before it can be used as input. In the notation of the constant lead time model, $\ell_i = 1.7$ and $\ell_{i,j} = 1.8$. Inventory of product i is sent to activity j just-in-time, and there are no other constraints on transfers. Consequently,

$$y_i(\tau) = z_i(\tau - 1.7) \text{ and } v_{i,j}(\tau) = 2z_j(\tau + 1.8),$$

and the inventory of product i at activity i at time t is

$$\begin{aligned} I^i_i(t) &= \int_0^t y_i(\tau)d\tau - \int_0^t v_{i,j}(\tau)d\tau \\ &= \int_0^t z_i(\tau - 1.7)d\tau - 2\int_0^t z_j(\tau + 1.8)d\tau \\ &= \int_0^{t-1.7} z_i(\tau)d\tau - 2\int_{1.8}^{t+1.8} z_j(\tau)d\tau \\ &= \int_0^{t-1.7} 100 \cdot 1_{[0,1)}(\cdot)d\tau - 2\int_{1.8}^{t+1.8} 40 \cdot 1_{[3,4)}(\cdot)d\tau. \end{aligned}$$

In this case, the inventories of product i at activity i at the time grid points are, respectively:

$$I_i^i(1) = \int_0^{-0.7} 250 \cdot 1_{[0,1)}(\cdot)d\tau - 2\int_{1.8}^{2.8} 40 \cdot 1_{[3,4)}(\cdot)d\tau = 0 - 0 = 0,$$
$$I_i^i(2) = \int_0^{0.3} 250 \cdot 1_{[0,1)}(\cdot)d\tau - 2\int_{1.8}^{3.8} 40 \cdot 1_{[3,4)}(\cdot)d\tau = 75 - 64 = 11,$$
$$I_i^i(3) = \int_0^{1.3} 250 \cdot 1_{[0,1)}(\cdot)d\tau - 2\int_{1.8}^{4.8} 40 \cdot 1_{[3,4)}(\cdot)d\tau = 250 - 80 = 170.$$

These inventories are all positive and all appears well. However, at the production level, the production plan simultaneously calls for an output schedule of product i and a transfer schedule of product i to activity j of

$$y_i(\tau) = 250 \cdot 1_{[1.7,2.7)}(\cdot) \text{ and } v_{i,j}(\tau) = 80 \cdot 1_{[1.2,2.2)}(\cdot).$$

Obviously, activity i cannot meet its required transfer schedule on the interval $[1.2, 1.7]$ because it will not have produced any output by this time! This production plan is most definitely not feasible, even though the projected inventories at the time grid points are all non-negative (as required).

All flow functions in the example, namely, $z_i(\tau)$, $y_i(\tau)$, $v_{i,j}(\tau)$, and $\hat{v}_{i,j}(\tau)$, are step functions. Consequently, their respective cumulative functions are all piecewise linear. A linear combination of piecewise linear functions is also piecewise linear. Consequently, the $FlowIn[0,t]$ and $FlowOut[0,t]$ functions in (21.1) are also piecewise linear, and thus $I(t)$ is piecewise linear, too. The following proposition provides necessary and sufficient conditions that ensure $I(t) \geq 0$ for all (relevant) time t.

Proposition 21.5. *Let $\mathcal{G} \subset [0,T]$ be a time grid and assume that $T \in \mathcal{G}$. Let $H(\cdot)$ be a piecewise linear function defined on $[0,T]$ whose points of non-differentiability (the "breakpoints") belong to $\mathcal{G}$. If $H(t_k) \geq 0$ for all $t_k \in \mathcal{G}$, then $H(t) \geq 0$ for all $t \in [0,\infty)$.*

Proof. Pick an arbitrary period $[t_{k-1}, t_k]$ and $t \in (t_{k-1}, t_k)$. Let s_k denote the constant slope of $H(\cdot)$ on the interval $[t_{k-1}, t_k]$. Since $H(\cdot)$ is piecewise linear, $H(t) = H(t_{k-1}) + s_k(t - t_{k-1})$. If $s_k \geq 0$, then clearly $H(t) \geq 0$, too. Suppose $s_k < 0$. Since $H(\cdot)$ is piecewise linear, it is also true that $H(t) = H(t_k) - s_k(t_k - t)$, which immediately implies that $H(t) \geq 0$ in this case. As t was chosen arbitrarily, the result follows. □

In light of Proposition 21.5, an explanation of the problem illustrated by Example 21.4 is simple. Let $\mathcal{G}_i^i$ denote the time grid whose points in time are used to ensure material balance of product i at activity i holds for all time. The points in time when $Z_i(\cdot)$ changes its slope belong to the set of positive integers, and so, by Proposition 21.5, this set belongs to $\mathcal{G}_i^i$. In addition, the points

in time when $Y_i(\cdot)$ changes its slope belong to the set $\{1.7, 2.7, 3.7, \ldots\}$, and the points it time when $V_{i,j}(\cdot)$ changes its slope belong to the set $\{1.2, 2.2, \ldots\}$. By Proposition 21.5, both of these sets of points must also be added to $\mathcal{G}_i^i$. With an expanded set $\mathcal{G}_i^i$, additional constraints are imposed, which renders the production plan of Example 21.4 infeasible.

Example 21.6. For the multi-event, lead time model, $t \in \mathcal{G}_i^i$ if and only if either t or $t - \ell_{i,r}^y$ belong to $\mathcal{G}$. Similarly, $t \in \mathcal{G}_i^j$ if and only if either t or $t + \ell_{j,i,r}$ belong to $\mathcal{G}$. One may use the derivation in Remark 18.15 on p. 320 to calculate all necessary integrals.

With respect to the service constraints

$$\sum_i a_i^s \left(\sum_r w_{i,r}^x z_i(\tau - \ell_{i,r}^x) \right) \leq \tilde{x}^s(\tau),$$

each side of this inequality is a step function. The points in time when the left-hand side function may change its rates are $\{\tau : \ \tau - \ell_{i,r}^x \in \mathcal{G}\}$ and the points in time when the right-hand side may change its rates belong to $\mathcal{G}$. Accordingly, the service constraints must be evaluated for all τ such that either τ or $\tau - \ell_{i,r}^x$ belong to $\mathcal{G}$.

The capacities of some services (e.g. labor) can be adjusted to meet the demand within a period of time. In such cases, it is only necessary to ensure that the total demand for the service does not exceed its supply within a period of time. (This relaxation of the detailed τ-constraints implicitly assumes there will be no excessively high spikes of demand within the period of time.) In this case, the service constraint for service s becomes: for each period k,

$$\int_{t_{k-1}}^{t_k} \sum_i a_i^s \left(\sum_r w_{i,r}^x z_i(\tau - \ell_{i,r}^x) \right) \leq \int_{t_{k-1}}^{t_k} \tilde{x}^s(\tau) d\tau := x_k^s.$$

The integrals on the left-hand side can be computed using the derivation in Remark 18.15 on p. 320. The right-hand side values are exogenously specified parameters.

Example 21.7. For the continuous lead time model, the cumulative flows, in general, will not be piecewise linear. Without further restrictions, material balance cannot be guaranteed at all points in time. Since the model will be run frequently, if computational time is an issue, then the time grid $\mathcal{G}$ should have more time grid points earlier in the planning horizon. For this model, one may use expressions of the form (18.19) on p. 325 to calculate the necessary integrals.

21.4 A Manufacturing Example

21.4.1 Production Process Description

A manufacturer produces several products fabricated from sheet metal. The production system consists of five production centers in series: stamping, drilling, assembly, finishing (painting) and packaging. The company operates one shift per day, seven days per week, 365 days per year. A shift equals eight hours. Additional information:

- A central storage facility receives all raw materials from the vendors and transports the materials to the different production centers. Limited space (in square feet) is available to store raw materials at central storage. The (key) raw materials (e.g. sheet metal) are supplied by three different vendors.
- Each center uses specialized equipment and skilled labor generic to its task. Each center also uses unskilled labor. The pool of unskilled labor can be used by any center at any time. There are a fixed number of machines at each center.
- Limited space (in square feet) is available to store inventory of completed semi-finished parts at each center.
- Stamped and drilled parts for any product can be subcontracted. The subcontractor ships the semi-finished parts to the manufacturer for assembly, finishing, and packaging. Either subcontractor has capacity to fulfill any requirements by the manufacturer.
- All parts drilled in-house are inspected. The inspection crew's capacity is sufficient to meet all demand. Statistical data are available on the percent of parts passing inspection. Failed parts are discarded.
- There are no transportation lead times between centers.
- The manufacturer currently employs a fixed number of full-time employees of each skill type and a number of unskilled laborers. Labor can be hired or dismissed only at the beginning of the planning horizon, which is 12 weeks.
- Daily demands for final packaged product over the planning horizon are pre-specified and must be met at all times.

21.4.2 Formulation

We formulate a dynamic model of production that can be used by management to plan short-term production levels at each center and subcontracting requirements.

A standard time grid $t = 1, 2, \ldots, T = 84$ is adopted. The length of each period is one day. Product flows across the production centers in series:

$$\text{stamp} \longrightarrow \text{drill} \longrightarrow \text{assembly} \longrightarrow \text{finish} \longrightarrow \text{package} \qquad (21.40)$$

In addition, there is a central storage activity C that stores all incoming materials from the three vendors, and transports all outbound materials to the different production centers. An instantaneous, index-based dynamic model of production will be used.

The decision variables, state variables, and parameters are as follows:

- *Decision variables.* There are two sets of decision variables: in-house production variables and transfers from the subcontractors. Let $z_{i,p,t}$ denote the constant rate of starts (in-house) of semi-finished product p at production center i = stamp, drill, assembly, finish, package in period t; let $v^p_{SD_j,A,t}$ denote the constant rate of stamped/drilled semi-finished product p sent by subcontractor $j = 1, 2$ to the assembly production center in period t; and let $v^m_{M_j,C,t}$ denote the constant rate of material m sent by vendor $j = 1, 2, 3$ to central storage in period t.
- *State variables.* Let $I^p_{i,t}$ denote the inventory of completed semi-finished product p at production center i at time t, and let $I^m_{C,t}$ denote the inventory of material m at central storage at time t.
- *Parameters.* There are three sets of parameters: technical coefficients, capacities, and final product requirements. (Given the short cycle times, no production lead time parameters will be necessary.)
  - Technical coefficients: Let $a^p_{i,q}$ denote the number of units of semi-finished product p required per unit of product q at production center i, $a^e_{i,p}$ denote the number of units of skilled labor type e required per unit of product p at production center i, $a^u_{i,p}$ denote the number of units of unskilled labor required per unit of product p at production center i, $a^m_{i,p}$ denote the number of units of raw material m required per unit of product p at production center i, $a^s_{i,p}$ denote the number of units of machine service s required per unit of product p at production center i, and let $\pi^p_{drill} \in [0, 1]$ denote the probability of an in-house drilled semi-finished product p passing inspection at the assembly production center.
  - Capacities: Let $\tilde{x}^e_t$ denote the pool of skilled labor type e available in period t, $\tilde{x}^u_t$ denote the pool of unskilled labor available in period t, $\tilde{x}^s_{i,t}$ denote the amount of machine service s available at production center i in period t, $\tilde{x}^m_t$ denote the available supply of raw material m available in period t, $\tilde{x}^{sq}_{i,t}$ denote the capacity in square feet to store semi-finished product at production center i in period t, let α^i_p denote the square feet required by semi-finished product p at production center i, and let α^m denote the square feet required by material m at central storage C.
  - Final product requirements: Let $\tilde{y}^p_t$ denote the number of units of final packaged product p required in period t.

The material balance, service, and storage capacity constraints are as follows:

- *Material balance constraints:* For each t,

$$I^p_{stamp,t} = I^p_{stamp,0} + \sum_{\tau=1}^{t} z^p_{stamp,\tau} - \sum_q a^p_{drill,q} \left(\sum_{\tau=1}^{t} z_{drill,q,\tau} \right) \geq 0,$$

$$I^p_{drill,\tau} = I^p_{drill,0} + \sum_{\tau=1}^{t} \pi^p_{drill} z^p_{drill,\tau} + \sum_{\tau=1}^{t} v^p_{SD_1,A,\tau} + \sum_{\tau=1}^{t} v^p_{SD_2,A,\tau} - \sum_q a^p_{assemble,q} \left(\sum_{\tau=1}^{t} z_{assemble,q,\tau} \right) \geq 0,$$

$$I^p_{assemble,t} = I^p_{assemble,0} + \sum_{\tau=1}^{t} z^p_{assemble,\tau} - \sum_q a^p_{finish,q} \left(\sum_{\tau=1}^{t} z_{finish,q,\tau} \right) \geq 0,$$

$$I^p_{finish,t} = I^p_{finish,0} + \sum_{\tau=1}^{t} z^p_{finish,\tau} - \sum_q a^p_{package,q} \left(\sum_{\tau=1}^{t} z_{package,q,\tau} \right) \geq 0,$$

$$I^p_{package,t} = I^p_{packge,0} + \sum_{\tau=1}^{t} z^p_{package,\tau} - \sum_{\tau=1}^{t} \tilde{y}^p_\tau \geq 0,$$

$$I^m_{C,t} = I^m_{C,0} + \sum_{\tau=1}^{t} \left(v^m_{M_1,C,\tau} + v^m_{M_2,C,\tau} + v^m_{M_3,C,\tau} \right) - \sum_{\tau=1}^{t} \left(\sum_i \sum_p a^m_{i,p} z_{i,p,\tau} \right) \geq 0.$$

- *Service capacity constraints:* For each t, production center i, and machine service s or labor type e,

$$\sum_p a^s_{i,p} z_{i,p,t} \leq \tilde{x}^s_{i,t},$$

$$\sum_p a^e_{i,p} z_{i,p,t} \leq \tilde{x}^e_{i,t},$$

and for each t,

$$\sum_i \sum_p a^u_{i,p} z_{i,p,t} \leq \tilde{x}^u_t.$$

- *Storage capacity constraints:* For each t, production center i or storage center C,

$$\sum_p \alpha^p_i I_{i,p,t} \leq \tilde{x}^{sq}_{i,t},$$

$$\sum_m \alpha^m I^m_{C,t} \leq \tilde{x}^{sq}_t.$$

21.4.3 Extensions

We extend the formulation to accommodate the following information:

- The constant shipping lead times for vendors 1, 2, and 3 are, respectively, 4, 5, and 10 days.
- The lead time for subcontractor one is a constant 2 days and the lead time for subcontractor 2 is 3 days.
- The inspection crew at the drilling production center is given one day to perform inspections.
- Regardless of where the parts were stamped and drilled, there is a one delay before assembly begins to process incoming receipts.
- It takes two days for the paint to dry in the finishing center.

A constant lead time model will be used. With respect to initial conditions, we shall adopt Approach I: the values of all variables defined below before time 0 are pre-specified parameters. Furthermore, we shall assume that all production variables after the current time 0 that required transfers prior to the current time 0 are frozen. See Remark 21.2, p. 398.

Only the material balance constraints for drilling, finishing, and central storage in the formulation above change. In this setting, these constraints become: for each t,

$$I^p_{drill,t} = I^p_{drill,0} + \sum_{\tau=1}^{t} \pi^p_{drill} z^p_{drill,\tau-1} + \sum_{\tau=1}^{t} v^p_{SD_1,A,\tau-2} + \sum_{\tau=1}^{t} v^p_{SD_2,A,\tau-3} - \sum_q a^p_{assemble,q} \left(\sum_{\tau=1}^{t} z_{assemble,q,\tau+1} \right) \geq 0,$$

$$I^p_{finish,t} = I^p_{finish,0} + \sum_{\tau=1}^{t} z^p_{finish,\tau-2} - \sum_q a^p_{package,q} \left(\sum_{\tau=1}^{t} z_{package,q,\tau} \right) \geq 0,$$

$$I^m_{C,t} = I^m_{C,0} + \sum_{\tau=1}^{t} \left(v^m_{M_1,C,\tau-4} + v^m_{M_2,C,\tau-5} + v^m_{M_3,C,\tau-10} \right) - \sum_{\tau=1}^{t} \left(\sum_i \sum_p a^m_{i,p} z_{i,p,\tau} \right) \geq 0.$$

Finally, we extend the previous formulation to accommodate the following information:

- The lead times (for each material m) for vendor 1 follow this distribution: materials arrive 20% of the time in 3 days, 50% of the time in 4 days, and 30% of the time in 5 days.
- The lead times (for each material m) for vendor 2 follow this distribution: materials arrive 65% of the time in 5 days and 35% of the time in 10 days.

- The lead times (for each material m) for vendor 3 follow this distribution: materials arrive 25% of the time in 5 days, 60% of the time in 10 days, and 15% of the time in 20 days.

Only the central storage constraint in the above formulation changes. Replace the left-hand side below with the right-hand side:

$$\begin{aligned} v^m_{M_1,C,\tau-4} &\longleftarrow 0.20v^m_{M_1,C,\tau-3} + 0.50v^m_{M_1,C,\tau-4} + 0.30v^m_{M_1,C,\tau-5}, \\ v^m_{M_2,C,\tau-4} &\longleftarrow 0.65v^m_{M_2,C,\tau-5} + 0.35v^m_{M_2,C,\tau-10}, \\ v^m_{M_3,C,\tau-4} &\longleftarrow 0.25v^m_{M_3,C,\tau-5} + 0.60v^m_{M_3,C,\tau-10} + 0.15v^m_{M_3,C,\tau-20}. \end{aligned}$$

Remark 21.8. The new material balance constraints include *expected* transfers of materials from the vendors. It does not account for risk. A more conservative approach is to hedge the empirical distributions. For example, with respect to vendor 1,

$$v^m_{M_1,C,\tau-4} \longleftarrow 0.10v^m_{M_1,C,\tau-3} + 0.40v^m_{M_1,C,\tau-4} + 0.50v^m_{M_1,C,\tau-5}$$

could be a suitable choice. The most conservative approach is to use the *maximum* lead time, e.g.,

$$v^m_{M_1,C,\tau-4} \longleftarrow v^m_{M_1,C,\tau-5}.$$

This approach maintains maximum internal customer service at the expense of possibly too much inventory. Since the maximum lead time for vendor 3 is significant in relation to the average lead time, the most conservative approach can be too costly.

21.5 Assembly with Rework Example

21.5.1 Production Process Description

A manufacturer assembles several hundred finished products in its final assembly plant. The plant normally operates two 8-hour shifts per day, five days per week. A third shift and/or weekend operations are possible, if required. A high-level planning system determines monthly target requirements for each final product. For some products, the monthly target requirements are broken down even further within the month. Within the plant there are four stages of production: initial assembly, test, rework, and final assembly. Additional information:

- Initial assembly consists of attaching basic components. The initial assembly area consists of up to four identical assembly lines of two-three workstations. Assembly cycle time ranges from 4.0 to 8.0 minutes. After a product has been initially assembled, it awaits to be tested. There is a limited capacity to store assembled products.

- When a product is scheduled for testing, it is routed via conveyor to an Automated Storage and Retrieval System (AS/RS) for testing. The capacity of the AS/RS is several thousand lots. Testing consists of conducting a sequence of software programs, some of which are run several times. Testing does not involve labor resources. The time to complete a successful test depends on the type of product, and ranges from a few hours up to 48 hours. (For certain products, there is some degree of flexibility with respect to the length of test.) Testing occurs continuously, that is, a test is not stopped when the plant is not operating, e.g., on third shifts, weekends, holidays, scheduled downtime. The exact times when each product begins and fails each test (if it does) are recorded; all such records are stored in a database. If a product fails its test, it awaits rework.
- The rework area consists of individual workstations where each worker diagnoses and executes the required rework, mainly consisting of reassembling or inserting new components. Rework cycle time ranges from 5.0 to 10.0 minutes. When rework has been completed, the product awaits to be tested. There is a limited capacity to store reworked products. Rework does not change the probability distribution of success or failure over time.
- The final assembly area consists of identical workstations, where each worker attaches final components and packages the product. Final assembly cycle time ranges from 3.0 to 5.0 minutes.

21.5.2 Formulation

We formulate a dynamic model of production that can be used by the manufacturer to plan short-term production levels by shift for the initial, final, and rework areas to meet end-of-month target requirements for finished products. In addition, it can assist senior-level management to answer the following operational, tactical and strategic questions:

- Should we increase manufacturing productivity by reducing test lead time? What will be the cost of customer dissatisfaction and potential future rework?
- Should we purchase better components?
- Should we increase the AS/RS capacity?

There are four production activities: initial assembly, test, rework, and final assembly. The outputs of the test facility are *both* pass and failed parts, and separate inventories must be counted to ensure that rework only works on failed parts and final assembly only works on parts that have passed the test. A standard time grid $t = 1, 2, \ldots, T$ is adopted. The length of each time period is a shift, only those shifts that are operational are represented, and T represents the number of operational shifts remaining in the planning horizon.

With respect to the starts variables, let

- $z_{initial,p,\tau}$ denote starts in initial assembly of product p in period τ,

- $z_{rework,p,\tau}$ denote starts in rework of product p in period τ, and
- $z_{final,p,\tau}$ denote starts in final assembly of product p in period τ.

There are no starts variables in the test activity, as all transfers into the test area immediately begin their respective tests.

With respect to the inventory state variables, let

- $I^p_{initial,t}$ denote the inventory of initial assembled part p (awaiting testing) at time t,
- $I^{(p,fail)}_{test,t}$ denote the inventory of part p that has failed its test (awaiting rework) at time t,
- $I^{(p,pass)}_{test,t}$ denote the inventory of part p that has passed its test (awaiting final assembly) at time t, and
- $I^p_{final,t}$ denote the inventory of finished part p (awaiting delivery) at time t.

A continuous-time lead time model is adopted for the test activity. Let

- $\pi^{(p,pass)}$ denote the probability of part p passing each test,
- ℓ_p denote the length of time, not necessarily integer, of the test for product p, and
- $W^p(t,\tau)$ denote probability that part p that began test at time τ will have failed by time $\tau + t$. Note that $W^p(\ell_p,\tau) = 1 - \pi^{(p,pass)}$.

Remark 21.9. In this setting, the cumulative lead time distribution $W(\cdot,\tau)$ is dependent on τ. For example, consider a τ corresponding to second shift on Fridays, and suppose $\tau+1$ corresponds to the first shift the following Monday. That is, Friday's third shift and the weekend shifts are not operational. Even if the test took say 6 shifts (48 hours), here $W^p(2,\tau) = 1$.

Let

$$\Pi^p_{t-\tau,\tau} = \int_{\tau-1}^{\tau} W^p(t-u,u)du$$

denote the proportion of test starts of product p in period τ that will fail by time t. The material balance constraints are as follows: For each t,

$$\begin{aligned}
I^p_{IA,t} &= I^p_{IA,0} + \sum_{\tau=1}^{t} z_{initial,p,\tau} - \sum_{\tau=1}^{t} v^p_{initial,test,\tau} \geq 0,\\
I^{(p,fail)}_{test,t} &= I^{(p,fail)}_{test,0} + \sum_{\tau\leq t} \Pi^p_{t-\tau,\tau} v^p_{initial,test,\tau}\\
&\quad + \sum_{\tau\leq t} \Pi^p_{t-\tau,\tau} v^p_{rework,test,\tau} - \sum_{\tau=1}^{t} v^p_{test,rework,\tau} \geq 0,\\
I^{(p,pass)}_{test,t} &= I^{(p,pass)}_{test,0} + \sum_{\tau\leq t} \pi^{(pass,p)} v^p_{initial,test,\tau-\ell_p}
\end{aligned}$$

$$+\sum_{\tau\le t}\pi^{(p,pass)}v^p_{rework,test,\tau-\ell_p}-\sum_{\tau=1}^{t}v^p_{test,final,\tau}\ge 0,$$

$$I^p_{rework,t}=I^p_{rework,0}+\sum_{\tau=1}^{t}v^p_{test,rework,\tau}-\sum_{\tau=1}^{t}v^p_{rework,test,\tau}\ge 0,$$

$$I^p_{test,t}=I^p_{test,0}+\sum_{\tau=1}^{t}v^p_{initial,test,\tau}+\sum_{\tau=1}^{t}v^p_{rework,test,\tau}$$
$$-\sum_{\tau=1}^{t}v^p_{test,rework,\tau}-\sum_{\tau=1}^{t}v^p_{test,final,\tau}\ge 0,$$

$$I^p_{final,\tau}=I^p_{final,0}+\sum_{\tau=1}^{t}z_{final,p,\tau}-\sum_{\tau=1}^{t}\tilde{y}^p_\tau\ge 0.$$

The raw material, service, and storage capacity constraints are similar to the previous example and will be left to the reader.

Remark 21.10. A simpler model to formulate is to include *all* shifts, operational or not. The $z_{\cdot,\cdot,\tau}$ variables must be set to zero for those non-operational periods τ. The benefit of this approach is that the Π parameters do not depend on τ, which can reduce implementation error. The disadvantage of this approach is that it includes additional material balance constraints that are not needed.

21.5.3 Extensions

Suppose the period lengths are taken to equal one day? Let $s(\cdot)$, $\tau\in[0,1]$, denote the distribution of starts in each period (day) t. (See Section 18.4, p. 324.) For example, if $s(\tau)=1.5$, $\tau\in[0,2/3]$, $s(\tau)=0$, $\tau\in(2/3,1]$, then the production rate within each period is constant over the first two shifts (i.e., two-thirds of each day). Without further information, this is a natural choice for $s(\cdot)$. A more general form is $s(\cdot,t)$, which permits this distribution to depend on the day. In this case, the Π parameters become

$$\Pi_{t-\tau,\tau}=\int_0^1 s(u)W(t-u,u)du.$$

21.6 Extensions to the Basic Model

21.6.1 Material Balance Constraints

The material balance constraints (21.2)-(21.5) of the basic model can be extended to accommodate the following phenomena (and more):

- Activities may produce several products.

- Several activities may produce the same product.
- Activities may transfer product they do not produce.
- Activities may receive product they do not use.
- Intermediate or final products produced by the system can be obtained externally (e.g., via a subcontractor).

The more flexible material balance constraints are as follows. For each activity $i = 1, 2, \ldots, N+1$, product or material k, and for all $t \geq 0$:

$$0 \leq I_i^k(t) = \overbrace{I_i^k(0)}^{\text{Initial inventory}} + \overbrace{\sum_{j=0}^{N} \int_0^t \hat{v}_{j,i}^k(\tau) d\tau}^{\text{Transfers In}} + \overbrace{\int_0^t y_i^k(\tau) d\tau}^{\text{Production}} - \overbrace{\sum_{j=1}^{N+1} \int_0^t v_{i,j}^k(\tau) d\tau}^{\text{Transfers Out}} - \overbrace{\int_0^t x_i^k(\tau) d\tau}^{\text{Usage}} - \overbrace{\int_0^t \tilde{y}^k(\tau) d\tau}^{\text{Final demand}} . \quad (21.41)$$

In (21.41), it is understood that when $i = N+1$, the $y_{N+1}^k(\cdot)$, $v_{N+1,j}^k(\cdot)$, and $x_{N+1}^k(\cdot)$ are all zero.

Remark 21.11. Realff et. al. [2004] use an example of this general type of material balance constraint in their "reverse production system" design model for carpet recycling.

21.6.2 Transfers of Product or Materials

The material balance constraints (21.41) can be further extended to accommodate the following phenomena:

- There can be several mechanisms (e.g., alternate shipping modes or routes) to transfer product.
- There can be several sources (e.g. vendors) to acquire raw materials.

Let r denote a generic transfer mechanism or exogenous supply source. Let $v_{i,j,r}^k(\tau)$ denote the transfer of product or material k sent by activity i to activity j via transfer mechanism r at time τ, and let $\hat{v}_{i,j,r}^k(\tau) = t_{i,j,r}^k[v_{i,j,r}](\tau)$ denote the transfer of product or material k received by activity i to activity j via transfer mechanism r. In (21.41), one replaces $v_{i,j}^k(\tau)$ and $\hat{v}_{i,j}^k(\tau)$ with $v_{i,j,r}^k(\tau)$ and $\hat{v}_{i,j,r}^k(\tau)$, respectively, and sums over both j and r.

21.6.3 Activity Constraints

Practical models include many specialized constraints depending on the application. For example:

- There can be limited capacity to store products.

- A machine may only be able to produce one product at a time.
- There can be a setup time to prepare a machine for production (e.g., clean, test) or a changeover time to prepare a machine for the next product's production run (e.g., reconfigure machine settings).

One simple model to accommodate the first constraint is

$$\sum_k \gamma_i^k I_i^k(\tau) \leq \tilde{x}^{\gamma_i}(\tau). \tag{21.42}$$

In (21.42), the parameters γ_i^k represent conversion factors that translate units of inventory into units of storage (e.g., square feet, cubic feet). The left-hand side represents the usage of the service input 'γ_i' at time τ, and the right-hand side $\tilde{x}^{\gamma_i}(\tau)$ represents the available service capacity of input γ_i at time τ.

The second constraint is an example of a machine scheduling constraint, appropriate from the shop floor perspective. Both this constraint and the third one can be represented in a mathematical programming formulation by introducing (binary) logical variables and constraints (see, for example, Wolsey and Nemhauser [1999]). Introducing such variables and constraints can make the problem difficult to solve. If setup and changeover times are not too significant, this class of physical phenomena can be ignored in the model of technology. Alternatively, a projection of the amount of setup/changeover time per period can be deducted from the available capacity of this resource.

21.6.4 Service Output

An activity may produce a service s that cannot be stored. This can be accommodated by adding constraints of the form

$$\sum_j v_{i,j}^s(\tau) \leq y_i^s(\tau), \tag{21.43}$$

and

$$\sum_j \hat{v}_{j,i}^s(\tau) \geq x_i^s(\tau). \tag{21.44}$$

In (21.43), the $v_{i,j}(\tau)$ represent an *allocation* of the service s output of activity i to the other activities at time τ. The $\hat{v}_{j,i}(\tau)$ permit a possible time lag in the receipt of this service.

21.6.5 Alternate Production Processes

Often several processes can be used to manufacture a product within a work center or plant. Furthermore, how this product is manufactured can be irrelevant to final customers or to activities that use this product as intermediate product input. In this case, product inventory does not distinguish the source

of production. To properly account for product inventory, all output produced by the different processes must be aggregated, and to properly account for resource use, all resources consumed by the separate processes must be aggregated, too.

To be concrete, suppose there are N alternate processes to produce product i. Let $x_{i_n}(\cdot)$ and $y_{i_n}(\cdot) = [f_{i_n}(x_{i_n})](\cdot)$ denote, respectively, the input vector and output of product i produced by process $n = 1, 2, \ldots, N$. Let activity i be identified with production of product i by all processes. Let $x_i^k(\tau)$ denote the aggregate input of resource k by all processes, and let $y_i(\tau)$ denote the aggregate output produced by all processes. In the material balance constraint for product i, $\sum_{n=1}^{N} y_{i_n}(\tau)$ replaces $y_i(\tau)$, and in the resource constraint associated with resource k, $\sum_{n=1}^{N} x_{i_n}^k(\tau)$ replaces $x_i^k(\tau)$.

Remark 21.12. Aggregating output, as described above, can be modeled by introducing a *consolidation* activity, as follows. Activity i_n, $n = 1, 2, \ldots, N$, produces product 'i_n'. There is no inventory of product i_n. This activity's input-output process is characterized by the dynamic production function $y_{i_n}(\cdot) = [f_{i_n}(x_{i_n})](\cdot)$. This activity's output is immediately transferred to and instantaneously received by activity i, i.e.,

$$y_{i_n}(\cdot) = v_{i_n,i}(\cdot) = \hat{v}_{i_n,i}(\cdot).$$

Activity i uses no service resource to produce its output; rather, it instantaneously consolidates all intermediate product input, i.e.,

$$y_i(\cdot) = \sum_{n=1}^{N} \hat{v}_{i_n,i}(\cdot).$$

The constraints associated with the basic model or its extensions can now be applied without further *ad-hoc* adjustments.

21.6.6 Load-Dependent, Multi-Product, Single-Stage Model

In Section 19.1.4, we described a practical way to represent a load-dependent single-input, single-output process. Here, we show how to extend this formulation when there are multiple products flowing through a single stage. (The extension involving multiple stages is relatively straightforward.) We shall assume a discrete-time approximation using a standard time grid.

Let $z_{i,\tau}$ and $y_{i,\tau}$ denote, respectively, the starts and output of product i in period τ. The work queue of product i at time τ is

$$q_{i,\tau} := q_{i,\tau-1} + z_{i,\tau} - y_{i,\tau}. \tag{21.45}$$

The *stage work queue* at time τ is a weighted sum of the product work queues, and is defined as

$$q_\tau := \sum_i \beta_{i,\tau} q_{i,\tau}. \tag{21.46}$$

The weights $\beta_{i,\tau}$ are pre-specified and are chosen to reflect the relative importance of each product on the aggregate workload at this stage. Let $y_{i,\tau,t}$ denote the output of product i in period t due to the input in period $\tau \le t$, and let $y_{i,t}$ denote the output in period t as a result of all past input including its initial queue, $q_{0,i}$. Let

$$0 := q^0 < q^1 < q^2 < \cdots < q^L$$

denote pre-specified *stage queue levels*. (The last level, q^L, is sufficiently high to bound above any realized queue size.) For each τ, t such that $\tau \le t$, the new model for output for product i is

$$y_{i,\tau,t} := z_{i,\tau} \Pi^\ell_{\tau,t}, \quad \text{if } q(\tau) \in [q^{\ell-1}, q^\ell), \tag{21.47}$$

$$y_{i,t} := \sum_{\tau \le t} y_{i,\tau,t} \; + \; q_{0,i} \Pi_{0,t}. \tag{21.48}$$

The parameters $\Pi_{0,t}$ define the initial lead time distribution, which is a function of the initial queue q_0. It remains to implement the logical expression (21.47). This can be achieved with the use of *binary* variables and adding two sets of linear constraints, as follows.

- *Output constraints.* For each τ, t such that $\tau \le t$, and each $\ell = 1, 2, \ldots, L$, add these two constraints:

$$y_{i,\tau,t} \le z_{i,\tau} \Pi^\ell_{\tau,t} \; + \; M(1 - \xi_{\tau,\ell}), \tag{21.49}$$

$$y_{i,\tau,t} \ge z_{i,\tau} \Pi^\ell_{\tau,t} \; - \; M(1 - \xi_{\tau,\ell}), \tag{21.50}$$

  where M is a sufficiently large number, and

$$\xi_{\tau,\ell} := \begin{cases} 1, & \text{if } q(\tau) \in [q^{\ell-1}, q^\ell), \\ 0, & \text{otherwise.} \end{cases} \tag{21.51}$$

  If $\xi_{\tau,\ell} = 0$, then (21.49) and (21.50) constrain $y_{i,\tau,t}$ to lie in the interval $(-M, M)$, which, to all intents and purposes, is equivalent to $(-\infty, \infty)$. In this case, the constraints (21.49) and (21.49) are automatically satisfied. On the other hand, if $\xi_{\tau,\ell} = 1$, then (21.49) and (21.50) constrain $y_{i,\tau,t}$ to lie both above and below $z_{i,\tau} \Pi^\ell_{\tau,t}$. Obviously, this can only happen if $y_{i,\tau,t} = z_{i,\tau} \Pi^\ell_{\tau,t}$.
- *Queue constraints.* The logical variables $\xi_{\tau,\ell}$ must be linked to the queue values, the q_τ, so as to be consistent with their intended purpose.[4] In conjunction with the logical constraints

[4] These constraints are identical to the ones presented in Section 19.1.4, but we reproduce them here for completeness.

$$\sum_{\ell=1}^{L} \xi_{\tau,\ell} = 1, \quad \text{for each } \tau, \tag{21.52}$$

this is accomplished by adding two new constraints for each time period τ:

$$q_\tau \leq \sum_{\ell=1}^{L} \xi_{\tau,\ell} q^\ell, \tag{21.53}$$

$$q_\tau \geq \sum_{\ell=1}^{L} \xi_{\tau,\ell} q^{\ell-1}. \tag{21.54}$$

Suppose $q_\tau \in (q^{\hat{\ell}-1}, q^{\hat{\ell}})$. Since the $\xi_{\tau,\ell}$ are binary variables, for each τ, (21.45) implies there will be exactly one index $\ell(\tau)$ such that $\xi_{\tau,\ell(\tau)} = 1$. To satisfy constraints (21.53) and (21.54), the only choice for $\ell(\tau)$ is $\hat{\ell}$.

With respect to material balance constraints, for each product i and each period t,

$$I_{i,t} = I_{i,0} + \sum_{\tau \leq t} y_{i,\tau} - \sum_{\tau \leq t} v_{i,N+1,\tau}. \tag{21.55}$$

In sum, in addition to the standard service constraints, the proposed load-dependent, multi-product, single-stage continuous lead time model uses constraints (21.45) and (21.49)-(21.55).

Remark 21.13. Modeling the lead time distribution as a function of system load is a nontrivial task. A practical way to capture this phenomena is via an iterative, *simulation-optimization* approach. The basic idea is as follows. The first step uses shop-floor statistics to estimate an *initial* candidate lead time distribution, $W^0(\cdot,\cdot)$. Next, the model of technology is optimized to produce a production plan of starts, z^1, for the activities. This production plan is input to a simulation model that estimates the new candidate lead time distribution, $W^1(\cdot,\cdot)$. This new distribution is used to model of technology, another optimization is undertaken to yield a new production plan, z^2, which leads to $W^2(\cdot,\cdot)$ and so forth. There are two conceptual problems:

- *Lack of convergence.* There are no guarantees that this iterative process will converge to a limiting distribution $W_\infty(\cdot,\cdot)$.
- *Lack of uniqueness.* Even if a limiting distribution is guaranteed to exist, it need not be unique. That is, the limit distribution may depend on the initial candidate distribution.

Problems aside, this approach can be very effective for simultaneously representing and optimizing a technology. See Hung and Leachman [1996] for an example of how to successfully use this approach for production planning in the semiconductor industry, an extremely sophisticated and challenging manufacturing environment.

21.7 Efficiency and Productivity Measurement

We have described a number of descriptions for technology $\mathcal{T} = \{(\tilde{x}, \tilde{y})\}$ involving a network of interrelated activities. Given $\mathcal{T}$, in principle, one may undertake efficiency and productivity measurement, as described in Parts I and II.

21.7.1 Input and Output Efficiency

The radial measure of input and output efficiency use, respectively, the input distance function $\mathcal{D}^{\mathcal{T}}(\tilde{x}, \tilde{y})$ and output distance function $\mathcal{O}^{\mathcal{T}}(\tilde{x}, \tilde{y})$. In this setting, the distance functions are identical to their counterparts previously defined in Part I, except that here their arguments are now vector-valued functions of time. Since inputs and output occur over time, Russell-type measures of distance (with time-varying weights) should also be investigated.

21.7.2 Cost and Allocative Efficiency

For a given set of (time-varying) prices, the traditional cost function

$$Q(\tilde{y}, \tilde{p}) = \min \Big\{ \sum_{i=1}^{n} \int_0^T p_i(\tau)\tilde{x}_i(\tau)d\tau : \ (\tilde{x}, \tilde{y}) \in \mathcal{T} \Big\}$$

can be computed, which can be used to assess cost and allocative efficiency, as described in Parts I and II.

When the production process involves a network of activities that generate intermediate products (i.e., work-in-process), the cost of holding inventory

$$\sum_{i,p} \int_0^T c_{i,p}(\tau) I_i^p(\tau) d\tau$$

is normally added. The constant $c_{i,p}(\tau)$ incorporates the value of product p at activity i and the "time value of money."

For final products (or finished goods inventory), sometimes the customer will accept product later, in which case the customer order is on *backorder*. Missing customer due dates can be expensive (e.g., loss of goodwill that reduces future demands), and can represent a different cost than the traditional inventory holding cost. To differentiate these costs and to permit backorders, final product inventories at activity $N + 1$ are permitted to be *negative*. In the cost function, one replaces

$$\sum_{p} \int_0^T c_{N+1,p}(\tau) I_{N+1}^p(\tau) d\tau$$

with

$$\sum_p \int_0^T c_{N+1,p}(\tau) \max\{I^p_{N+1}(\tau), 0\} d\tau - \sum_p \int_0^T b_{N+1,p}(\tau) \min\{I^p_{N+1}(\tau), 0\} d\tau.$$

The constant $b_{N+1,p}(\tau)$ incorporates the cost of not meeting customer demand of product p and the "time value of money."

21.7.3 Productivity Assessment

The distance functions $\mathcal{D}^{\mathcal{T}}(\tilde{x}, \tilde{y})$ and $\mathcal{O}^{\mathcal{T}}(\tilde{x}, \tilde{y})$ may be used to calculate technical and efficiency change, as described in Part II.

21.7.4 Computation

All discrete-time approximations of technology described in this chapter yield a set of linear inequalities in the core index variables. Coupled with an appropriate objective function, the measures of efficiency and productivity listed above can be computed via mathematical programming software. For example, if the objective functions are all linear and there are no complicating constraints that require integer variables, linear programming can be used; if the objective functions are all convex (concave) and there are no complicating constraints that require integer variables, convex (concave) programming can be used; finally, if integer variables are needed, then integer programming optimization software must be used, or a heuristic algorithm must be developed or applied. In the last case, it may not be possible to guarantee an optimal solution in reasonable amount of time. However, a provably good solution can often be obtained in a satisfactory amount of time.

21.8 Bibliographical Notes

Dynamic models using activity analysis, coined *Dynamic Linear Activity Analysis Models*, originates with Shephard et. al. [1977]. The treatment here extends and refines the earlier work of Hackman [1990] and Hackman and Leachman [1989].

Johnson and Montgomery's [1974] provide concrete textbook examples of dynamic, multi-stage models of production described in this chapter. The handbook of Graves et. al. [1993] contains several chapters describing a variety of production planning models. See also de Kok and Graves [2003] and Simchi-Levi et. al. [2002] for models appropriate to supply chains.

Leachman et. al. [1996] and Leachman [2002] describe extensive, in-depth, award-winning models for production planning currently in use by the semiconductor industry.

Voss and Woodruff [2006] provide an extensive discussion of practical considerations associated with production planning, including a thorough treatment of Manufacturing Resources Planning (MRP), as well as a detailed formulation of a computational model for production planning involving load-dependent, lead times.

A different approach to modeling lead time as a function of system load uses the concept of a *clearing function*, first introduced in Graves [1986]. A clearing function specifies the fraction of the work-in-process that can be completed (or "cleared") by a resource in a time period. Pahl et. al. [2007] provide a thorough review of the state-of-the art production planning models with load-dependent lead times, including relevant references pertaining to clearing functions.

Riano [2002] describes an iterative scheme to obtain the load dependent lead time distributions in a multi-stage system that contains explicit approximations to the underlying queue process.

Reveliotis [2004] develops a new control paradigm for the real-time management of resource allocation systems.

22

Optimizing Labor Resources Within a Warehouse

We apply the ideas presented in Part IV to develop a dynamic, multi-stage model of warehouse operations to optimize labor resources.

22.1 Introduction

In today's marketplace, warehouse systems store thousands of stock keeping units (skus), have tight delivery times and demands for value-added services. In lieu of simply throwing more money at the problem, competitive pressures force warehouse managers to squeeze the most out of their labor, capital, space, and information to meet these increased demands.

A retrofit of an existing facility (or a new design) begins by forecasting the short-, medium- and long-term requirements, which are largely dictated by the number of skus, their different sizes, the number of lines, customer order patterns, and customer service requirements. A design consists of two interrelated components: a *physical* design—the configuration of labor, capital and space; and a *material flow* design—the inbound and outbound processes, namely, how product is received (unloaded, staged, transported to storage) and shipped (picked, packed, and loaded). For a given set of requirements, there are many possible designs due to the myriad of substitution possibilities among resource categories. Examples include:

- *Information vs. Labor.* Use of bar-codes, light aids, or a paperless Warehouse Management System (WMS) to reduce labor requirements.
- *Capital vs. Labor.* Use of a semi-automated or automated picking system (e.g., carousel, miniload, or A-frame) to replace a walk-and-pick system. Use of conveyor and sortation equipment to replace manual assembly.
- *Capital vs. Space.* Use of narrow or very narrow aisle trucks, overhead conveyor, mezzanines, person-aboard AS/RS, or facility expansion to create more space.

The most economical choice depends in large part on the availability and cost of land and labor, the type of inventory, and the customer service time window. Each of the above options requires a significant expenditure, so any recommendation for change should be economically justified relative to the best use of the current design.

There are a number of practical ways to significantly improve a current system *without* a major capital expenditure. Each is designed to reduce cost by simply taking work out of the system. In this chapter, we develop a dynamic model of a warehouse system for the purpose of *labor staffing and workforce scheduling*. We shall concentrate on the picking/shipping process, as it represents the largest share of the operations cost. We develop an optimization model that will determine the various times personnel (pickers and packers) report to work throughout the day, and how to strategically use overtime and part-time staff. By better matching workers to the *timing* of work requirements, significant reductions in both the number of workers and overtime will be achieved. As a by-product, the model suggests *order release guidelines* that will improve labor efficiency and ease demands for space by reducing unnecessary work-in-process.

22.2 System Description

We begin by describing the class of warehouse systems we have in mind, and the sources of inefficiencies that commonly occur. Our generic system covers the key aspects found in conventional facilities, but is concrete enough to illustrate the optimization-based approach.

22.2.1 Business Environment

Our example facility ships a large variety of product to small businesses via parcel post and to large corporate clients via trailers. The facility houses over 25,000 sku's and ships about two million lines per year. Product is stored in shelving, case flow rack, or pallet rack, and a walk-and-pick system is used for order picking. The average lines per order is 4.0 with 40% of all orders being single-line but with many orders involve many lines. Broken case picking represents 75% of all lines with the remainder being full case lines.

The business has seen a tremendous growth in sku's. The facility, which at one time had ample space, is now cramped. During peak times, there is a large amount of work-in-process, which has slowed workers down and increased order cycle times. Since expansion of the facility is not possible, management would like to increase the use of space, which they feel would increase labor productivity. Management also feels they can increase business if they could extend the order cutoff time from 6:00pm to 7:00-8:00pm, but this would require a reduction in the order cycle time. (Any order, if received by 6:00pm, will be processed and shipped for next day delivery.) Finally, there has been a

significant increase in overtime use, which management would obviously like to reduce.

22.2.2 Material Flow

Approximately 20 trailers depart the same time each day. Each trailer follows the same route, and each stop corresponds to a corporate customer. Trailer departure times are set primarily according to route distances in that the trailers with the longest routes leave first. Some departure times are set to satisfy the customer's receiving time window. There are three "route waves" of trailer departures centered around the times 7:30pm, 10:00pm, and 2:00am.

There are three distinct zones that respectively fulfill the broken case, full case, and full pallet picking requirements. As orders arrive into the system, the WMS prints and routes the suborders to each zone. Order batching occurs within the broken and full case zones. Order pickers manually assemble a batch of orders to roughly correspond to an hour's worth of work. Orders are prioritized according to their route wave. Lines are picked to totes, and partitions are used to maintain order integrity. When a batch has been picked, bar-code labels are affixed, and the totes are then dropped onto conveyors that route them to their appropriate check/pack station. All packers check product to ensure accuracy. Some packers pack product into cases; others must use a shrink-wrap machine. Packers stage completed product onto pallets that await loading onto trailers.

22.2.3 Workforce Schedule

Most order pickers and all packers report at 3:30pm. Some order pickers report at noon to build up a queue of work in front of the packers. Full-time workers are scheduled for 40-hr work weeks and receive full benefits. There are a few part-timer workers who work less than 25 hours per week and who do not receive full benefits. Labor markets are tight, but management feels that they can hire part-time workers. Since a part-time worker hour is cheaper than a full time worker hour, management would like to hire part-timers, but it does not know the best times during the day or week for them to be scheduled.

22.2.4 Sources of Inefficiency

If

- all order-lines arrived early enough in the day,
- if there were enough staging capacity, and
- if all order-lines arrived smoothly throughout the day with little variability from day-to-day,

then it would be possible to efficiently assign workers to the pick and pack stages to smooth material flow, and to build the pallets for each route and stage them well in advance of loading. Scheduling the workforce would be relatively easy.

Unfortunately, this is not the case. Approximately 25% of all lines arrive early (before 2:00pm), but the majority (over 50%) arrive in the *last* two hours. Variability of demand within each day and each week is reasonably high. Within each week Mondays and Fridays are the light-demand days. There is one highly seasonal period (December-January).

In recognition of the significant workload that must be completed each evening by 7:30pm (the first route wave), the second shift begins at the usual time of 3:30pm. Due to the high variability of demand, and since a worker-hour cannot be inventoried, the number of full time workers is hired to meet peak demand. Normal labor capacity is high on days when demand is low, which results in a significant amount of idle time. Normal labor capacity is insufficient on days when demand is high, which results in a significant amount of overtime to complete the workload associated with the final route wave. (Some labor inefficiency is inevitable with an inflexible workforce schedule.)

Order pickers naturally first pick those orders associated with the first route. During the first few hours of the second shift, however, it is often the case that the arrival queue is empty for those orders. Since order pickers are measured against strict productivity standards, they do not wish to be idle, so they continue to pick orders for the second and third route waves, if necessary. Essentially, then, the facility operates an implicit *push order release policy.* This causes a high degree of work-in-process that congests the system, as vast numbers of totes are staged on conveyor or even on the floor. Labor inefficiency results, as packers must search for the totes that belong to the early route waves.

Some work-in-process is inevitable, since a strict pull system will not work for the following reasons. With so many sku's, the facility's footprint is large, which necessitates order batching to achieve picking efficiency. On the shipping side, all orders for all customers on a route must be accumulated before the trailer may depart. Reducing the batch size will lower work-in-process but, while helpful, it does not attack a root problem. As is well-known in manufacturing, an imbalance among production rates along a production line will either result in too much work-in-process or possible starvation in front of a stage. For a warehouse system, production rates for pick and pack stages are determined by the amount of labor assigned. To ensure that the packing stage will never be starved, there is a natural tendency to assign too many workers to picking "to be on the safe side."

Labor staffing determines aggregate capacity. Workforce scheduling determines the *flow* of labor, namely, who should work *where* and *when.* To make it work, however, an order release policy must be determined that will strategically build the *right* work-in-process, which is possible since a signif-

icant number of lines do arrive early enough. The next section develops an optimization model to achieve this objective.

22.3 An Optimization Model

Conceptually, the process flow of a warehouse system is analogous to the process flow of a manufacturing system. That is, raw material is processed through a variety of value-added production stages, building up work-in-process that eventually is transformed into final output for use by the end customer. There are three main production or processing stages: picking, packing, and loading. The picking stage takes the raw material "customer-lines" that are "stored" in the Warehouse Management System (WMS) and transforms them into physical product stored in totes. The packing process transforms the totes into packaged quantities stored on pallets that await loading onto trailers according to the route schedule. The loading process transforms the pallets that are staged on the floor onto the trailers for shipment. There is work-in-process between the pick and pack stages (the totes) and between the pack and load stages (the pallets).

The optimization model we develop below takes the queue of line arrivals as *known* for each day $d \in \mathsf{D}$. It determines the amount and flow of workers, which defines the labor capacities. It determines which lines should be processed when (the production schedule) to meet the due dates (the trailer departure times). The production schedule will be consistent with labor capacities and space constraints. The objective of the optimization is to meet the requirements at minimum labor cost.

The set D could correspond to the actual line arrivals for every day in a past week, month, season, etc. We shall assume that D has been partitioned into subsets D_i. The labor capacities and flow of workers will be constant for each $d \in \mathsf{D}_i$. For example, D_1 could correspond to all Mondays and Fridays, and D_2 could correspond to all Tuesdays, Wednesdays and Thursdays. The optimization model is separately formulated for each D_i.

Since the optimization model *assumes* knowledge of the line arrivals, which in actuality management will not know, the production schedule output provides order release *guidelines*. We will discuss this in more detail in the subsequent section.

22.3.1 Parameters

The parameter classes involve space conversion factors, labor rates, capacities, and the customer lines.

- *Space conversion factors.*
  - $ToteLines[r, z, p]$ denotes the average number of lines per tote on each route r, in pick zone z, using packing process p. These parameters are

used to convert lines into requirements for space to store tote work-in-process between the pick and pack stages. (It would be more accurate to record conversion factors at the customer-line level, which is often difficult and expensive to estimate.)

- *Labor rates.*
  - $PickRate[z]$ denotes the average pick rate for each zone z measured in lines per hour.
  - $PackRate[p]$ denotes the average pack rate for each process p measured in lines per hour.
- *Labor types and costs.* A worker of type "H" works H consecutive hours. For full time workers $H \geq 8$ and for part time workers $H \leq 4$.
  - C_H denotes the cost per hour of an H type worker.
  - C_O denotes the cost per overtime hour.
  - $PartTimeBound$ denotes the maximum number of part time workers the company may hire.
- *Capacities.*
  - $ToteCapacity[p]$ denotes the tote capacity of work-in-process between the pick and pack stages.
  - $ProcessRateBound[p]$ denotes the bound on the production rate of process p. Some packing processes require a machine and/or there can be a limit on the size of a crew.
- *Customer lines.*
  - $LINES[r, z, p, t, d]$ denotes the *cumulative* number of customer lines for route r in zone z requiring process p up to time period t on day d. This represents the queue of arrivals that can be inducted into the system.
  - S_r denotes the period by which all lines for route r must be packed.
- *Earliest start times.*
  - τ_{rp} denotes the earliest period by which any lines on route r requiring process p can be packed. These parameters ensure that there will be sufficient space to stage pallets awaiting loading, since a pallet must be preset on the floor to store even a single line.

22.3.2 Decision Variables

The decision variable classes determine the production schedule, labor capacities, and workforce schedule.

- *Production quantities.* The rate of induction of order-lines and subsequent processing will determine the amount of labor required to meet the shipping due dates.
  - $Pick[z, r, p, t, d]$ denotes the number of lines in zone z on route r requiring process p that are picked within period t on day d. $PICK[z, r, p, t, d]$ denotes the cumulative number of such lines up to time t.

  - $Pack[r, p, t, d]$ denotes the number of lines on route r requiring process p that are packed within period t on day d. $PACK[z, r, p, t, d]$ denotes the cumulative number of such lines up to time t.
- *Labor capacities.*
  - $Labor[H, t, \mathsf{D}_i]$ denotes the (integer) number of workers who begin their shift at the beginning of period t for each day $d \in \mathsf{D}_i$ and who work for H consecutive periods.
- *Labor assignments.*
  - $Pickers[z, t, \mathsf{D}_i]$ denotes the (integer) number of workers assigned to pick zone z in period t for each day $d \in \mathsf{D}_i$.
  - $Packers[t, \mathsf{D}_i]$ denotes the (integer) number of workers assigned to pack zone z in period t for each day $d \in \mathsf{D}_i$.

22.3.3 Constraints

The constraint classes involve labor, inventory, space, and process bound requirements.

- *Labor requirements.*

  A production schedule for picking and packing requires a minimum amount of labor hours, and must not exceed the amount of workers assigned to the respective picking and packing zones. For all zones z, periods t, and days d,

$$\sum_r \sum_p \frac{Pick[r, z, p, t, d]}{PickRate[z]} \;\leq\; Pickers[z, t, d]. \tag{22.1}$$

$$\sum_r \sum_p \frac{Pack[r, p, t, d]}{PackRate[z]} \;\leq\; Packers[t, d]. \tag{22.2}$$

  Workforce assignments cannot exceed the total amount of full and part time workers actually present during each period.

$$\sum_z Pickers[z, t, d] \;+\; Packers[t, d] \;\leq\; \sum_H \sum_{\tau=t-H+1}^{t} Labor[H, \tau, \mathsf{D}_i]. \tag{22.3}$$

  The number of part time workers cannot exceed the maximum amount that can be hired. For all periods t,

$$\sum_{H \leq 4} Labor[H, t, \mathsf{D}_i] \;\leq\; PartTimeBound. \tag{22.4}$$

- *Inventory requirements.*

  No line can be packed until it has been picked. To be conservative a one-period lead time is assumed between the pick and pack stages; that is, only those lines that have been picked prior to the *beginning* of a period can be packed during the period. For all routes r, packing processes p, periods t, and days d,

$$\sum_z PICK[r, z, p, t-1, d] \;\geq\; PACK[r, p, t, d]. \tag{22.5}$$

  All lines for a route must be packed by its due date S_r. For all routes r, packing processes p, and days d,

$$PACK[r, p, S_r, d] \;=\; \sum_z LINES[r, z, p, S_r - 1, d]. \tag{22.6}$$

- *Space requirements.*

  To control congestion, the work-in-process in front of each packing process is bounded. The work-in-process is the difference between the cumulative number of totes picked and packed. For all packing processes p, periods t, and days d,

$$\frac{\sum_r \sum_z PICK[r, z, p, t, d] - PACK[r, p, t, d]}{ToteLines[p]} \;\leq\; ToteCapacity[p]. \tag{22.7}$$

  The limits on pallet staging capacity, proxied by the earliest packing start times for each route, must not be exceeded. For all routes r, packing processes p, periods $t < \tau_{rp}$, and days d,

$$PACK[r, p, t, d] = 0. \tag{22.8}$$

  (The variable $PACK[r, p, t, d]$ only enters the first equation as long as $t \leq S_r$.)

- *Process rate bounds.*

  The limits on how much can be packed must not be exceeded. For all packing processes p, periods t, and days d,

$$\sum_r Pack[r, p, t, d] \;\leq\; ProcessRateBound[p]. \tag{22.9}$$

22.3.4 Objective Function

Since there are no capital expenditures (by fiat), and since space is given and accounted for in the constraints, the workforce schedule and labor staffing are chosen to minimize the cost of labor.

$$MIN \sum_H \sum_t [C_H \min(H, 8) \;+\; C_O \max(H - 8, 0)] * Labor[H, t, \mathsf{D}_i] \tag{22.10}$$

22.4 Implementation

22.4.1 Computational Issues

The prototype optimization model has linear constraints and objective, but uses integer variables for labor staffing and workforce scheduling. Integer variables are used to model the reasonable requirement that if a worker is assigned to a zone for a period, then all of that worker's time for the period cannot be used elsewhere. It is not practical from either a management or efficiency perspective to assume that workers are freely movable, minute-by-minute.

Generally, the computational time grows exponentially with the number of integer variables. With three possible worker assignments in any period—broken or full case picking and packing—and using 30-minute periods, the number of integer variables required to model the possibilities for labor staffing and workforce scheduling is about 150 *for each cluster of days* D_i. Since the use of integer variables poses computational difficulties, any reasonable limitation that reduces the number of such variables will be beneficial. Here are some options:

- *Limit shift length.* As a reasonable starting point, assume that a full time worker can work at most 12 hours per day, thus limiting a worker's overtime to 4 hours per night. Assume that a part time worker must work 4 hours per shift. Thus H can take on 6 values.
- *Expand period length.* Expand the period length to say 1 or 2 hours. Since a worker's time is billed to the zone for the entire period regardless of actual use, lengthening the time period will lead to an incremental loss in labor efficiency.
- *Limit shift starting times.* Assume that a full time worker can begin his shift at any time between noon and 6:30pm, and that a part time worker can begin his shift at either noon, 2pm, 4pm, 6pm, 8pm, 10pm. (Recall that the third route wave departs at 2:00am.)
- *Implement worker waves.* A worker wave is a block of time during the day in which all assignments are frozen. A worker wave eases the burdens associated with managing a highly variable flow of workers. For example, in lieu of a workforce schedule that results in 12, 4 and 8 pickers assigned to a pick zone in three consecutive periods, a smoother assignment would assign 8 pickers in each period. To establish the worker waves, one could first solve a few problems to see where natural breaks occur in the schedule.
- *Solve one day at a time.* One could solve the prototype model for each day of the week but only use one day at a time. By examining the output, one can observe the general trend to obtain a suitable labor staffing and workforce schedule for that day and every other day in the week. Pooling all days for a week results in a loss of labor efficiency, but the weekly schedule is easier to manage.

22.4.2 Using the Prototype Model: A Case Study

To illustrate how to use the prototype model, all options discussed above were implemented on a real data set, with the exception of expanding the period length. The total cost was measured in worker hours with the cost of a regular time worker hour being set to 1.0, the cost of an overtime hour being set to 1.5, and the cost of a part time worker hour being set to 0.4. (Part time workers are not paid full benefits.) A maximum of 10 part time workers could be hired. The earliest start time for packing any lines associated with the second route wave was set at 8:00pm and was set at 10:00pm for the third route wave.

The optimization model was separately run for each of 19 days within a peak month. An AMPL front-end for model maintenance was used in conjunction with the CPLEX mixed-integer linear programming engine. Here are the steps:

- *Step 1: Determine the worker waves.* Four worker waves revealed themselves naturally: noon-4:30pm, 4:30pm-8:30pm, 8:30pm-12:30am, 12:30am-2:30am.
- *Step 2: Determine sensible start times for workers.* The worker waves were fixed and the model was run again for each day. The start times for most full time workers were at noon, 4:30pm and 6:30pm, and the start times for most part time workers were at 4:00pm and at 10:00pm. These start times were subsequently fixed.
- *Step 3: Determine baseline full time staff to handle non-peak days.* The worker waves and start times were fixed, and the model was run again for each day. The results showed that no part time workers were hired and no overtime was used for those days whose line total was less than 8,000 lines. With respect to full time workers 7-8 were hired at noon, 14-16 were hired to start at 4:30pm, and 4-5 were hired to start at 6:30pm. Since the model does not include down time for dinner, breaks, and sick time/vacation, it wad decided to add 15% to the workforce. Since the data were collected for the peak month (in which some overtime and temporary workers are to be expected), and since most days throughout the year fell below the 8,000 line level, it was decided to set a baseline full time staff to meet the 8,000 line level, as follows: 9@noon, 18@4:30pm, and 6@6:30pm.
- *Step 4: Determine temporary and part time staff to handle peak days.* The model was then run again for each day with the baseline staff *fixed* and only permitting additional full time workers to be hired at 6:30pm. These workers would be viewed as temporary staff required to meet the peak demand during the peak month.
- *Step 5: Determine order release guidelines.*

22.4.3 Benefits and Other Applications

Using the optimization-based approach is a systematic way to staff labor and schedule the workforce.

The following results were reported to management:

- *Labor staffing:* the total number of lines, the theoretical minimum number of full-time equivalent workers, the optimum labor cost, the distribution of the labor staffing, and the amount of overtime for each day.
- *Workforce schedule:* the number of workers assigned to each pick and pack zone within each worker wave for each day.
- *Order release guidelines:* cumulative number of lines picked for each route. For example, roughly 20% of all lines for route 1 should be released and picked by 2:00pm, 35% by 4:00pm, and 80% by 6:00pm; the corresponding numbers for route 2 are 5%, 15%, and 20% respectively.
- *Packing guidelines:* cumulative number of lines packed for each route. For example, roughly 60% of all route 1 lines that have been picked by 2:00pm should be packed by this time, 70% by 4:00pm, and 80% by 6:00pm; for route 2 roughly 40% of all lines that have been picked by 8:00pm should be packed by this time and 80% should be packed by 10:00pm.

After reviewing the results, it was determined that (i) due to the late line arrivals and early departure time for the first route wave, a strategic use of part timer workers at 4:00pm helped to reduce the overall need for full time staff at 4:30pm; and (ii) the 6:30pm start times for the full time workers and the 10:00pm start times for the 4-hour part timers helped to eliminate the overtime used by the original 3:30pm crew to handle occasional peak loads for the third route wave.

The projected benefits included a 20% reduction in full time staff for normal days (41 down to 33), and an 80% reduction in overtime hours for peak days (100 down to 20). Once the core model has been built, a number of scenarios can be examined to assess additional economic benefits, such as:

- Reorganizing routes and route departure times.
- Extending the order cutoff time.
- Adding staging capacity.
- Increasing packing efficiency.
- Using 3 or 4 day work weeks.

22.5 Bibliographical Notes

This application was jointly undertaken with John J. Bartholdi. For an in-depth description of warehouse systems, consult Bartholdi and Hackman [2007], Frazelle [2001], Tompkins and Smith [1998], Mulcahy [1993], or Jenkins [1990].

Part V

Mathematical Appendix

A

Notation and Mathematical Preliminaries

A.1 Logical Statements

- The notation := is shorthand for **by definition**. This means the object to the left of := is by definition equal to whatever is written to the right of :=.
- Let A and B denote two sets of logical conditions. The statement A **if and only if** B and symbolized by $A \iff B$ represents two logical statements:
  - the **if part** which means that if B holds, then A must hold. Also expressed as "B implies A" and symbolized by $A \Longleftarrow B$.
  - the **only if part**, which means that A holds only if B holds or, equivalently, if B holds, then A must hold. Also expressed as "A implies B" and symbolized by $A \Longrightarrow B$.

A.2 Sets

A set is taken as a primitive concept. The objects that constitute a set are called the **elements** of the set.

Membership

- $x \in S$ means x is an element of S or x belongs to S.
- $x \notin S$ means x is not an element of S or x does not belong to S.
- $\{a, b, c, \ldots, z\}$ denotes the set whose elements are those listed inside the braces, i.e., a, b, c up to z—the meaning of "up to" must be clear in the context. For example, given that the symbol n is a positive integer, the set $\{1, 2, \ldots, n\}$ is the set of integers from 1 up to n.
- $\{x : \ P(x)\}$ means the set of all x for which the proposition $P(x)$ is true. The symbol $\{x \in S : \ P(x)\}$ means the set of all $x \in S$ for which the proposition $P(x)$ is true.

- $S \subset T$ means every element of S is also an element of T. The set S is said to be a **subset** of T or **contained** in T.
- $S = T$ means the sets S and T are **equal**, namely, $S \subset T$ and $T \subset S$.

Union, intersection, complements

- $S \cup T$ is the set of all elements x that belong to S or T, otherwise known as the **union** of the sets S and T. It can also be expressed as $\{x : \ x \in S \text{ or } x \in T\}$.
- $S \cap T$ is the set of all elements x that belong to both S and T, otherwise known as the **intersection** of the sets S and T. It can also be expressed as $\{x : \ x \in S \text{ and } x \in T\}$.
- A set S is said to **meet** the set T if their intersection $S \cap T$ is not empty.
- $S \backslash T$ is the set of elements x that belong to S but do not belong to T, known as the **complement of T with respect to S**. It can also be expressed as $\{x : \ x \in S \text{ and } x \notin T\}$.

Sum and product

- $S+T$ is the set $\{s+t : \ s \in S, \ t \in T\}$, known as the **sum or set-theoretic addition** of the sets S and T. *The sets S and T must be such that the sum makes sense.*
- (x, y) is an **ordered pair** taken to be a primitive concept. Two ordered pairs (x_1, y_1), (x_2, y_2) are equal if $x_1 = x_2$ and $y_1 = y_2$.
- $S \times T$ is the set $\{(x, y) : \ x \in S \text{ and } y \in T\}$, known as the **cartesian product** of the sets S and T. More generally, $\prod_{i=1}^{N} S_i$ is the set $\{(x_1, x_2, \ldots, x_N) : \ x_i \in S_i \text{ for every } i = 1, 2, \ldots, N\}$, the cartesian product of the sets $S_1, S_2, \ldots, S_N$. If the S_i are identical to a set S, then $\prod_{i=1}^{N} S_i$ is denoted by S^N and is called the N-fold cartesian product of S.

Families

- A set $\mathcal{F} = \{S_i\}_{i \in \mathcal{I}}$ whose elements S_i are sets for each $i \in \mathcal{I}$ is called a **family** of sets. The set $\mathcal{I}$ is known as the **index set**, which can be uncountable. A family is said to be a **finite family** of sets when the index set is finite. The index set will often be suppressed from the notation.
- Let $\mathcal{F}$ be a family of sets. If $\mathcal{F}' \subset \mathcal{F}$, i.e., every set in the family $\mathcal{F}'$ is also a member of $\mathcal{F}$, then $\mathcal{F}'$ is said to be a **subfamily** of $\mathcal{F}$.
- Given a family $\{S_i\}_{i \in \mathcal{I}}$ of sets, its **intersection** is the set $\{x : \ x \in S_i \text{ for every } i \in \mathcal{I}\}$ and its **union** is the set $\{x : \ x \in S_i \text{ for some } i \in \mathcal{I}\}$.

Finite-dimensional spaces

- The symbols $\mathbb{R}$, $\mathbb{R}_+$, $\mathbb{R}_-$, $\mathbb{R}_{++}$ and $\mathbb{R}_{--}$ denote, respectively, the set of real numbers, the set of nonnegative numbers, the set of negative numbers, the set of positive numbers, and the set of negative numbers.
- The symbols $\mathbb{R}^n$, $\mathbb{R}^n_+$, $\mathbb{R}^n_-$, $\mathbb{R}^n_{++}$ and $\mathbb{R}^n_{--}$ denote, respectively, the n-fold cartesian product of the sets $\mathbb{R}$, $\mathbb{R}_+$, $\mathbb{R}_-$, $\mathbb{R}_{++}$ and $\mathbb{R}_{--}$.
  - $\mathbb{R}^n_+$ is called the **nonnegative orthant** of R^n.
  - $\mathbb{R}^n_{++}$ is called the **positive orthant** of R^n.
  - $\mathbb{R}^n_-$ is called the **nonpositive orthant** of R^n.
  - $\mathbb{R}^n_{--}$ is called the **negative orthant** of R^n.

Supremum and infimum

- Let S be a subset of the real line.
  - A real number α is said to be an **upper bound** of S if $x \leq \alpha$ for every $x \in S$. A set $S \subset \mathbb{R}$ is not guaranteed to have an upper bound; if it does, however, the set S is said to be **bounded above**.
  - A real number α is said to be a **lower bound** of S if $x \geq \alpha$ for every $x \in S$. A set $S \subset \mathbb{R}$ is not guaranteed to have a lower bound; if it does, however, the set S is said to be **bounded below**.

  Remark A.1. As a logical consequence of the definitions, every real number is both an upper bound and lower bound of the empty set.

- Let S be a subset of the real line.
  - If S is bounded above, then an upper bound of S is said to be a **supremum or least upper bound** of S if it is less than any other upper bound of S. If the supremum of S also belongs to S, it is said to be a **maximum** of S.
  - If S is bounded below, then a lower bound of S is said to be an **infimum or greatest lower bound** of S if it is greater than any other lower bound of S. If the infimum of S also belongs to S, it is said to be a **minimum** of S.

  Remark A.2. A fundamental property of the real number system, called *completeness*, guarantees that every nonempty set of real numbers that is (i) bounded above has a supremum and (ii) bounded below has an infimum. The supremum and infimum must be unique, and will be respectively denoted by $\sup S$ and $\inf S$.

Special sets

- $\emptyset$ denotes the set without any elements, otherwise known as the **empty set**.
- $\mathbf{N}$ denotes the set of positive integers $\{1, 2, 3, \ldots\}$.

- $\mathbf{Z}$ denotes the set of integers $\{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$.
- The notation for intervals of the real line are as follows:
  - (a, b) denotes the set $\{x : a < x < b\}$, referred to as the **open interval between a and b**;
  - $[a, b)$ denotes the set $\{x : a \leq x < b\}$ and $(a, b]$ denotes the set $\{x : a < x \leq b\}$. Both sets are referred to as **half-open intervals between a and b**;
  - $[a, b]$ denotes the set $\{x : a \leq x \leq b\}$, referred to as the **closed interval between a and b**.
- $\mathcal{R}(x)$ denotes the set $\{sx : s \geq 0\}$, which is the **ray emanating from the origin and passing through the point** $x \in \mathbb{R}^n_+$.
- A nonempty set $S \subset \mathbb{R}^n$ is a **cone** (with vertex zero) if $x \in S$ implies that $\lambda x \in S$ for all $\lambda \geq 0$.

A.3 Vectors

Representation

- Each $x \in \mathbb{R}^n$ is called a **point** or **vector**. Each $x \in \mathbb{R}^n$ will sometimes be written as a **row vector** $x = (x_1, x_2, \ldots, x_n)$ and sometimes be written as a **column vector**

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

  The entry x_i is called the i^{th} **coordinate** of x.
- x^T denotes the **transpose** of the vector x. If x is understood to be a column (row) vector, then x^T is x written as a row (column) vector.

Special vectors

- The **zero** vector is the point in $\mathbb{R}^n$ whose coordinates all equal zero. It will be denoted by the zero symbol 0.
- $x \in \mathbb{R}^n$ is **nontrivial** if $x \neq 0$.
- The **sum** vector is the point in $\mathbb{R}^n$ whose coordinates all equal the number one. It will be denoted by the letter e.
- The i^{th} **unit or coordinate vector** is the point in $\mathbb{R}^n$ whose coordinates all equal zero *except* coordinate i, whose value equals the number one. It will be denoted by the symbol e_i.

Addition and multiplication

- Let $x, y \in \mathbb{R}^n$. The symbol $x + y$ denotes the **sum of the two vectors** x **and** y, which is defined to be the vector $z \in \mathbb{R}^n$ whose i^{th} coordinate z_i is $x_i + y_i$.
- Let $\alpha \in \mathbb{R}$ and let $x \in \mathbb{R}^n$. The symbol αx denotes the **multiplication of the real number** α **with the vector** $x \in \mathbb{R}^n$, which is defined to be the vector $y \in \mathbb{R}^n$ whose i^{th} coordinate y_i is αx_i.
- $x \cdot y = \sum_{i=1}^{n} x_i y_i$ is called the **dot or inner product** of the vectors $x, y \in \mathbb{R}^n$. See Remark A.5, p. 443.

Relationships

- For $x, y \in \mathbb{R}^n$:
  - $x \geq y$ means that $x_i \geq y_i$ for each $i = 1, 2, \ldots, n$.
  - $x \gneqq y$ means that $x \geq y$ and $x \neq y$.
  - $x > y$ means that $x_i > y_i$ for each $i = 1, 2, \ldots, n$.
  - $x \leq y$ means that $x_i \leq y_i$ for each $i = 1, 2, \ldots, n$.
  - $x \lneqq y$ means that $x \leq y$ and $x \neq y$.
  - $x < y$ means that $x_i < y_i$ for each $i = 1, 2, \ldots, n$.

A.4 Correspondences

Relations

- A **relation** ϕ of a set S into a set T is a subset of $S \times T$; that is, a relation ϕ is a set of ordered pairs (x, y) where $x \in S$ and $y \in T$.
- $dom(\phi)$ is the **domain** of a relation ϕ, which is defined to be the set $\{x \in S : (x, y) \in \phi \text{ for some } y \in T\}$.
- $range(\phi)$ is the **range** of a relation ϕ, which is defined to be the set $\{y \in S : (x, y) \in \phi \text{ for some } x \in S\}$.
- $\phi(x)$ denotes the **image** of $x \in S$ with respect to the relation ϕ, which is defined to be the set $\{y \in T : (x, y) \in \phi\}$.
- $\phi(A)$ denotes the **image** of $A \subset S$ with respect to the relation ϕ, which is defined to be the set $\{y \in T : (x, y) \in \phi \text{ for some } x \in A\}$.

Definition

- A relation ϕ of S into T is called a **correspondence** if $dom(\phi) = S$.

A.5 Functions

Definition

- A correspondence ϕ of S into T is called a **function** or **mapping** if for every $x \in S$ there is a *unique* $y \in T$ such that $(x, y) \in \phi$. That is, for every $x \in S$, the image of x with respect to ϕ, $\phi(x)$, consists of a single element. The unique element y is called the **value** of ϕ at x. With slight abuse of notation, we write $y = \phi(x)$ and use the symbol $\phi(x)$ to denote y. We write $f : S \to T$ to denote a mapping f of S into T.

  Remark A.3. Let 2^T denote the set of all nonempty subsets of the set T. A correspondence ϕ of S into T can be thought of as a mapping of S into 2^T.

Indicator Function

- For a fixed $A \subset \mathbb{R}^n_+$ define

$$1_A(t) := \begin{cases} 1 & \text{if } t \in A, \\ 0 & \text{if } t \notin A. \end{cases}$$

  The function $1_A(\cdot)$ is called the **indicator function of the set** A. It will also be denoted as $1(t \in A)$. The indicator function of the set A takes on the value 1 if the event A is true and takes on the value 0 if the event A is false. For example, the function

$$20 \cdot 1_{[2,5]}(\cdot) = 20 \cdot 1(2 \leq t \leq 5)$$

  is zero everywhere on the real line, except on the interval $[2, 5]$ where it takes on the value of 20.

Sequences

- A mapping $x : \mathbf{N} \to S$ of the positive integers $\mathbf{N}$ into a set S is called a **sequence** of points in S. It will be denoted by the expression $\{x_n\}$ or the expression $x_1, x_2, \ldots, x_n, \ldots$. (In context there will be no confusion as to whether we are referring to a sequence or a set whose only element is x_n.) A **subsequence** of the sequence x is the restriction of x to a subset $K \subset \mathbf{N}$ such that for each $n \in \mathbf{N}$ there is an $k \in K$ such that $k \geq n$. It will be denoted by the expression $x_{n_k} : k = 1, 2, \ldots$.

Level sets

- The **level sets** of a function $f : S \to \mathbb{R}$ at $\alpha \in \mathbb{R}$ are defined as:

$$L_f^{\geq}(\alpha) := \{x \in S : \ f(x) \geq \alpha\} \quad (\textbf{upper level set}).$$
$$L_f^{>}(\alpha) := \{x \in S : \ f(x) > \alpha\} \quad (\textbf{strict upper level set}).$$
$$L_f^{\leq}(\alpha) := \{x \in S : \ f(x) \leq \alpha\} \quad (\textbf{lower level set}).$$
$$L_f^{<}(\alpha) := \{x \in S : \ f(x) < \alpha\} \quad (\textbf{strict lower level set}).$$

Graphs

- The **graph** of the function $f : S \to \mathbb{R}$ is the set

$$Gr(f) := \{(x, f(x)) \in S \times R : \ x \in S\}.$$

- The **hypograph** of the function $f : S \to \mathbb{R}$ is the set

$$hypo(f) := \{(x, \gamma) \in S \times R : \ \gamma \leq f(x)\}.$$

  It is the collection of all points in $S \times \mathbb{R}$ that "lie on or below" the graph of $f(\cdot)$.
- The **epigraph** of the function $f : S \to \mathbb{R}$ is the set

$$epi(f) := \{(x, \gamma) \in S \times R : \ \gamma \geq f(x)\}.$$

  It is the collection of all points in $S \times \mathbb{R}$ that "lie on or above" the graph of $f(\cdot)$.

Monotonicity

- Let $S \subset \mathbb{R}^n$ and let $f : S \to \mathbb{R}$.
  - $f(\cdot)$ is said to be **nondecreasing** on S if $f(y) \geq f(x)$ whenever $y \geq x$.
  - $f(\cdot)$ is said to be **increasing** on S if $f(y) > f(x)$ whenever $y \geqneq x$.
  - $f(\cdot)$ is said to be **nonincreasing** on S if $f(y) \leq f(x)$ whenever $y \geq x$.
  - $f(\cdot)$ is said to be **decreasing** on S if $f(y) < f(x)$ whenever $y \geqneq x$.

Homogeneity

- Let S be a cone in $\mathbb{R}^n$. A function $f : S \to \mathbb{R}$ is **homogeneous of degree** k if $f(\lambda x) = \lambda^k f(x)$ for every $x \in S$ and $\lambda \geq 0$.
- A function is **linearly homogeneous** if it is homogeneous of degree one.

A.6 Matrices

Definitions

- A **matrix** is a rectangular array of real numbers. If the matrix has m rows and n columns, it is called an $m \times n$ matrix. The entry in row i and column j of a matrix A is denoted by a_{ij}. For example, the matrix

$$A = \begin{bmatrix} 1 & 3 & 4 \\ 2 & 8 & 5 \end{bmatrix} \tag{A.1}$$

  has two rows and three columns.
- The **transpose of an** $m \times n$ **matrix** A, denoted by the symbol A^T, is the $n \times n$ matrix whose $(i,j)^{th}$ entry is a_{ji}. The i^{th} column of A becomes the i^{th} row of A^T.

  Remark A.4. A row vector $x \in \mathbb{R}^n$ can be thought of as an $1 \times n$ matrix, and its transpose, x^T, can be thought of as an $n \times 1$ matrix. Similarly, a column vector $x \in \mathbb{R}^n$ can be thought of as an $n \times 1$ matrix, and its transpose, x^T, can be thought of as an $1 \times n$ matrix.

Addition and multiplication

- Let A and B be two $m \times n$ matrices. The **sum** of A and B is the matrix whose $(i,j)^{th}$ entry is $a_{ij} + b_{ij}$.
- Let A be an $m \times n$ matrix and let $\lambda \in \mathbb{R}$. The symbol λA denotes the $m \times n$ matrix whose $(i,j)^{th}$ entry is λa_{ij}.
- Let A be an $m \times n$ matrix and let $x \in R^n$. The symbol Ax denotes the **multiplication of the** $m \times n$ **matrix** A **with the vector** $x \in \mathbb{R}^n$. It is defined to be the vector $y \in \mathbb{R}^m$ whose coordinates are given by

$$y_i := \sum_{j=1}^{n} a_{ij} x_j.$$

  This multiplication is often written down as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \dots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

  For example, the multiplication of the matrix A given in (A.1) with the vector $x^T = (4, 1, 3)$ is

$$\begin{bmatrix} 1 & 3 & 4 \\ 2 & 8 & 5 \end{bmatrix} \begin{pmatrix} 4 \\ 1 \\ 3 \end{pmatrix} = \begin{pmatrix} (1)(4) + (3)(1) + (4)(3) \\ (2)(4) + (8)(1) + (5)(3) \end{pmatrix} = \begin{pmatrix} 19 \\ 31 \end{pmatrix}.$$

The multiplication of a matrix A with a vector x only makes sense when the number of columns of A equals the number of coordinates of x.

Remark A.5. The dot product of two vectors x and y can be represented via matrix multiplication. If x, y are column vectors, then $x \cdot y = x^T y = y^T x$. If, on the other hand, x, y are row vectors, then $x \cdot y = xy^T = yx^T$.

- Let A be an $m \times n$ matrix and let B be an $n \times p$ matrix. The symbol AB denotes the **multiplication of the matrix A with the matrix B**. It is defined to be the $m \times p$ matrix C whose $(i,j)^{th}$ entry is

$$c_{ij} := \sum_{k=1}^{n} a_{ik} b_{kj}.$$

This multiplication is often written down as

$$\begin{bmatrix} c_{11} & c_{12} & \dots & c_{1p} \\ c_{21} & c_{22} & \dots & c_{2p} \\ \vdots & \vdots & \dots & \vdots \\ c_{m1} & c_{m2} & \dots & c_{mp} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \dots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1p} \\ b_{21} & b_{22} & \dots & b_{2p} \\ \vdots & \vdots & \dots & \vdots \\ b_{n1} & b_{n2} & \dots & b_{np} \end{bmatrix}.$$

The i^{th} column of C is the multiplication of the matrix A with the column vector given by the i^{th} column of the matrix B. For example,

$$\begin{bmatrix} 1 & 3 & 4 \\ 2 & 8 & 5 \end{bmatrix} \begin{bmatrix} 2 & 7 & 3 & 5 \\ 6 & 5 & 2 & 6 \\ 0 & 3 & 1 & 4 \end{bmatrix} = \begin{bmatrix} 20 & 34 & 13 & 39 \\ 52 & 69 & 27 & 78 \end{bmatrix}.$$

The multiplication of a matrix A with a matrix B only makes sense when the number of columns of A equals the number of rows of B.

- Let A be an $n \times m$ matrix. It is possible to subdivide A into **submatrices** as follows:
  - An $n \times k$ matrix B consisting of $k \leq m$ columns of A *ordered from left to right as they appear in A* is called a **column submatrix** of A. The notation $A = [A_1 \quad A_2]$ means that A_1 and A_2 are column submatrices of A consisting, respectively, of the first k_1 columns of A and the last $m - k_1$ columns of A.
  - A $k \times m$ matrix B consisting of $k \leq n$ rows of A *ordered from top to bottom as they appear in A* is called a **row submatrix** of A. The notation $A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$ means that A_1 and A_2 are row submatrices of A consisting, respectively, of the first k_1 row of A and the last $n - k_1$ columns of A.
  - Since each submatrix can be further subdivided into its own submatrices, it is possible to subdivide an $n \times m$ matrix A as

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

The number of rows of A_{11} and A_{12} respectively equals the number of rows of A_{12} and A_{22}. Likewise, the number of columns of A_{11} and A_{12} respectively equals the number of columns of A_{21} and A_{22}.

- The product of two (appropriately dimensioned) subdivided matrices A and B can be multiplied in the following way:

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} A_{11}B_{11} & A_{12}B_{12} \\ A_{21}B_{21} & A_{22}B_{22} \end{bmatrix}.$$

Rank

- Vectors $x_i \in \mathbb{R}^n$, $i = 1, 2, \ldots, m$, are said to be **linearly independent** if there does not exist real numbers α_i, $i = 1, 2, \ldots, n$, such that $\sum_i \alpha_i x_i = 0$. (If the vectors are linearly independent, then $m \leq n$.)
- The **rank** of an $m \times n$ matrix A is the maximum number of linearly independent rows or, equivalently, the maximum number of linearly independent columns of A.
- The matrix A is said to be of **full rank** if its rank equals $\min\{m, n\}$.

Special matrices

- A **square** matrix has the same number of rows as columns.
- A square matrix A is **symmetric** if $A = A^T$.
- The square matrix A is called the **identity** matrix if $a_{ij} = 0$ if $i \neq j$ and $a_{ii} = 1$ for each i. It is denoted by the symbol I.
- A square matrix A is called **nonsingular** if there is a matrix B, called the **inverse** matrix, such that $AB = BA = I$. The inverse of a square matrix, if it exists, is unique, and is denoted by the symbol A^{-1}. (A nonsingular matrix must be of full rank.)
- An $n \times n$ symmetric matrix A is **positive definite** if $x^T Ax > 0$ for all $x \neq 0$. An $n \times n$ symmetric matrix A is **negative definite** if $x^T Ax < 0$ for all $x \neq 0$.

A.7 Differentiability

Please Note: Throughout this section S shall denote a nonempty set in $\mathbb{R}^N$ with nonempty interior, and $f : S \to \mathbb{R}$ shall denote a real-valued function defined on S. Also, $||x|| := (\sum_{i=1}^n x_i^2)^{1/2}$, known as the **Euclidean norm** of the vector $x \in \mathbb{R}^n$.

Subgradients

- The vector $\bar{x}^*$ is a **subgradient** of $f(\cdot)$ at $\bar{x} \in S$ if

$$f(x) \geq f(\bar{x}) + \bar{x}^* \cdot (x - \bar{x}) \tag{A.2}$$

  holds for all $x \in S$.
- Vector $\bar{x}^*$ is a **supergradient** of $f(\cdot)$ at $\bar{x} \in S$ if

$$f(x) \leq f(\bar{x}) + \bar{x}^* \cdot (x - \bar{x}) \tag{A.3}$$

  holds for all $x \in S$.

  Remark A.6. Here is a geometrical interpretation. Suppose $S \subset \mathbb{R}$ so that $f(\cdot)$ is a function of one variable. If the real number $\bar{x}^*$ is a supergradient of $f(\cdot)$ at $\bar{x}$, then the line

$$y = \bar{x}^*(x - \bar{x}) + f(\bar{x}) \tag{A.4}$$

  passes through the point $(\bar{x}, f(\bar{x}))$ and lies on or above the graph of $f(\cdot)$, i.e., $y = y(x) \geq f(x)$ for all $x \in S$. Figure A.1 depicts a supergradient of a nondifferentiable function $f(\cdot)$ at the point 3.

- The **subdifferential of** $f(\cdot)$ **at** $\bar{x}$ is the collection of all subgradients of $f(\cdot)$ at $\bar{x}$, and the **superdifferential of** $f(\cdot)$ **at** $\bar{x}$ is the collection of all supergradients of f at $\bar{x}$. Both the subdifferential and superdifferential are denoted by the symbol $\partial f(\bar{x})$.

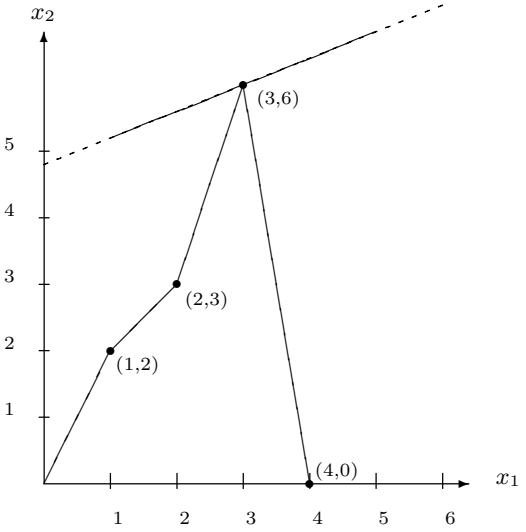


Fig. A.1. Example of a supergradient.

"Little oh" $o(\cdot)$ functions

The **little oh** functions are standard notational devices in analysis defined as follows:

- Let $\alpha \in \mathbb{R}$. The symbol "$o(\alpha)$" is a generic expression for any real-valued function $h(\alpha)$ that satisfies the property that $h(\alpha)/\alpha \to 0$ as $\alpha \to 0$.
- Let $d \in \mathbb{R}^n$. The symbol "$o(||d||)$" is a generic expression for any real-valued function $h(d)$ that satisfies the property that $h(d)/||d|| \to 0$ as $||d|| \to 0$. This says that as $||d|| \to 0$ the remainder term $o(||d||)$ goes to 0 faster than $||d||$ does. For example, the function y^n of the scalar y is $o(y)$ as long as $n > 1$.
- Let $d \in \mathbb{R}^n$. The symbol "$o(||d||^2)$" is a generic expression for any real-valued function $h(d)$ that satisfies the property that $h(d)/||d||^2 \to 0$ as $||d|| \to 0$. This says that as $||d|| \to 0$ the remainder term $o(||d||^2)$ goes to 0 faster than $||d||^2$ does. For example, the function y^n of the scalar y is $o(y^2)$ as long as $n > 2$.

  Remark A.7. The sum or difference of two $o(\cdot)$ functions and a scalar multiple of an $o(\cdot)$ function are still $o(\cdot)$ functions.

Differentiable functions

- A function $f(\cdot)$ is **differentiable** at x in the interior of S if there exists a vector $\nabla f(x)$, called the **gradient vector**, for which

$$f(x+d) = f(x) + \nabla f(x) \cdot d + o(||d||) \tag{A.5}$$

  holds for all d for which $x + d \in S$. A function $f(\cdot)$ is *differentiable* if it is differentiable at each point in the interior of S.
- When $S \subset \mathbb{R}$ and $f(\cdot)$ is differentiable at x, the derivative of $f(\cdot)$ at x is denoted by the symbol $f'(x)$.
- Suppose $f(\cdot)$ is differentiable at x. The linear function $f(x) + \nabla f(x) \cdot d$ of $d \in \mathbb{R}^n$ is called the **first-order Taylor series approximation of** $f(\cdot)$ **at** x.

  Remark A.8. Often d refers to the direction between x and another vector y, and one wishes to approximate the function $f(\cdot)$ near x in the direction d. One may use (A.5) to write

$$f(x + \alpha d) = f(x) + \alpha \nabla f(x) \cdot d + o(\alpha). \tag{A.6}$$

  (It is understood that α is sufficiently small so that $x + \alpha d \in S$.)
- Suppose $f(\cdot)$ is differentiable at x. The **partial derivative of** $f(\cdot)$ **with respect to** x_i **at** x is defined as

$$\frac{\partial f(x)}{\partial x_i} := \lim_{\alpha \to 0} \frac{f(x + \alpha e_i) - f(x)}{\alpha}. \tag{A.7}$$

  It coincides with the i^{th} coordinate of the vector $\nabla f(x)$.

Remark A.9. Differentiability of $f(\cdot)$ at x guarantees existence of all partial derivatives. The converse, however, is not true: the existence of all partial derivatives at a point does not guarantee that the function will be differentiable at that point.

- If a first-order Taylor series approximation exists, it should be unique; that is, there can only be one gradient vector. Furthermore, every supergradient or subgradient must coincide with the gradient.

  Theorem A.10. *If $f(\cdot)$ be differentiable at x, then (a) $\nabla f(x)$ is unique, and (b) if x^* is a subgradient or supergradient, then $x^* = \nabla f(x)$.*

- The following two well-known theorems have many applications.

  Theorem A.11. Mean-value theorem. *Suppose $f(\cdot)$ is differentiable and S is open. For every x_1 and x_2 in S there exists an $x = \lambda x_1 + (1-\lambda)x_2$ for which*

$$f(x_2) = f(x_1) + \nabla f(x) \cdot (x_2 - x_1). \tag{A.8}$$

  Theorem A.12. L'Hôpital's rule. *Let $f(\cdot)$ and $g(\cdot)$ be two real-valued differentiable functions defined on $S \subset \mathbb{R}$ such that $a \in S$. If $f(a) = g(a) = 0$, and if the limit of the ratio $f'(x)/g'(x)$ as x approaches a exists, then*

$$\lim_{x \to a} \frac{f(x)}{g(x)} = \lim_{x \to a} \frac{f'(x)}{g'(x)}. \tag{A.9}$$

Twice differentiable functions

- A function $f(\cdot)$ is **twice differentiable** if it is differentiable and each partial derivative is differentiable.
- For a twice differentiable function $f(\cdot)$ the **Hessian matrix** $H(x)$ is the $n \times n$ matrix whose $(i,j)^{th}$ entry is $\frac{\partial(\partial f(x)/\partial x_j)}{\partial x_i}$. It is more commonly denoted by the symbol $\frac{\partial^2 f(x)}{\partial x_i \partial x_j}$.

  Remark A.13. The Hessian matrix is symmetric. That is, $\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x)}{\partial x_j \partial x_i}$ so that $H(x) = H(x)^T$.

- A twice differentiable function $f(\cdot)$ is **twice continuously differentiable** if $H(x)$ is continuous for every $x \in S$.

  Remark A.14. As a consequence of *Taylor's theorem*, for twice continuously differentiable functions

$$f(x+d) = f(x) + d \cdot \nabla f(x) + 1/2\, d^T H(x) d + o(||d||^2) \tag{A.10}$$

  holds for all d for which $x + d \in S$.

- Let $f(\cdot)$ be a twice differentiable function at x. The quadratic function $f(x) + \nabla f(x) \cdot d + \frac{1}{2} d^T H(x) d$ of $d \in real^n$ is called the **second-order Taylor series approximation of** $f(\cdot)$ **at** x.

  Remark A.15. Often d refers to the direction between x and another vector y, and we wish to approximate the function $f(\cdot)$ near x in the direction d. We then use (A.10) to write

$$f(x + \alpha d) = f(x) + \alpha \nabla f(x) \cdot d + 1/2\ \alpha^2 d^T H(x) d + o(\alpha^2). \qquad \text{(A.11)}$$

  A useful property is the following second-order form of Taylor's theorem.

  Theorem A.16. Second-order form of Taylor's theorem. *Assume $f(\cdot)$ is twice-differentiable and S is open. For each x_1 and x_2 in S there exists an $x = \lambda x_1 + (1 - \lambda) x_2$ for which*

$$f(x_2) = f(x_1) + \nabla f(x_1) \cdot (x_2 - x_1) + 1/2\ (x_2 - x_1)^T H(x)(x_2 - x_1). \quad \text{(A.12)}$$

B

Real Analysis

B.1 Linear Spaces

Vectors in $\mathbb{R}^n$ can be added together and multiplied by a scalar. These addition and multiplication operations are commutative, associative, and distributive. In addition, a vector $x \in \mathbb{R}^n$ has "length," often given by $||x|| := \sqrt{\sum_{i=1}^{n} x_i^2}$. A linear space is any set that possesses these types of operations, and a norm abstracts the notion of length.

B.1.1 Definition

Definition B.1. *A nonempty set L of elements $x, y, z, \ldots$ is a* **real linear or vector** *space if it satisfies the following properties:*

a) Any two elements $x, y \in L$ uniquely determine a third element $x + y \in L$, called the **sum** *of x and y, such that*
- $x + y = y + x$ *(***commutativity***);*
- $(x + y) + z = x + (y + z)$ *(***associativity***);*
- *There exists an element $0 \in L$, called the* **zero** *element, such that $x + 0 = x$ for every $x \in L$;*
- *For every $x \in L$ there exists an element $-x \in L$, called the* **negative** *of x, such that $x + (-x) = 0$.*

b) For any real number α and element $x \in L$ there exists a unique element $\alpha x \in L$, called the **product** *of α and x, such that*
- $\alpha(\beta x) = (\alpha\beta)x$;
- $1x = x$.

c) The addition and multiplication laws satisfy the following two **distributive** *laws:*
- $(\alpha + \beta)x = \alpha x + \beta x$;
- $\alpha(x + y) = \alpha x + \alpha y$.

The elements of L are called **points** or **vectors** and the real numbers $\alpha, \beta, \ldots$ are called **scalars**.

B.1.2 Examples

Example B.2. The real line with the usual arithmetic operations of addition and multiplication is a linear space.

Example B.3. The set of ordered n-tuples $x = (x_1, x_2, \ldots, x_n)$ of real numbers with addition and scalar multiplication defined **coordinatewise** by

$$\begin{aligned}(x_1, x_2, \ldots, x_n) + (y_1, y_2, \ldots, y_n) &:= (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)\\ \alpha(x_1, x_2, \ldots, x_n) &:= (\alpha x_1, \alpha x_2, \ldots, \alpha_n x_n)\end{aligned}$$

is a linear space. Of course, this space is recognized as $\mathbb{R}^n$.

Example B.4. The set $\mathbb{R}^\infty$ of all infinite sequences $x = (x_1, x_2, \ldots, x_k, \ldots)$ of real numbers with addition and scalar multiplication defined coordinatewise is a linear space.

Example B.5. There are a number of subsets of $\mathbb{R}^\infty$ that define linear spaces. Here are a few notable examples.

- The subset ℓ_1 of $\mathbb{R}^\infty$ is the collection of all x such that $\sum_{i=1}^\infty |x_i| < \infty$.[1] Here, one needs to verify that if $x, y \in \ell_1$, then

$$\sum_{i=1}^\infty |x_i + y_i| < \infty.$$

  This follows since the **triangle inequality** $|a+b| \le |a| + |b|$ holds for any two real numbers a and b.
- The subset ℓ_2 of $\mathbb{R}^\infty$ is the collection of all x such that $\sum_{i=1}^\infty x_i^2 < \infty$. Here, one needs to verify that if $x, y \in \ell_1$, then

$$\sum_{i=1}^\infty (x_i + y_i)^2 < \infty.$$

  This follows since $(a+b)^2 \le 2(a^2 + b^2)$ for any two real numbers a and b.
- The subset ℓ_∞ of $\mathbb{R}^\infty$ is the collection of all x such that $\sup_i x_i < \infty$. Each sequence $x \in \ell_\infty$ is **uniformly bounded**, i.e., for each $x \in \ell_\infty$, there exists a $B > 0$ such that $|x_i| < B$ for all i.

Example B.6. Let $S \subset \mathbb{R}^n$. The set L of all mappings $f : S \longrightarrow R$ with addition and multiplication defined by

$$\begin{aligned}(f + g)(x) &:= f(x) + g(x),\\ (\alpha f)(x) &:= \alpha f(x)\end{aligned}$$

is a linear space. Since the elements of this space are functions, this space is an example of a **function space**.

[1] The symbol $|\cdot|$ means take the absolute value.

Example B.7. Here are some notable function spaces that are linear spaces.

- The set L_1 consisting of all functions $f : \mathbb{R} \longrightarrow \mathbb{R}$ such that $\int_{-\infty}^{\infty} |f(x)|dx < \infty$. Such functions are called **integrable** functions.
- The set L_∞ consisting of all bounded functions $f : \mathbb{R} \longrightarrow \mathbb{R}$.
- The set $C[0,1]$ consisting of all continuous functions $f : [0,1] \longrightarrow \mathbb{R}$. Since $[0,1]$ is compact, the set $C[0,1] \subset L_\infty$.

Example B.8. The set of all $m \times n$ matrices is a linear space with addition and multiplication defined in the usual way.

B.2 Linear Independence and Dimension

Definition B.9. *A* **linear combination** *of the elements $x_1, x_2, \ldots, x_n$ of a linear space L is an element of the form $\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_n x_n$ for some real numbers $\lambda_1, \lambda_2, \ldots, \lambda_n$. A linear combination is* **nontrivial** *if not all of the λ_i are zero.*

Definition B.10. *The distinct elements $x_1, x_2, \ldots, x_n$ of a linear space L are said to be* **linearly dependent** *if there exists a nontrivial linear combination of the x_i that equals zero. If no such linear combination exists, then the elements $x_1, x_2, \ldots, x_n$ are said to be* **linearly independent**.

Definition B.11. *A linear space L is said to be n-***dimensional or of dimension** *n if there exists n linearly independent elements in L, but any $n+1$ distinct elements in L are linearly dependent. If n linearly independent elements can be found in L for every integer n, then L is said to be* **infinite-dimensional**; *otherwise, L is said to be* **finite dimensional**. *Any set of n linearly independent elements of an n-dimensional space is called a* **basis** *of L.*

Example B.12. $\mathbb{R}^n$ is n-dimensional.

Example B.13. All of the examples of function spaces are infinite-dimensional.

B.3 Normed Linear Spaces

B.3.1 Definition

Definition B.14. *A* **norm** *defined on a linear space L is a finite real-valued mapping on L, denoted by the symbol $||\ ||$, that possesses the following properties:*

a) $||x|| \geq 0$ *for all* $x \in L$ *and* $||x|| = 0$ *if and only if* $x = 0$.
b) $||\alpha x|| = |\alpha|||x||$ *for all* $x \in L$ *and* $\alpha \in \mathbb{R}$.
c) $||x + y|| \leq ||x|| + ||y||$ *for all* $x, y \in L$ *(***Triangle Inequality***).*

Remark B.15. A function $f : L \longrightarrow \mathbb{R}$ is **convex** if $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$ for every $x, y \in L$ and $\lambda \in [0,1]$. Properties (b) and (c) imply that a norm is a convex function. A convex function that satisfies property (b) automatically satisfies property (c).

B.3.2 Examples

Here are some notable examples of normed linear spaces.

Example B.16. The absolute value function is a norm defined on the real line.

Example B.17. On $\mathbb{R}^n$ three common norms are used:

- ℓ_1 **norm** defined by $||x|| := \sum_{i=1}^{n} |x_i|$.
- ℓ_2 **norm** defined by $||x|| := \sqrt{\sum_{i=1}^{n} x_i^2}$. Here, one must verify the inequality

$$\sqrt{\sum_{i=1}^{n}(x_i + y_i)^2} \leq \sqrt{\sum_{i=1}^{n} x_i^2} + \sqrt{\sum_{i=1}^{n} y_i^2}\ ,$$

  also known as the *Minkowski inequality for* $p = 2$. By squaring both sides, the inequality holds if and only if the **Cauchy-Schwarz inequality**

$$\Big(\sum_{i=1}^{n} x_i y_i\Big)^2 \leq \Big(\sum_{i=1}^{n} x_i^2\Big)\Big(\sum_{i=1}^{n} y_i^2\Big)$$

  holds. The Cauchy-Schwarz inequality follows here since it can be readily verified that

$$\Big(\sum_{i=1}^{n} x_i y_i\Big)^2 = \Big(\sum_{i=1}^{n} x_i^2\Big)\Big(\sum_{i=1}^{n} y_i^2\Big) - 1/2 \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i y_j - x_j y_i)^2.$$

- **sup norm** defined by $||x|| := \max\limits_{1 \leq i \leq n} |x_i|$.

Example B.18. For $p \geq 1$ let L_p denote the set of all functions $f : \mathbb{R} \longrightarrow \mathbb{R}$ such that $\int_{-\infty}^{\infty} |f(x)|^p \, dx < \infty$. The L_p**-norm** defined on L_p is

$$||f||_p := \left(\int_{-\infty}^{\infty} |f(x)|^p \, dx\right)^{1/p}.$$

It can be shown that L_p is a linear space and that $||f||_p$ is, in fact, a norm.[2]

[2] When $p = 1$ this is simple. When $p = 2$ one establishes the integral version of Cauchy-Schwarz inequality.

B.4 Metric Spaces

The operation of taking limits is used throughout these notes, and is a fundamental operation of mathematical analysis in general. The concept of limit is inexorably tied to a notion of "closeness," which can make sense if the underlying space is equipped with a distance or "metric" to measure the degree of closeness between points in the space. In this section, we make precise the notion of distance for a normed linear space and establish fundamental properties about limits.[3]

B.4.1 Definition

The **Euclidean distance** between two points in $X = \mathbb{R}^n$ is defined as

$$d(x,y) := \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}. \tag{B.1}$$

It equals the ℓ_2-norm of the vector $x - y$. Given the definition of a norm, it follows immediately that this distance function $d(\cdot,\cdot)$ defined on $X \times X$ has the following three properties:

a) $d(x,y) \geq 0$ for every $x, y \in X$ and $d(x,y) = 0$ if and only if $x = y$;
b) $d(x,y) = d(y,x)$ for every $x, y \in X$ (**Symmetry**);
c) $d(x,z) \leq d(x,y) + d(y,z)$ for every $x, y, z \in X$ (**Triangle Inequality**).[4]

Definition B.19. *A function $d(\cdot,\cdot)$ defined on a linear space X that satisfies properties (a)-(c) above is called a* **metric**, *the number $d(x,y)$ is called the* **distance** *between points $x, y \in X$, and the pair (X,d) is called a* **metric space**. *The set X is said to be* **equipped** *with the metric d.*

Every normed linear space X with norm $||\cdot||$ becomes a metric space (X,d) by defining the metric $d(x,y) := ||x - y||$ for every $x, y \in X$. The metric d so defined will be referred to as the *metric induced* from this norm. Given a normed linear space X with norm $||\cdot||$ and induced metric d, it is natural to measure the degree of closeness of two points x and y by the distance $d(x,y) = ||x - y||$ and to say that x and y are "close" if $d(x,y) < \epsilon$ for "sufficiently small" ϵ.

The same set X can be "metrized" in a number of *different* ways. For example, consider $\mathbb{R}^N$ equipped with the metrics d_1, d_2, and d_∞ respectively induced from the ℓ_1, ℓ_2, or ℓ_∞ norms. What determines the degree of closeness between points depends on the metric chosen.

In the subsections to follow, X shall denote a nonempty metric space and S shall denote a subset of X.

[3] Several definitions, concepts, and results introduced in this section apply to general topological spaces.

[4] Follows from the triangle inequality for a norm and the identity $||x - z|| = ||(x - y) + (y - z)||$.

B.4.2 Open and Closed Sets

Definition B.20. *Let $x \in X$ and let r be a positive real number. The set*

$$B(x,r) := \{y \in X : \ d(x,y) < r\}$$

is called the **open ball of radius r centered at x,** *and the set*

$$B[x,r] := \{y \in X : \ d(x,y) \leq r\}$$

is called the **closed ball of radius r centered at x.**

Definition B.21. *Let $x \in X$. A set $U \subset X$ is an* **open neighborhood of x** *if $U = B(x,r)$ for some r.*

Definition B.22. *S is said to be* **open** *if for every $x \in S$ there exists an open neighborhood U of x such that $U \subset S$. A set S is said to be* **closed** *if its complement $X \backslash S = \{y \in X : \ y \notin S\}$ is open.*

Example B.23. The subset $S = (0,1)$ of the real line is open, whereas the subset $S = (0,1]$ is not, since every neighborhood of one contains a point not in S. The subset $S = [0,1]$ is closed.

Example B.24. The infinite intervals $(-\infty, \infty), (a, \infty), (-\infty, b)$ are open.

Example B.25. The subset $S = \{0, 1, \frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{7}{8}, \frac{1}{16}, \frac{15}{16}, \ldots, \}$ of the real line is closed.

Proposition B.26. *Open and closed sets have the following union and intersection properties:*

a) A union of a arbitrary family of open sets is open.
b) An intersection of a finite number of open sets is open.
c) An intersection of a arbitrary family of closed sets is closed.
d) A union of a finite number of closed sets is closed.

The following proposition characterizes the open sets on the real line equipped with the absolute value metric.

Proposition B.27. *Every open set of the real line is the union of a finite or countable number of pairwise disjoint open intervals.*

B.4.3 Closure and Boundary

Definition B.28. *A point x is said to be a* **limit point** *of the set S if each of its neighborhoods (minus the point itself) meets S; that is, $S \cap (U \backslash \{x\}) \neq \emptyset$ for every neighborhood U of x. The symbol S' will denote the set of limit points of the set S.*

A limit point of a set may or may not belong to the set, as the following example illustrates.

Example B.29. The points 0 and 1 are the only two limits points for the set S defined in Example B.25. If these two points are *removed* from S, each point will still be a limit point of S.

Definition B.30. *The* **closure** *of a set S is the union of S with S'. The symbol $cl(S)$ will denote the closure of S.*

Proposition B.31. *The closure operation has the following properties:*

a) $cl(S)$ *is closed.*
b) $cl(S) \subset cl(T)$ *whenever* $S \subset T$.
c) $cl(S)$ *is the smallest closed set containing* S.
d) S *is closed if and only if* $S = cl(S)$.
e) $cl(cl(S)) = cl(S)$.
f) $cl(S \cup T) = cl(S) \cup cl(T)$.

Remark B.32. Part (e) implies that the closure operation only needs to be performed once, and part (f) implies that the closure of a finite union of sets is the union of the closures of each set. By part (b), the closure of a finite intersection of sets is a subset of the intersection of the closures of each set; however, the former set can be a *strict* subset of the latter set. Consider, for example, the sets $S = [0, 1/2)$ and $T = [1/2, 1]$: $cl(S \cap T) = \emptyset$ but $cl(S) \cap cl(T) = \{1/2\}$.

Definition B.33. *The* **boundary** *of the set S is the set of all points $x \in S$ for which each of its neighborhoods meets both S and its complement $X \backslash S$. The symbol ∂S will denote the boundary of S. A point $x \in \partial S$ is said to* **belong or lie on the boundary** *of S.*

Example B.34. Figure B.4.3 depicts the same triangle with some, none, or all of its boundary included in the set. In (a), none of its boundary is included, and therefore the triangle is an open set; in (b), all of the boundary is included, and therefore the triangle is a closed set; finally, in (c), some but not all of the boundary is included, and so the triangle is neither open nor closed.

Proposition B.35. *The boundary of S possesses the following properties:*

a) $\partial S = \partial(X \backslash S)$.
b) $cl(S) = S \cap \partial S$.
c) S *is closed if and only if* $\partial S \subset S$.
d) S *is open if and only if* $\partial S \cap S = \emptyset$.
e) $\partial S = cl(S) \cap cl(X \backslash S)$.

Definition B.36. *The* **interior** *of the set S is the set $S \backslash \partial S$, namely, the set of all points in S that do not lie on the boundary of S. The symbol $int(S)$ will denote the interior of S.*

(a) Open Set

(b) Closed Set

(c) Not Open - Not Closed

Proposition B.37. *The interior of S has the following properties:*

a) $int(S)$ is open.
b) $int(S)$ is the largest open set contained within S.

B.4.4 Convergence and Limits

Definition B.38. *An infinite sequence $\{x_n\} \subset X$ is said to* **converge** *to a point $x \in X$ if for every $\epsilon > 0$ there exists an N_ϵ such that $x_k \in B(x, \epsilon)$ for all $k > N_\epsilon$. The notation $x_n \longrightarrow x$ will be used to denote this convergence or $x_n \xrightarrow{d} x$ will be used when it is important to emphasize the metric. The point x is called the* **limit point** *of the convergent sequence. An infinite sequence $\{x_n\} \subset X$ that does not converge is said to* **diverge**.

Proposition B.39. *Convergent sequences of a metric space have the following properties:*

a) The limit point of a convergent sequence is unique.
b) If a sequence $\{x_n\}$ converges to x, then so does every subsequence of $\{x_n\}$.

The following proposition provides an equivalent, useful definition of a closed set.

Proposition B.40. *S is closed if and only if whenever the infinite sequence $\{x_n\} \subset S$ converges, its limit point belongs to S.*

Proposition B.41. *On $\mathbb{R}^n$ $x_n \xrightarrow{d_1} x \iff x_n \xrightarrow{d_2} x \iff x_n \xrightarrow{d_\infty} x$.*[5]

Remark B.42. This proposition shows that a metric is not an intrinsic tool for analyzing convergence properties of sequences.

B.4.5 Completeness

Often, one wishes to establish that a particular sequence converges. To show that a sequence converges in a *complete* metric space, it is sufficient to show that eventually all of the remaining points in the sequence can be made arbitrarily close to one another. In a complete metric space, one does not have to know what the limit is in order to prove that it exists.

[5] The d_1, d_2 and d_∞ metrics are the ones respectively induced from the ℓ_1, ℓ_2 and ℓ_∞ norms.

Definition B.43. *Let (X, d) be a metric space. A sequence of points $\{x_n\} \subset X$ is said to be a* **Cauchy sequence** *if for every $\epsilon > 0$, there exists an integer N_ϵ such that $d(x_n, x_m) < \epsilon$ for all $n, m > N_\epsilon$.*

Remark B.44. For a Cauchy sequence, $\lim_{m,n \longrightarrow \infty} d(x_n, x_m) = 0$.

Definition B.45. *A metric space (X, d) is said to be* **complete** *if every Cauchy sequence converges to a point in X.*

Example B.46. Here are some notable examples of complete metric spaces:

- $\mathbb{R}^n$.
- The L_p spaces.
- The space of real-valued continuous functions of one variable defined on a closed and bounded set under the sup norm.

Proposition B.47. *A closed subset of a complete metric space is itself a complete metric space.*

B.4.6 Compactness

Definition B.48. *A family $\mathcal{F} = \{U_\alpha\}$ is a* **cover** *of S if $S \subset \cup_\alpha U_\alpha$. If every set in a cover is open, then the cover is said to be an* **open cover**. *A* **subcover** *of a cover is a subfamily of $\mathcal{F}$ that also covers S.*

Definition B.49. *S is compact if every open cover of S has a finite subcover.*

Definition B.50. *The* **diameter** *of S is defined to be*

$$diam(S) := \sup\{d(x, y) : \ x, y \in S\}.$$

Definition B.51. *S is* **bounded** *if its diameter $diam(S)$ is finite.*

Remark B.52. If X is a normed linear space and d is the metric induced from the norm on X, then $S \subset X$ is bounded if $S \subset B(0, r)$ for some finite real number r.

The following theorem provides an equivalent definition of compactness in $\mathbb{R}^n$ equipped with any of the three metrics d_1, d_2 or d_∞.

Theorem B.53. Heine-Borel Theorem *Let $S \subset \mathbb{R}^n$. S is compact if and only if S is closed and bounded.*

Corollary B.54. *A closed subset of a compact set in $\mathbb{R}^n$ is itself compact.*

One of the most useful properties of compact subsets of complete metric spaces (such as $\mathbb{R}^n$) is given in the following theorem.

Theorem B.55. *Let X be a complete metric space and let $S \subset X$ be compact. Every infinite sequence $\{x_n\} \subset S$ has a convergent subsequence.*

B.4.7 Continuity

Definition B.56. *Let (X, d) be a metric space and let $f : X \longrightarrow \mathbb{R}$. The real-valued function f defined on X is said to be* **continuous at the point** *$x_0 \in X$ if for every $\epsilon > 0$ there exists a $\delta > 0$ such that*

$$|f(x) - f(x_0)| < \epsilon \text{ whenever } d(x, x_0) < \delta.$$

The mapping $f(\cdot)$ is said to be **continuous on** *X if it is continuous at every point $x_0 \in X$.*

Remark B.57. This definition is consistent with the usual definition of a continuous real-valued function of one variable. In this case, $X = S \subset \mathbb{R}$ and the metric d is the absolute value metric. In this setting, $f(\cdot)$ is continuous at x_0 if for every $\epsilon > 0$ there exists a $\delta > 0$ such that

$$|f(x) - f(x_0)| < \epsilon \text{ whenever } |x - x_0| < \delta.$$

More generally, if X is a subset of $\mathbb{R}^n$ equipped with the metric d induced from the usual norms (Euclidean, sup, or ℓ_1), then $|x - x_0|$ is replaced with $||x - x_0||$.

Definition B.58. *Let (X, d) be a metric space and let $f : X \longrightarrow \mathbb{R}$.*

- *The real-valued function $f(\cdot)$ defined on X is said to be* **upper semicontinuous at the point** *$x_0 \in X$ if for every $\epsilon > 0$ there exists a $\delta > 0$ such that*
$$f(x) - f(x_0) < \epsilon \text{ whenever } d(x, x_0) < \delta.$$
The mapping $f(\cdot)$ is said to be **upper semicontinuous on** *X if it is upper semicontinuous at every point $x_0 \in X$.*
- *The real-valued function $f(\cdot)$ defined on X is said to be* **lower semicontinuous at the point** *$x_0 \in X$ if for every $\epsilon > 0$ there exists a $\delta > 0$ such that*
$$f(x) - f(x_0) > \epsilon \text{ whenever } d(x, x_0) < \delta.$$
The mapping $f(\cdot)$ is said to be **lower semicontinuous on** *X if it is lower semicontinuous at every point $x_0 \in X$.*

The following proposition is central to our study, and so we shall supply the proof.

Proposition B.59. *Let (X, d) be a metric space and let $f : X \longrightarrow \mathbb{R}$.*

a) The mapping $f(\cdot)$ is upper semicontinuous on X if and only if every upper level set $L_f^{\geq}(\alpha) = \{x \in X : f(x) \geq \alpha\}$ is closed.
b) The mapping $f(\cdot)$ is lower semicontinuous on X if and only if every lower level set $L_f^{\leq}(\alpha) = \{x \in X : f(x) \leq \alpha\}$ is closed.

Proof. We shall only prove part (a), as the proof of part (b) follows by considering the negative of the function $f(\cdot)$.

Suppose first that f is upper semicontinuous. We must show each upper level set is closed. Pick an $\alpha \in \mathbb{R}$ for which the upper level set $L_f^{\geq}(\alpha)$ is nonempty,[6] and let $\{x_n\}$ be an infinite sequence of points in the upper level set that converges to the point x_0. We must show that $x_0 \in L_f^{\geq}(\alpha)$. Pick an $\epsilon > 0$ and use the upper semicontinuity of $f(\cdot)$ at x to find a $\delta > 0$ such that $f(x) - f(x_0) < \epsilon$ whenever $d(x, x_0) < \delta$. Since $x_n \longrightarrow x_0$ eventually $d(x_n, x_0) < \delta$ for all n sufficiently large. This means that $f(x_n) < f(x_0) + \epsilon$ for all n sufficiently large. Since $f(x_n) \geq \alpha$ for every n and since ϵ was chosen arbitrarily, it now follows that $f(x_0) \geq \alpha$, as required. (If $f(x_0) < \alpha$, then $\epsilon = (\alpha - f(x_0))/2$ will lead to a contradiction.)

Now suppose each upper level set is closed. Pick an $x_0 \in X$. We must show $f(\cdot)$ is upper semicontinuous at x_0. Pick an $\epsilon > 0$ and set $\alpha := f(x_0)$. Since $x_0 \notin L_f^{\geq}(\alpha + \epsilon)$, a closed set, there must exist a $\delta > 0$ such that the open ball $B(x_0, \delta)$ centered at x_0 does not intersect $L_f^{\geq}(\alpha+\epsilon)$. But this means that $f(x) < \alpha + \epsilon = f(x_0) + \epsilon$ whenever $d(x, x_0) < \delta$, which establishes upper semicontinuity. □

The following proposition makes clear the connection between these different concepts of continuity. The proof is an immediate consequence of the definitions.

Proposition B.60. *Let (X, d) be a metric space and let $f : X \longrightarrow \mathbb{R}$. The mapping $f(\cdot)$ is continuous on X if and only if it is both upper semicontinuous and lowersemicontinous on X.*

Theorem B.61. *Let (X, d) be a metric space and let $f : X \longrightarrow \mathbb{R}$. If $S \subset X$ is compact, then $f(S)$ is compact.*

Remark B.62. This theorem is true for arbitrary topological spaces X and Y. That is, if $f : X \longrightarrow Y$ and $S \subset X$ is compact, then $f(S) \subset Y$ is compact.

Definition B.63. *Let (X, d) for a metric space, let $f : X \longrightarrow \mathbb{R}$ and let $S \subset X$.*

- *The function $f(\cdot)$ is said to* **attain its minimum on** *S if there exists an $x_m \in S$ such that*
$$f(x_m) = \inf\{f(x) : \ x \in S\}.$$
By definition of infimum $f(x_m) \leq f(x)$ for all $x \in S$.
- *The function $f(\cdot)$ is said to* **attain its maximum on** *S if there exists an $x_M \in S$ such that*
$$f(x_M) = \sup\{f(x) : \ x \in S\}.$$
By definition of supremum $f(x_M) \geq f(x)$ for all $x \in S$.

[6] Recall that the empty set is closed.

Theorem B.64. *Let (X, d) for a metric space and let $f : X \longrightarrow \mathbb{R}$. If $S \subset X$ is compact, then $f(\cdot)$ attains both its minimum and maximum on S.*

Remark B.65. In light of this theorem, when S is compact it is acceptable to replace the "inf" with a "min" and write $\min\{f(x) : x \in S\}$ and the "sup" with a "max" and write $\max\{f(x) : x \in S\}$.

B.4.8 Connectedness

Definition B.66. *Two sets A and B are said to be* **separated** *if $A \neq B \neq \emptyset$ and $cl(A) \cap B = A \cap cl(B) = \emptyset$.*

Definition B.67. *S is said to be* **connected** *if there does not exist a decomposition $S = A \cup B$ such that A and B are separated.*

The following proposition characterizes connected subsets of the real line equipped with the absolute value metric.

Proposition B.68. *A nonempty subset $S \subset \mathbb{R}$ is connected if and only if it consists of a single element or is an interval.*

As the following theorem shows, continuous functions preserve the property of connectedness.

Theorem B.69. *Let $f : X \longrightarrow \mathbb{R}$ be continuous on X. If $S \subset X$ is connected, then $f(S)$ is a connected subset of the real line.*

B.5 Bibliographical Notes

Classic references are Kelly [1955], Kolmogorov and Fomin [1970], and Royden [1968]. An accessible but less general treatment of these concepts can be found in the excellent text by Bazarra et. al. [1993].

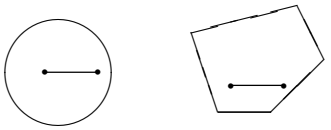
C

Convex Sets

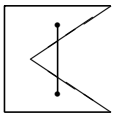
Convexity is essential to modern economic analysis. In this chapter, we establish basic properties about convex sets and the basic separation theory in Euclidean space (finite-dimensional spaces) used throughout this book.

C.1 Definition and Examples

Geometrically, a set $S \subseteq \mathbb{R}^n$ is convex if it contains the line segment joining any of its two points. See Figure C.1.



(d) Convex Sets



(e) A nonconvex set

Fig. C.1. Example of a convex set and a nonconvex set.

Definition C.1. *A set $S \subseteq \mathbb{R}^n$ is* **convex** *if $\lambda x_1 + (1-\lambda)x_2 \in S$ for each $\lambda \in [0,1]$ and $x_1, x_2 \in S$.*

Example C.2. The following sets are examples of convex sets:

- $\{(x_1, x_2) \in \mathbb{R}^2 : \ x_1^2 + x_2^2 \leq 9\}$. It is the collection of points that lie on or inside the circle with center $(0,0)$ and radius 3.

- $\mathcal{B}_\epsilon(x) := \{y \in \mathbb{R}^n : \ ||y-x|| \leq \epsilon$, the closed ball of radius ϵ about $x \in \mathbb{R}^n$. Here, $||\cdot||$ denotes the Euclidean norm. Convexity follows from the triangle inequality of the norm.
- $\{x \in \mathbb{R}^n : \ Ax \leq b\}$, where A is an $m \times n$ matrix and $b \in \mathbb{R}^m$. (One may replace $\leq$ with $<$, $=$, $\geq$ or $>$.)

An important class of convex sets are hyperplanes and their associated closed and open halfspaces.

Definition C.3. *A* **hyperplane** *in* $\mathbb{R}^n$ *is the set defined by*

$$H(p, \alpha) := \{x \in \mathbb{R}^n : \ p \cdot x = \alpha\}$$

for some nontrivial $p \in \mathbb{R}^n$ *and real number* α*. The vector* p *is called the* **normal** *to the hyperplane. Each hyperplane* $H(p, \alpha)$ *induces four* **halfspaces** *defined as:*

$H^{\geq}(p, \alpha) := \{x : \ p \cdot x \geq \alpha\}$ **(all points lying on or above the hyperplane)**.
$H^{>}(p, \alpha) := \{x : \ p \cdot x > \alpha\}$ **(all points lying above the hyperplane)**.
$H^{\leq}(p, \alpha) := \{x : \ p \cdot x \leq \alpha\}$ **(all points lying on or below the hyperplane)**.
$H^{<}(p, \alpha) := \{x : \ p \cdot x < \alpha\}$ **(all points lying below the hyperplane)**.

Definition C.4. *A* **convex combination** *of points* $x_1, x_2, \ldots, x_N$ *is a weighted average of the form* $\sum_{i=1}^N \lambda_i x_i$ *in which* $\sum_{i=1}^N \lambda_i = 1$ *and each* λ_i *is nonnegative.*

A routine induction argument shows that a convex combination of points in a convex set S also belong to S.[1]

Proposition C.5. *The following sets are convex:*

(a) An intersection of a family of convex sets.
(b) A finite algebraic sum of convex sets.
(c) The closure of a convex set.
(d) The interior of a convex set.

C.2 Convexification

Often one is given a collection of points S that is not convex, and then "convexifies" this set by adding to it all convex combinations of its elements.

Definition C.6. The **convex hull** of a set S is the collection of all convex combinations of points in S. It will be denoted by $Conv(S)$.

Proposition C.7. *$Conv(S)$ is the intersection of all convex sets that contain S.*

[1] Note that the $\sum_{i=1}^{n+1} \lambda_i x_i$ can be expressed as $\lambda z + (1-\lambda)y$, where $\lambda = \sum_{i=1}^n \lambda_i$, $z = \sum_{i=1}^n (\lambda_i/\lambda) x_i$ and $y = x_{n+1}$.

Remark C.8. $Conv(S)$ is the *smallest* convex set containing S. That is, if F is convex and $S \subseteq F$ then $Conv(S) \subseteq F$.

Proof. Let $\mathcal{S}$ denote the intersection of all convex sets F that contain S. It is convex by Proposition C.5 (a). It immediately follows that $\mathcal{S}$ is the smallest convex set containing S. In particular, $\mathcal{S} \subseteq Conv(S)$ since $Conv(S)$ is itself a convex set that contains S. Since an arbitrary convex set F that contains S includes all convex combinations of points in S, $Conv(S) \subseteq F$, from which it follows that $Conv(S) \subseteq \mathcal{S}$. Thus, $Conv(S) = \mathcal{S}$, as claimed. □

C.3 Separation of a Convex Set and a Point

C.3.1 Strict Separation

Lemma C.9. *Let X be a nonempty, closed subset of $\mathbb{R}^n$. If X does not contain the origin, then there exists a nontrivial point $x^* \in X$ that minimizes the Euclidean distance from X to the origin, i.e.,*

$$0 \ < \ x^* \cdot x^* \ \leq \ x \cdot x \text{ for all } \ x \in X.$$

Proof. Pick a nonzero $z \in X$, and let E denote the collection of points in X whose norm is bounded by the norm of z. E is closed and bounded and hence compact. Since the (quadratic) function $f(x) = x \cdot x$ is continuous, it achieves its minimum over E, say at x^*. Clearly, x^* satisfies the requisite properties. □

Theorem C.10. Strict Separation Theorem *Let X be a nonempty, closed convex subset of $\mathbb{R}^n$ that does not contain the origin. Then there exists a (nontrivial) $p \in \mathbb{R}^n$ and an $\alpha > 0$ such that $0 < \alpha < p \cdot x$ for all $x \in X$.*

In Theorem C.10, the hyperplane $H(p, \alpha)$ is said to **strictly separate** the origin from X. That is, X is contained in the closed halfspace $H^{\geq}(p, \alpha)$ but the origin does not belong to this closed halfspace.

Proof. Lemma C.9 guarantees the existence of a nonzero point $a \in X$ for which

$$a \cdot a \leq ((1-t)a + tb) \cdot ((1-t)a + tb) = (a \cdot a) + 2ta \cdot (b - a) + t^2(b \cdot b)$$

for all $b \in X$ and $t \in [0, 1]$. Since the function $f(t) := 2ta \cdot (b - a) + t^2(b \cdot b)$ is nonnegative on $[0, 1]$ and since $f(0) = 0$, its derivative at zero, $2a \cdot (b - a)$, must be nonnegative. Thus, $a \cdot b \geq a \cdot a > 0$, and the result follows by taking p to be a and setting $\alpha = p \cdot p/2$. □

Corollary C.11. *Let X be a nonempty, closed convex subset of $\mathbb{R}^n$. For each $x_0 \notin X$ there exists a (nontrivial) $p_0 \in \mathbb{R}^n$ and $\gamma \in \mathbb{R}$ such that $p_0 \cdot x_0 < \gamma < p_0 \cdot x$ for all $x \in X$.*

In Corollary C.11, the hyperplane $H(p_0, \gamma)$ is said to **strictly separate** x_0 from X. That is, X is contained in the closed halfspace $H^{\geq}(p_0, \gamma)$ but x_0 does not belong to this closed halfspace.

Proof. Apply Theorem C.10 to the set $X - x_0 = \{y : y + x_0 \in X\}$. Set $\gamma := \alpha + p \cdot x_0$. □

Corollary C.12. *A closed convex set $X \subset \mathbb{R}^n$ is the intersection of all closed halfspaces that contain it.*

Proof. Let S denote the intersection of all closed halfspaces that contain X. Clearly, $X \subset S$. If $X \neq S$, then there exists a $z \in S$ such that $z \notin X$. However, the strict separation theorem guarantees existence of a closed halfspace that contains X but not z, which implies $z \notin S$, a contradiction. □

C.3.2 Supporting Hyperplanes

Theorem C.13. Supporting Hyperplane Theorem *Let X be a nonempty closed convex subset of $\mathbb{R}^n$. For each x_0 that lies on the boundary of X there exists a nontrivial $p_0 \in \mathbb{R}^n$ such that $p_0 \cdot x_0 \leq p_0 \cdot x$ for all $x \in X$.*

Remark C.14. Equivalently, in terms of sets, $X \subset H^{\geq}(p_0, p_0 \cdot x_0)$ and $X \cap H(p_0, p_0 \cdot x_0) \neq \emptyset$. The hyperplane $H(p_0, p_0 \cdot x_0)$ is said to **support** X at x_0, hence the name of the theorem.

Proof. Since x_0 lies on the boundary of X, it is possible to find an infinite sequence of points $\{x_n\}$ not in X that converges to x_0. Corollary C.11 guarantees the existence of a sequence of nontrivial $\{p_n\}$'s for which

$$p_n \cdot x_n \leq p_n \cdot z \text{ for all } z \in X.$$

Let $\hat{p}_n = p_n / ||p_n||$. Since the $\hat{p}_n$'s belong to the boundary of the unit ball, which is compact, it is possible to extract a convergent subsequence $\{\hat{p}_{n_k}\} \to p_0$. Clearly, p_0 is nonzero since its norm is one. Moreover, since the inner product is continuous $\hat{p}_{n_k} \cdot x_{n_k} \to p_0 \cdot x_0$ and $\hat{p}_{n_k} \cdot z \to p_0 \cdot z$ for each $z \in X$. The result now follows from the easily established fact that if $\{a_n\}$ and $\{b_n\}$ are two infinite sequences of numbers such that (i) $a_n \leq b_n$ for all n and (ii) $a_n \to a$ and $b_n \to b$, then $a \leq b$. □

C.3.3 Polar Cones

Definition C.15. *The* **polar cone** *of a set $S \subset \mathbb{R}^n$ is defined as*

$$S^* := \{p \in \mathbb{R}^n : \ p \cdot x \leq 0 \ \text{ for all } \ x \in S\}.$$

If S is empty, then $S^* := \mathbb{R}^n$.

A cone that is also convex is called, quite naturally, a **convex cone**. The following characterization of nonempty, closed convex cones is at the heart of the dual characterizations of technology.

Theorem C.16. *If S is a nonempty closed convex cone, then $S = S^{**}$.*

Proof. By definition,

$$S^* = \{p : \ p \cdot x \leq 0 \text{ for all } x \in S\},$$
$$S^{**} = \{y : \ y \cdot p \leq 0 \text{ for all } p \in S^*\}.$$

Inspection of these two definitions shows that $S \subset S^{**}$. To establish the reverse inclusion $S^{**} \subset S$, we shall show that if $x \notin S$, then $x \notin S^{**}$. If $x \notin S$, then by Theorem C.10 there exists a nontrivial p for which $p \cdot x > p \cdot z$ for all $z \in S$. Since S is a cone, it must be the case that $p \cdot z \leq 0$ for all $z \in S$, which shows that $p \in S^*$. Since S is also closed, the origin belongs to S, which implies that $p \cdot x > 0$. Thus, $x \notin S^{**}$, as claimed. □

C.4 Polyhedra

Corollary C.12 shows that any closed convex set is the intersection of all closed halfspaces that contain it. For a polyhedron only a finite number of closed halfspaces is needed to represent it.

C.4.1 Definition and Examples

Definition C.17. *A* **polyhedron** $S \subset \mathbb{R}^k$ *is a finite intersection of closed halfspaces (and hence closed and convex). That is,*

$$S = \{x \in \mathbb{R}^k : \ a_i \cdot x \leq b_i, \ 1 \leq i \leq M\},$$

where the $a_i \in \mathbb{R}^k$ are nontrivial and each $b_i \in \mathbb{R}$.

Since an equation $a \cdot x = b$ can be represented via two inequalities, namely, $a \cdot x \geq b$ and $a \cdot x \leq b$, a polyhedron can be expressed via a finite number of inequalities and/or equalities.

Example C.18. Let A be an $n \times m$ matrix and let $b \in \mathbb{R}^m$. Sets of the form $\{x \in \mathbb{R}^n : \ Ax \leq b\}$, $\{x \in \mathbb{R}^n : \ Ax = b, \ x \geq 0\}$ or $\{x \in \mathbb{R}^n : \ Ax \geq b, \ x \geq 0\}$ are examples of polyhedron. Since $\mathbb{R}^n_+ = \{x \in \mathbb{R}^n : \ x \geq 0\}$, the last set, for example, can be expressed as $\{x \in \mathbb{R}^n_+ : \ Ax \geq b\}$.

Definition C.19. *A* **polytope** *is a bounded polyhedron (and hence compact and convex).*

Example C.20. The subset of the plane defined by

$$\{(x_1, x_2) : \ x_1 + x_2 \leq 5, \ 2x_1 + x_2 \leq 9, \ x_1 + 2x_2 \leq 9, x_1 \geq 0, \ x_2 \geq 0\}$$

is a polytope. See Figure C.2.

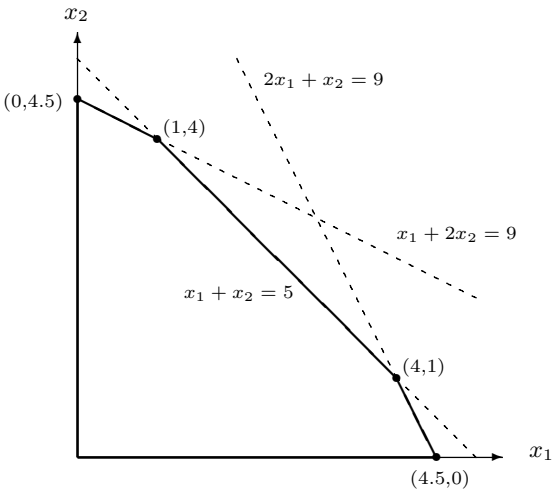


Fig. C.2. Example of a polytope.

C.4.2 Extreme Points and Directions

Definition C.21. *Let C be a convex set. A point $c \in C$ is an* **extreme point** *of C if it cannot be represented as a convex combination of two other points in C. Equivalently, c is an extreme point if whenever $c = \lambda c_1 + (1-\lambda)c_2$ with $c_1, c_2 \in C$ and $\lambda \in (0,1)$, then $c_1 = c_2 = c$.*

Example C.22. In Figure C.2, the points $(0, 4.5)$, $(1, 4)$, $(4, 1)$ and $(4.5, 0)$ are the extreme points of this polytope.

Each point in the polytope in Figure C.2 can be represented as a convex combination of the four extreme points. This is true for polytopes, since they are *compact*, but not true for unbounded polyhedron, as the following example illustrates.

Example C.23. Consider the subset of the plane defined by

$$C := \{(x_1, x_2) : \ -x_1 + 2x_2 \leq 0,\ 2x_1 - x_2 \leq 0,\ x_1 \geq 0,\ x_2 \geq 0\}$$

and depicted in Figure C.3. C is a convex cone with a *single* vertex, namely, the origin.

Obviously, the set C in Figure C.3 cannot be represented as a convex combination of its extreme points. However, each point in C *can* be represented as a nonnegative linear combination of the vectors $d_1 := (2, 1)$ and $d_2 := (1, 2)$, i.e., for each $c \in C$ there exists $\mu_1, \mu_2 \geq 0$ such that $c = \mu_1 d_1 + \mu_2 d_2$. The d_i here are called extreme directions of C.

Definition C.24. *Let $S \subset \mathbb{R}^n$ be nonempty and closed.*

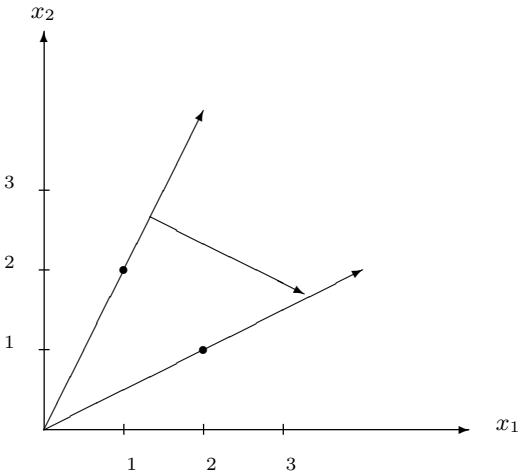


Fig. C.3. Example of an unbounded polyhedron.

- *A nontrivial vector $d \in \mathbb{R}^n$ is a* **direction** *of S if for each $x \in S$ the set $\{x + \lambda d : \lambda \geq 0\}$ is contained within S.*
- *Two directions d_1 and d_2 are* **distinct** *if there does not exist a positive α for which $d_1 = \alpha d_2$.*
- *A direction d of S is an* **extreme direction** *if it cannot be expressed as a positive linear combination of two distinct directions.*

Geometrically, d is a direction of S if for each $x \in S$ the ray emanating from x and passing through d also belongs to S. Obviously, a direction is unique up to a positive scalar multiplication, hence the definition of distinct direction. Algebraically, an extreme direction d has the property that whenever $d = \lambda_1 d_1 + \lambda_2 d_2$ for two directions d_1, d_2 with $\lambda_1, \lambda_2 > 0$, then $d_1 = \alpha d_2$ for some positive α.

C.4.3 Characterization of Extreme Points and Directions

It will be useful for the development to follow to define a standard format for a polyhedron.

Definition C.25. *The* **standard format** *for a polyhedron is the representation given by $\{x : Ax = b,\ x \geq 0\}$. The matrix A is of full rank.*

Remark C.26. By removing redundant equations, we may assume the rank of A is m when A is an $m \times n$. Moreover, $m \leq n$.

Every polyhedron S can be equivalently expressed in the standard format. For example, the inequality $a \cdot x \leq b$ can be transformed to an equality by adding a nonnegative slack variable s so that $a \cdot x + s = b$. Similarly, the inequality

$a \cdot x \geq b$ can be transformed into an equality by adding a nonnegative slack variable t so that $a \cdot x - t = b$. In either case, note that the dimension of x increases by one and a column corresponding to the coordinate vector e_{k+1} is added to the end of the original matrix A. In this book, the economic variables of interest are almost always nonnegative. However, if a variable, say x_j, is originally "unrestricted," namely, it is not constrained to be nonnegative, then one simply replaces it with the difference of two nonnegative variables $x_j^+ - x_j^-$. Again, the dimension of x will increase by one and the j^{th} column of A is multiplied by minus one and inserted into the original matrix A next to the j^{th} column.

Definition C.27. *Let S be a nonempty polyhedron in standard format. Suppose the columns of A can be permuted so that $A = [B \quad N]$ such that the submatrix B is an $m \times m$ invertible matrix satisfying $B^{-1}b \geq 0$. The vector*

$$x = \begin{bmatrix} x_B \\ x_N \end{bmatrix} = \begin{bmatrix} B^{-1}b \\ 0 \end{bmatrix}$$

is called a **basic feasible solution for S with basis B.**

Remark C.28. An arbitrary selection of m columns from the n columns of A will not necessarily be linearly independent.

Theorem C.29. *Let S be a nonempty polyhedron in standard format. A point x is an extreme point of S if and only if it is a basic feasible solution of S for some basis B.*

Corollary C.30. *The number of extreme points of S is finite.*

Proof. The number of ways to select m columns from the n columns in A is $\binom{n}{m} = \frac{n!}{(n-m)!m!}$, which is finite.[2]

The proof of the following proposition is immediate from the definitions.

Proposition C.31. *Let S be a nonempty polyhedron in standard format. The nonzero vector d is a direction of S if and only if $d \geq 0$ and $Ad = 0$.*

Characterizing the *extreme* directions of S takes a little more work. We motivate with the following example. Consider the polyhedron in the plane defined by $\{(x_1, x_2) : \ -x_1 + x_2 \leq 1, \ -x_1 + 2x_2 \leq 10, \ x_1 \geq 0, \ x_2 \geq 0\}$. See Figure C.4. There is one extreme direction given by $d = (10, 10) - (8, 9) = (2, 1)$. In standard format the polyhedron is represented as

$$\left\{ (x_1, x_2, s_1, s_2) : \ \begin{bmatrix} -1 & 1 & 1 & 0 \\ -1 & 2 & 0 & 1 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ s_1 \\ s_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 10 \end{pmatrix} \right\}.$$

[2] $n!$ denotes "n factorial," i.e., $n(n-1)(n-2)\cdots 1$.

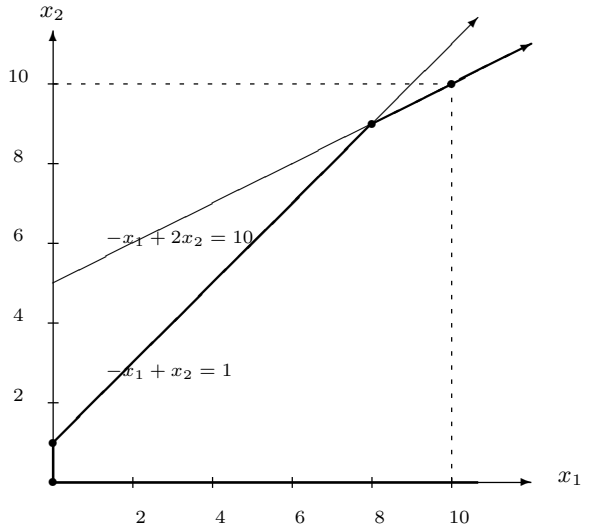


Fig. C.4. Example of an extreme direction of a polyhedron.

Construct the basis from the first two columns[3] so that

$$B = \begin{bmatrix} -1 & 1 \\ -1 & 2 \end{bmatrix} \text{ and } B^{-1} = \begin{bmatrix} -2 & 1 \\ -1 & 1 \end{bmatrix}.$$

Now multiply both sides of the equation $Ax = b$ by B^{-1} to obtain

$$\begin{bmatrix} 1 & 0 & -2 & 1 \\ 0 & 1 & -1 & 1 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ s_1 \\ s_2 \end{pmatrix} = \begin{pmatrix} 8 \\ 9 \end{pmatrix}.$$

Consistent with Theorem C.29, the extreme point $(8, 9)$ is a basic feasible solution with basis B. Note how the third column has nonpositive entries (in fact, negative entries). The extreme direction that is generated from this nonpositive *nonbasic* column is

$$d := \begin{pmatrix} 2 \\ 1 \\ 1 \\ 0 \end{pmatrix}$$

and is constructed in the following way. One multiplies the nonbasic column by minus one and places these two entries at the "top" of d. This produces the subvector $(2, 1)^T$. To complete the definition of d—keep in mind it has four entries—one places a one in the third position, since this is the position

[3] The x_1 and x_2 entries must be positive so this is the only basis.

of the nonbasic column, and places zeroes everywhere else (the last position of d in this case). That $d = (2, 1, 1, 0)^T$ is a direction of the polyhedron is easily verified since $Ad = 0$.

Remark C.32. That d is a direction is no mystery. Here is the algebraic proof. Let j denote the index of the nonbasic column whose entries are nonpositive, and let a_j denote the nonbasic column. (In this example $j = 3$.) Let $\hat{e}_j$ denote the j^{th} unit vector e_j truncated by removing the first m zeroes. The length of the vector $\hat{e}_j$ is $n - m$. (Here, $e_j^T = (0, 0, 1, 0)$ and $\hat{e}_j^T = (1, 0)$.) In this notation,

$$d = \begin{pmatrix} -B^{-1}a_j \\ \hat{e}_j \end{pmatrix} \text{ and } B^{-1}N\hat{e}_j = B^{-1}a_j.$$

(Keep in mind the nonbasic column is obtained *after* multiplying A by B^{-1}.) We have

$$B^{-1}Ad = B^{-1}[B \quad N]d = [I \quad B^{-1}N] \begin{pmatrix} -B^{-1}a_j \\ \hat{e}_j \end{pmatrix} = -B^{-1}a_j + B^{-1}a_j = 0,$$ [4]

which implies that $Ad = 0$; otherwise, B would not be invertible.

It can be shown that the d so constructed is, in fact, an extreme direction. Thus, if a nonbasic column associated with a basic feasible solution has nonpositive entries, then it generates an extreme direction of the polyhedron. *Conversely, every extreme direction can be generated in this fashion.* This characterizes extreme directions. For proofs of these facts, see Bazarra et al. [1993], p. 59.

Proposition C.33. *The number of extreme directions is finite.*

Proof. For every choice of matrix B there are $n - m$ choices for the nonbasic column. Thus, the number of extreme directions is bounded above by $(n - m)\binom{n}{m}$, which is finite. □

C.4.4 Representation Theorem for Polyhedra

The following theorem characterizes nonempty polyhedra via extreme points and extreme directions. For a proof, see Bazarra et al. [1993], pp. 60–61.

Theorem C.34. *Let S be a nonempty polyhedron in standard format. Let $x_1, x_2, \ldots, x_K$ be the extreme points of S and let $d_1, d_2, \ldots, d_L$ be the extreme directions of S. Then $x \in S$ if and only if there exists a $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_K) \geq 0$ and $\mu = (\mu_1, \mu_2, \ldots, \mu_L) \geq 0$ such that*

$$x = \sum_{k=1}^{K} \lambda_k x_k + \sum_{\ell=1}^{L} \mu_\ell d_\ell \quad \textit{and} \quad \sum_{k=1}^{K} \lambda_k = 1.$$

[4] Algebraically $[A_1 \quad A_2]\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = A_1a_1 + A_2a_2$ whenever A_i is an $m \times n_i$ matrix and a_i is an n_i-dimensional vector, $i = 1, 2$.

Corollary C.35. *S has at least one extreme direction if and only if S is unbounded.*

Proof. Clearly, if S has an extreme direction, it is unbounded. Now suppose S has no extreme directions. Then x must belong to the convex hull of a finite number of points, and so its norm must be finite.[5]

A separate proof establishes that a nonempty polyhedron S will always have at least one extreme point—see Bazarra et al. [1993], p. 58.

C.5 Application to Linear Programming

A linear programming problem is the minimization or maximization of a linear function over a polyhedron. Without loss of generality, we consider the linear program

$$(P): \qquad \min\{c^T x: \ Ax = b, \ x \geq 0\}. \tag{C.1}$$

We make the following assumptions:

Assumption 10 *(i) A is an $m \times n$ matrix of full rank m. (ii) The feasible region $\{x: \ Ax = b, \ x \geq 0\}$ is not empty.*

Let $x_1, x_2, \ldots, x_K$ denote the extreme points and $d_1, d_2, \ldots, d_L$ denote the extreme directions of the feasible region.

Theorem C.36. *Under Assumption 10:*

a) The linear program (P) has a finite optimal solution if and only if $c^T d_\ell \geq 0$ for $1 \leq \ell \leq L$.
b) If the linear program (P) has a finite optimal solution, then at least one extreme point is an optimal solution.

Proof. As a direct application of the representation theorem, the linear programming problem (P) can be reformulated as

$$(P') \quad \min\left\{\sum_{k=1}^{K} \lambda_k (c^T x_k) + \sum_{\ell=1}^{L} \mu_\ell (c^T d_\ell): \ \sum_k \lambda_k = 1, \ \lambda_k, \ \mu_\ell \geq 0 \ \text{ for all } \ k, \ell\right\}. \tag{C.2}$$

If $c^T d_\ell < 0$ for any ℓ, then by choosing μ_ℓ to be arbitrarily large, the objective function can be made to equal an arbitrarily large negative number, which implies that (P) has no finite solution.

Conversely, suppose $c^T d_\ell \geq 0$ for all ℓ. Pick a point $(\lambda, \mu) \in \mathbb{R}_+^K \times \mathbb{R}_+^L$ in the feasible region for (P'). The point $(\lambda, 0)$ is feasible and will have an objective function value no higher than (λ, μ). Consequently, under the assumed hypothesis, without loss of generality one may consider only those points in

[5] Repeated application of the triangle inequality for norms shows that $||x|| = ||\sum_{k=1}^K \lambda_k x_k|| \leq \sum_{k=1}^K \lambda_k ||x_k|| \leq \sum_{k=1}^K ||x_k||$.

the feasible region for (P') such that $\mu = 0$. It now follows that the optimal objective function value equals

$$\min_{1 \leq k \leq K} c^T x_k,$$

which completes the proof of part (a) and immediately proves part (b). □

Remark C.37. If the number of extreme points is small in number and there is a way to determine each one, then the linear program can be solved by *enumeration*. In general, the number of extreme points is too large for this approach to be practical. It is possible to start with an extreme point and execute **pivoting operations or pivots** to sequentially move from one extreme point to an **adjacent** extreme point in such a manner that the objective function never decreases. Since there are a finite number of extreme points, the algorithm will eventually terminate, as long as it does not move to an extreme point already visited. In the language of linear programming, the algorithm must not **cycle**. There are a number of proven, simple ways to avoid cycling.

C.6 Bibliographical Notes

The classic references on convexity and linear programming are Rockafellar [1970] and Dantzig [1963], respectively. Bazarra et. al. [1993] provides an in-depth, accessible coverage of this material.

D

Concave, Convex Functions and Generalizations

The classes of concave (convex) and quasiconcave (quasiconvex) functions are essential to modeling economic problems. Here, we establish their basic properties.[1]

In this chapter, S denotes a nonempty convex set in $\mathbb{R}^n$ with nonempty interior, and $f(\cdot)$ denotes a real-valued function defined on S.

D.1 Definitions

Definition D.1. *The function $f(\cdot)$ is* **concave** *if for each x_1 and x_2 in S*

$$f(\lambda x_1 + (1-\lambda)x_2) \geq \lambda f(x_1) + (1-\lambda)f(x_2)$$

holds for all $\lambda \in [0,1]$. The function $f(\cdot)$ is **strictly concave** *if the inequality holds as a strict inequality for all $\lambda \in (0,1)$ and $x_1 \neq x_2$. A function $f(\cdot)$ is* **convex** *or* **strictly convex** *if $-f(\cdot)$ is concave or strictly concave.*

Proposition D.2. *The following functions are concave or convex:*

a) A finite, positive linear combination of concave (convex) functions is concave (convex).
b) The minimum of a finite collection of concave functions is concave, and the maximum of a finite collection of convex functions is convex.

Theorem D.3. Characterization of concave (convex) functions via level sets.

a) $f(\cdot)$ is concave if and only if hypo(f) is convex.
b) $f(\cdot)$ is convex if and only if epi(f) is convex.

[1] Consult Bazarra et. al. [1993] for more detail on this subject matter.

Every concave (convex) function has at least one supergradient (subgradient) at each point in the interior of S. Conversely, if each point in the interior of S has at least one supergradient, then $f(\cdot)$ is concave on the interior of S (but not necessarily on all of S).

Theorem D.4. *If $f(\cdot)$ is concave, then the superdifferential $\partial f(\bar{x})$ is not empty for each $\bar{x}$ in the interior of S.*

Proof. The point $(\bar{x}, f(\bar{x}))$ lies on the boundary of $hypo(f)$, which is a convex set. It has a supporting hyperplane, which can be identified with a point $(\xi, \mu) \in \mathbb{R}^n \times \mathbb{R}$ for which

$$\xi \cdot \bar{x} + \mu f(\bar{x}) \leq \xi \cdot x + \mu\gamma \text{ for all } (x, \gamma) \in hypo(f).$$

Since $\bar{x}$ lies in the interior of S, $\bar{x} - \delta\xi \in S$ for sufficiently small positive δ, which shows that μ cannot be zero. Since γ can take on arbitrarily high negative values, we conclude that μ must be negative. The vector $\bar{x}^* = \xi/(-\mu)$ belongs to the superdifferential of $f(\cdot)$ at $\bar{x}$. □

Theorem D.5. *If the superdifferential $\partial f(\bar{x})$ is not empty for each $\bar{x}$ in the interior of S, then $f(\cdot)$ is concave on the interior of S.*

Proof. Let x_1 and x_2 be two points that lie in the interior of S, and pick a point $\bar{x} = \lambda x_1 + (1-\lambda)x_2$ that lies in the interior of the line segment joining these two points. Since the interior of S is convex, there is a supergradient of $f(\cdot)$ at $\bar{x}$. Accordingly,

$$f(x_1) \leq f(\bar{x}) + \bar{x}^* \cdot (x_1 - \bar{x}) = f(\bar{x}) + (1-\lambda)\bar{x}^* \cdot (x_1 - x_2), \quad \text{(D.1)}$$
$$f(x_2) \leq f(\bar{x}) + \bar{x}^* \cdot (x_2 - \bar{x}) = f(\bar{x}) - \lambda\bar{x}^* \cdot (x_1 - x_2). \quad \text{(D.2)}$$

Now multiply the first equation by λ, the second by $(1-\lambda)$, and add to obtain the desired result. □

Theorem D.6. *A concave function is continuous on the interior of its domain.*

D.2 Quasiconcavity and Quasiconvexity

The properties of convex sets make them extremely useful as modeling approximations. It is possible to retain the analytical power of convexity with the following class of functions.

Definition D.7. *The function $f(\cdot)$ is* **quasiconcave** *if each upper level set $L_f^{\geq}(\alpha)$ is convex. The function $f(\cdot)$ is* **quasiconvex** *if each lower level set $L_f^{\leq}(\alpha)$ is convex.*

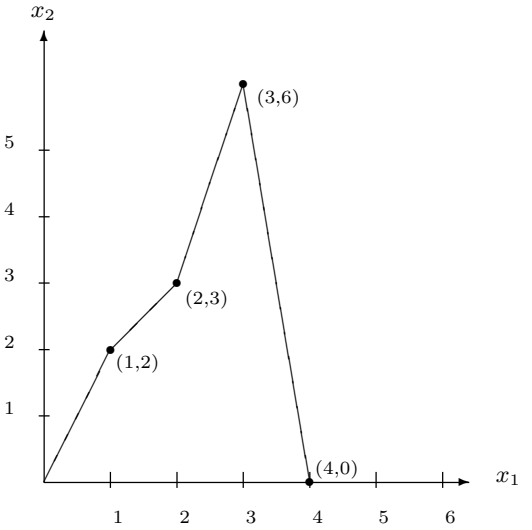


Fig. D.1. Example of a quasiconcave function that is not concave.

Example D.8. An example of function that is *not* concave but is quasiconcave is depicted in Figure D.1. Each upper level is a closed interval. The hypograph is not convex, and so this function is not concave.

Theorem D.9. *If $f(\cdot)$ is concave (convex), then $f(\cdot)$ is quasiconcave (quasiconvex).*

Proof. The result follows from the easily verified fact that each upper (lower) level set of a concave (convex) function is convex. □

An equivalent definition of quasiconcavity (quasiconvexity) is stated in the following theorem.

Theorem D.10. Characterization of quasiconcavity (quasiconvexity).

a) $f(\cdot)$ is quasiconcave if and only if

$$f(\lambda x_1 + (1-\lambda)x_2) \geq \min\{f(x_1), f(x_2)\}$$

holds for each x_1 and x_2 in S and for all $\lambda \in [0,1]$.

b) $f(\cdot)$ is quasiconvex if and only if

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \max\{f(x_1), f(x_2)\}$$

holds for each x_1 and x_2 in S and for all $\lambda \in [0,1]$.

D.3 Differential Characterizations

There are a number of powerful characterizations for concavity and quasiconcavity (convexity and quasiconvexity) when the function under consideration is sufficiently differentiable.

Theorem D.11. *Let $f(\cdot)$ be differentiable.*

a) If $f(\cdot)$ is concave, then for all $x \in S$

$$f(x) \leq f(\bar{x}) + \nabla f(x) \cdot (x - \bar{x}). \tag{D.3}$$

b) If $f(\cdot)$ is convex, then for all $x \in S$

$$f(x) \geq f(\bar{x}) + \nabla f(x) \cdot (x - \bar{x}). \tag{D.4}$$

Proof. Theorem D.4 established that a concave (convex) function has at least one supergradient (subgradient) at each point in the interior of S. It therefore follows that when $f(\cdot)$ is differentiable at x, the superdifferential (subdifferential) of $f(\cdot)$ at x must exactly coincide with the gradient vector. The result follows from the definition of superdifferential (subdifferential). □

We turn to establishing a characterization of differentiable quasiconcave (quasiconvex) functions.

Theorem D.12. *Suppose S is open and $f(\cdot)$ is differentiable. Then:*

a) $f(\cdot)$ is quasiconcave if and only if for each $x \in S$ the hyperplane $H(p, \alpha)$ supports the upper level set $L_f^{\geq}(f(x))$ at x from below, where $p := \nabla f(x)$ and $\alpha := \nabla f(x) \cdot x$. That is,

$$L_f^{\geq}(f(x)) \subset H^{\geq}(\nabla f(x), \nabla f(x) \cdot x).$$

b) $f(\cdot)$ is quasiconvex if and only if for each $x \in S$ the hyperplane $H(p, \alpha)$ supports the lower level set $L_f^{\geq}(f(x))$ at x from above, where $p := \nabla f(x)$ and $\alpha := \nabla f(x) \cdot x$. That is,

$$L_f^{\leq}(f(x)) \subset H^{\leq}(\nabla f(x), \nabla f(x) \cdot x).$$

Proof. It is sufficient to establish part (a). Pick x_1 and x_2 in S and, without loss of generality, assume that $f(x_1) \leq f(x_2)$. Set $d = x_2 - x_1$.

Suppose first that $f(\cdot)$ is quasiconcave. Due to the quasiconcavity of $f(\cdot)$, we know $L^{\geq}(f(x_1))$ is convex, from which it follows that $f(x_1 + \alpha d) \geq f(x_1)$ for all $\alpha \in [0, 1]$. Since $f(\cdot)$ is differentiable, we also know that

$$f(x_1 + \alpha d) = f(x_1) + \alpha d \cdot \nabla f(x_1) + o(\alpha)$$

for sufficiently small α. This can only be true if $d \cdot \nabla f(x_1) = (x_2 - x_1) \cdot \nabla f(x_1)$ is nonnegative, as required.

To establish the converse, we argue by contradiction. Define the function

$$\theta(\gamma) := f(x_1 + \gamma d) - f(x_1),\ \gamma \in [0,1].$$

We suppose a γ exists for which $\theta(\gamma) < 0$, and then show this leads to a contradiction. Note that $\theta(0) = 0$ and $\theta(1) \geq 0$. Since $\theta(\cdot)$ is continuous it achieves its minimum on $[0,1]$, say at γ^*. Clearly, $\gamma^* \in (0,1)$. The set $\theta^{-1}\{0\} \cap [0,\gamma^*]$ is compact and therefore achieves its supremum, say at γ_m. Obviously, $\theta(\gamma_m) = 0$ and $\gamma_m \neq \gamma^*$. The continuity of $\theta(\cdot)$ ensures that $\theta(\gamma) < 0$ for all $\gamma \in (\gamma_m, \gamma^*)$. Invoking the Mean-Value Theorem A.11, p. 447, there exists a $\lambda \in [\gamma_m, \gamma^*]$ such that

$$\theta(\gamma^*) = \theta(\gamma_m) + (\gamma^* - \gamma_m)\theta'(\lambda) = (\gamma^* - \gamma_m)\theta'(\lambda).$$

This in turn implies that

$$0 > \theta'(\lambda) = \lambda \nabla f(x_\lambda) \cdot d,$$

where $x_\lambda := x_1 + \lambda d$. Thus, $\nabla f(x_\lambda) \cdot d$ is negative. Since $\theta(\lambda) < 0$, it follows that $f(x_2) \geq f(x_\lambda)$, and so $x_2 \in L^{\geq}(f(x_\lambda))$. Now we invoke the assumption that the gradient $\nabla f(x_\lambda)$ supports $L^{\geq}(f(x_\lambda))$ at x_λ to obtain that

$$\nabla f(x_\lambda) \cdot (x_\lambda) \leq \nabla f(x_\lambda) \cdot x_2$$

or, equivalently,

$$0 \leq (1 - \lambda)\nabla f(x_\lambda) \cdot d.$$

This implies that $\nabla f(x_\lambda) \cdot d$ is nonnegative. Obviously, $\nabla f(x_\lambda) \cdot d$ cannot be both negative and nonnegative, and so the desired contradiction has been reached. □

We now come to a very important characterization of twice continuously differentiable concave (convex) functions.

Theorem D.13. *Suppose S is open and $f(\cdot)$ is twice differentiable on S. Then, $f(\cdot)$ is concave (convex) if and only if the Hessian matrix is negative (positive) semidefinite at each point in S.*

Proof. Suppose first that $f(\cdot)$ is concave. Pick an $x \in S$ and a $d \in R^n$. Since S is open, we may assume $x + \alpha d \in S$ for all sufficiently small α. Since the gradient vector is a supergradient,

$$f(x + \alpha d) \leq f(x) + \alpha \nabla f(x) \cdot d.$$

As $\alpha \to 0$, it must be the case that $d^T H(x) d \leq 0$ in (A.11), p. 448, which shows that $H(x)$ is indeed negative semidefinite. The converse is an immediate consequence of the second-order form of Taylor's theorem (A.16), p. 448. □

Remark D.14. As a special case of Theorem D.13, we note the important, well-known fact from ordinary calculus about a twice continuously differentiable function $f(\cdot)$ of a single variable, namely, that $f(\cdot)$ is concave (convex) if and only if its second derivative is always nonpositive (nonnegative).

Remark D.15. When $f(x) = a + c \cdot x + 1/2\ x^T Q x$ is quadratic the Hessian is Q, which is independent of x. Without loss of generality, we may assume Q is symmetric. (If not, replace it with the symmetric matrix $1/2(Q + Q^T)$.) Thus, a quadratic function is concave (convex) if and only if Q is negative (positive) semidefinite. It is a fact of linear algebra that a matrix Q is negative (positive) semidefinite if and only if all of its eigenvalues are non-positive (non-negative).

E

Optimality Conditions

Not surprisingly, economists assume economic agents (e.g., consumers, producers) make rational decisions about what to consume, how much to save, invest, produce, etc. Such problems are formulated as optimization problems, which typically involve an objective function that measures the value to the economic agent of making a particular choice, and a collection of constraints that define what choices are feasible. In this chapter, we define classes of optimization problems, provide necessary and sufficient conditions for optimality, and show how to solve such problems.

E.1 Unconstrained Problems

Definition E.1. *Let* $f : \mathbb{R}^n \to \mathbb{R}$.

- *If* $f(x^*) \geq f(x)$ *for all* $x \in \mathbb{R}^n$, *then* x^* *is a* **global maximizer**.
- *If* $f(x^*) \geq f(x)$ *for all* x *in some neighborhood of* x, *then* x^* *is a* **local maximizer**.
- *If* $f(x^*) > f(x)$ *for all* x *in some neighborhood of* x *(excluding* x^* *of course), then* x^* *is a* **strict local maximizer**.

Definition E.2. *The gradient at* x *is said to* **vanish** *if* $\nabla f(x) = 0$.

Theorem E.3. *Suppose* $f : \mathbb{R}^n \to \mathbb{R}$ *is differentiable at* x^*. *If* x^* *is a local maximizer, then the gradient at* x^* *vanishes.*

Proof. If $\nabla f(x^*) \neq 0$, then the direction $d = \nabla f(x^*)$ has positive norm. For α sufficiently small, $x^* + \alpha d$ lies in the neighborhood where x^* is a maximum, and thus $f(x^* + \alpha d) \leq f(x^*)$. Since $f(\cdot)$ is differentiable at x^*,

$$f(x^* + \alpha d) = f(x^*) + \alpha \nabla f(x^*) \cdot d + o(\alpha) = f(x^*) + \alpha ||\nabla f(x^*)||^2 + o(\alpha),$$

which contradicts the local maximality of x^* for α sufficiently small. □

Theorem E.3 is an example of a *necessary* condition. The following theorem establishes a *sufficient* condition for x^* to be a strict local maximizer.

Theorem E.4. *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable at x^*. If $\nabla f(x^*) = 0$ and $H(x^*)$ is negative definite, then x^* is a strict local maximizer.*

Proof. If x^* is not a strict local maximizer, then there exists an infinite sequence x_k converging to x^* for which $f(x_k) \geq f(x^*)$, $x_k \neq x^*$. Let $d_k = x_k - x^*$ for each k. Using (A.10), p. 447, and the fact that $\nabla f(x^*)$ vanishes,

$$f(x^* + d_k) = f(x^*) + 1/2\ d_k^T H(x^*) d_k + o(||d_k||^2).$$

Since $f(x_k) \geq f(x^*)$,

$$1/2\ d_k^T H(x^*) d_k + o(||d_k||^2) \geq 0,$$

or, equivalently,

$$1/2\ \hat{d}_k^T H(x^*) \hat{d}_k + \frac{o(||d_k||^2)}{||d_k||^2} \geq 0 \tag{E.1}$$

where $\hat{d}_k := d_k/||d_k||$. The $\{\hat{d}_k\}$'s belong to the unit ball, which is compact, and so we may extract a convergent subsequence, say $\hat{d}_{n_k} \to \hat{d}$. By letting $n_k \to \infty$ in (E.1), we conclude that

$$\hat{d}^T H(x^*) \hat{d} \geq 0.$$

This contradicts the negative semidefiniteness of $H(x^*)$. □

Remark E.5. Theorems E.3 and E.4 both hold if $f : \mathbb{R}^n \to \mathbb{R}$ is defined on an open set $S \subseteq \mathbb{R}^n$.

E.2 Problems with Inequality Constraints

We consider the general maximization problem defined by

$$(P): \qquad \{\max f(x) : g_i(x) \geq 0,\ i = 1, 2, \ldots, m\}. \tag{E.2}$$

In this section we make the following assumptions.

Assumption 11

- *$f(\cdot)$ and each $g_i(\cdot)$ are real-valued, differentiable functions defined on an open set $S \subseteq \mathbb{R}^n$.*
- *The vectors $\nabla g_i(x)$, $i \in I(x)$, are linearly independent,*[1] *where $I(x) := \{i : g_i(x) = 0\}$ for each feasible x. (This condition is an example of what is termed a* **constraint qualification***.)*

[1] This means that if $\sum_{i \in I(x)} \mu_i \nabla g_i(x) = 0$, then each $\mu_i = 0$.

Let $g(x)$ denote the vector $(g_1(x), \ldots, g_m(x))$ and $\nabla g(x)$ denote the vector $(\nabla g_1(x), \ldots, \nabla g_m(x))$.

Definition E.6. *Vectors $x \in S$ and $\lambda \in \mathbb{R}^m$ are said to satisfy the* **complementary slackness conditions** *if $\lambda_i g_i(x) = 0$ for $i = 1, 2, \ldots, m$.*

Definition E.7. *A vector x^* is said to satisfy the* **first-order optimality conditions** *[or the* **(KKT) conditions***] if*

i) x^* is feasible for (P).
ii) There exists a non-negative vector $\lambda^* \in \mathbb{R}^m$ for which (x^*, λ^*) satisfies the complementary slackness conditions.
iii) $\nabla f(x^*) + \lambda^* \cdot \nabla g(x^*) = 0$.

The following theorem establishes the *Karush-Kuhn-Tucker (KKT) Necessary Conditions* for a vector x^* to be a local maximizer of (P).

Theorem E.8. *Under Assumption 11, if x^* is a local maximizer of (P), then x^* satisfies the (KKT) conditions.*

Proof. Let A denote the matrix whose rows consist of the vector $\nabla f(x^*)$ and $\nabla g_i(x^*)$ for $i \in I(x^*)$. We first establish the claim that there cannot be a vector $d \in \mathbb{R}^n$ for which $Ad > 0$. To this end consider the set

$$F = \{d \in \mathbb{R}^n : \nabla g_i(x^*) \cdot d > 0 \text{ for all } i \in I(x^*)\}, \tag{E.3}$$

and suppose F is non-empty. Pick a $d \in F$. For each $i \in I(x^*)$, there exists a $\delta > 0$ for which $g(x^* + \alpha d) \geq 0$ for all $\alpha \in [0, \delta]$. For each $i \notin I(x^*)$, the continuity of g_i ensures existence of a neighborhood about x^* for which $g_i(x^*)$ remains positive. Given that S is open, there exists a $\delta^* > 0$ for which $x + \alpha d$ is feasible for (P) for all $\alpha \in [0, \delta^*]$. Since x^* is a local maximum, it immediately follows that $\nabla f(x^*) \cdot d$ cannot be positive. Thus, the claim has been shown.

If the system $Ad > 0$ has no solution, then application of the separation theorem for convex sets shows there must exist a *nonnegative* vector y for which $y^T A = 0$.[2] That is,

$$y_0 \nabla f(x^*) + \sum_{i \in I(x^*)} y_i \nabla g_i(x^*) = 0. \tag{E.4}$$

With a slight abuse of notation extend y to all of $\mathbb{R}^{m+1}$ by defining $y_j = 0$ for $j \notin I(x^*)$ so that

$$y_0 \nabla f(x^*) + \sum_{i=1}^{m} y_i \nabla g_i(x^*) = 0. \tag{E.5}$$

The linear independence assumption ensures that y_0 is positive, and so we may divide both sides of (E.5) by y_0 to obtain the desired result. □

[2] See Exercise 8.4, p. 143.

The (KKT) conditions are extremely useful for identifying *possible* (local) solutions to (P). When the set S is closed, one must also check the boundary of S. In many economic problems, it will be clear a solution exists and cannot lie on the boundary.

We now establish an example of *sufficient* conditions for optimality of (P). The statement below only assumes that each $g_i(\cdot)$ is quasiconcave.

Theorem E.9. *Suppose $f(x) = c^T x$ is linear and each $g_i(\cdot)$ is quasiconcave. Under Assumption 11, if x^* satisfies the (KKT) conditions, then x^* is optimal for (P).*

Proof. Using Theorem D.12, we know the gradient $\nabla g_i(x^*)$ induces a supporting hyperplane to the upper level set $L^{\geq}_{g_i}(g_i(x^*))$ at x^*. In particular, for each $k \in I(x^*)$, we have that

$$\nabla g_k(x^*) \cdot (z - x^*) \geq 0$$

for each feasible z. Since $\lambda_k^* = 0$ for each $k \notin I(x^*)$ it then follows that

$$\sum_{k=1}^{m} \lambda_k^* \nabla g_k(x^*) \cdot (z - x^*) \geq 0$$

for each feasible z. Since the inner product

$$(\nabla f(x^*) + \lambda^* \cdot \nabla g(x^*)) \cdot (z - x^*)$$

is obviously zero, it now follows that

$$\nabla f(x^*) \cdot (z - x) \leq 0.$$

Optimality follows from the linearity of $f(\cdot)$ since $\nabla f(x^*) = c^T$. □

E.3 Lagrangian Duality

We once again consider problem (P). In this section, we impose the following additional assumptions.

Assumption 12

- *$f(\cdot)$ and each $g_i(\cdot)$ are concave.*[3]
- *S is convex.*
- *(P) has a finite optimal solution.*
- **Slater's condition** *holds, namely, there exists an $x \in S$ for which $g_i(x) > 0$ for each i.*

[3] Concavity is not required for several of the results below.

Under Assumptions 11 and 12, problem (P) is intimately related to an *unconstrained* problem defined by the *Lagrangian* function.

Definition E.10. *The function*

$$L(x, \lambda) := f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) = f(x) + \lambda \cdot g(x) \tag{E.6}$$

defined on $S \times \mathbb{R}^m$ *is called the* **Lagrangian** *function. The point* $(x^*, \lambda^*) \in S \times \mathbb{R}^m$ *is called a* **saddle point** *of the Lagrangian if*

$$L(x, \lambda^*) \leq L(x^*, \lambda^*) \leq L(x^*, \lambda) \tag{E.7}$$

holds for all $(x, \lambda) \in S \times \mathbb{R}^m_+$.

In the developments to follow, we shall find it convenient to use the following two expressions for a given (x^*, λ^*):

Condition A: $L(x, \lambda^*) \leq L(x^*, \lambda^*)$ holds for all $x \in S$.

Condition B: $L(x^*, \lambda^*) \leq L(x^*, \lambda)$ holds for all $\lambda \in \mathbb{R}^m_+$.

Lemma E.11. *Under Assumptions 11 and 12, if* (x^*, λ^*) *is a saddle point of the Lagrangian, then* (x^*, λ^*) *satisfies complementary slackness.*

Proof. Setting $\lambda = 0$ in Condition B and noting that $\lambda^* \cdot g(x^*)$ is always nonnegative shows that (x^*, λ^*) satisfies complementary slackness. □

Theorem E.12. *Under Assumptions 11 and 12, if* (x^*, λ^*) *is a saddle point of the Lagrangian, then* x^* *is an optimal solution to* (P).

Proof. For Condition B to hold each $g_i(x^*)$ must be nonnegative, since each λ_i may take on arbitrarily large positive values. Thus, x^* is feasible. Lemma E.11 (complementary slackness), Condition A, and the fact that $\lambda^* \cdot g(x)$ is nonnegative for each feasible x shows that $f(x) \leq f(x^*)$ for each $x \in S$. Since x^* is feasible, it is obviously optimal. □

Theorem E.13. *Under Assumptions 11 and 12, if* x^* *is an optimal solution to* (P)*, then a* λ^* *exists for which* (x^*, λ^*) *is a saddle point of the Lagrangian.*

Proof. Define the set

$$C \equiv \{(\alpha, \beta) \in \mathbb{R} \times \mathbb{R}^m : \alpha \leq f(x) - f(x^*),\ \beta \leq g(x) \text{ for some } x \in S\}. \tag{E.8}$$

C is convex and does not intersect the convex cone $D = \mathbb{R}_{++} \times \mathbb{R}^m_+$. The separation theorem for convex sets ensures existence of a (non-trivial) hyperplane $\{(\alpha, \beta) : \alpha^* \alpha + \beta^* \cdot \beta = \gamma\}$ that (weakly) separates C from D, namely,

$$\alpha^* \alpha + \beta^* \cdot \beta \geq \gamma \geq \alpha^* \bar{\alpha} + \beta^* \cdot \bar{\beta} \tag{E.9}$$

holds for all $(\alpha, \beta) \in D$ and $(\bar{\alpha}, \bar{\beta}) \in C$. Since there exists points in C that can take on arbitrarily high negative values, (E.9) can only hold if $(\alpha^*, \beta^*) \geq 0$. Furthermore, since $(0, 0) \in C$ and the left-hand side of the inequality in (E.9) can be arbitrarily close to 0, it follows that $\gamma = 0$. Slater's condition ensures that $\alpha^* \neq 0$. Now set $\lambda^* = \beta^*/\alpha^*$, and use the right-hand side of (E.9) to establish that $f(x) - f(x^*) + \lambda^* g(x) \leq 0$, which is Condition A. Since $(0, g(x^*)) \in C$ the right-hand side of (E.9) shows that $\lambda^* \cdot g(x^*) \leq 0$. Since $\lambda^* \cdot g(x^*)$ is always non-negative, complementary slackness holds, and Condition B immediately follows. □

We turn to the differentiable setting. A real-valued, differentiable concave function $h(\cdot)$ defined on an open convex set S is characterized by the following intuitive geometrical property: for all $x \in S$ and $y \in S$,

$$h(y) \leq h(x) + \nabla h(x) \cdot (y - x). \tag{E.10}$$

Note that if $\nabla h(x^*) = 0$, then x^* maximizes $h(\cdot)$ on S. A necessary condition for x^* to be a local maximizer is that its gradient must vanish; for concave functions defined on open sets this property is sufficient for x^* to be a global maximum.

Theorem E.14. *Suppose $f(\cdot)$ and each $g_i(\cdot)$ are differentiable. Under Assumptions 11 and 12, x^* optimizes (P) if and only if x^* satisfies the first-order optimality conditions.*

Proof. If x^* optimizes (P), then obviously it is feasible. From Theorem E.13 a λ^* exists for which (x^*, λ^*) is a saddle point of the Langrangian. Complementary slackness follows from feasibility and Condition B. Condition A and complementary slackness shows that x^* optimizes $L(\cdot, \lambda^*)$ on S, an open set. The gradient of $L(\cdot, \lambda^*)$ must vanish, which is condition (iii).

As for the converse, the function $L(\cdot, \lambda^*)$ is concave and its gradient at x^* vanishes. Consequently, x^* optimizes $L(\cdot, \lambda^*)$. Feasibility of x^* and complementary slackness now show that x^* optimizes (P). □

Definition E.15. *The* **dual problem** *to problem (P) is given as*

$$(D) : \inf_{\lambda \geq 0} \; V(\lambda) \tag{E.11}$$

where

$$V(\lambda) := sup_{x \in S} \, L(x, \lambda). \tag{E.12}$$

for each $\lambda \in \mathbb{R}^m_+$. We say that (x^*, λ^*) solves (D) if $V(\lambda^*) \leq V(\lambda)$ for all $\lambda \geq 0$ and $V(\lambda^*) = L(x^*, \lambda^*)$.

Theorem E.16. Weak duality *Under Assumptions 11 and 12, the optimal objective function value for (D) is an upper bound to the optimal objective function value for (P).*

Proof. Observe that $V(\lambda) \geq f(x)$ for each feasible x and nonnegative λ. □

Theorem E.17. Strong duality *Under Assumptions 11 and 12, the optimal objective function values for the primal (P) and dual (D) problems coincide; that is, $f(x^*) = V(\lambda^*)$ for respective optimal solutions x^* and λ^*. In particular, x^* solves the maximization problem defined by $V(\lambda^*)$ and (x^*, λ^*) satisfies the complementary slackness condition.*

Proof. Let x^* denote an optimal solution to (P). By Theorem E.13, a λ^* exists for which (x^*, λ^*) is a saddle point of the Lagrangian. Proposition E.11 (x^*, λ^*) and Condition A immediately implies that $V(\lambda^*) \leq f(x^*)$. The result now follows from Theorem E.16. □

For many optimization problems it is possible to efficiently solve for $V(\lambda)$ for a given λ. It is often the case that the number of decision variables in the dual problem is far less than the number of decision variables in the primal problem. Consequently, solving the dual problem can be much easier from a computational perspective. The following theorem shows that the dual problem is a convex minimization problem.

Theorem E.18. *Under Assumptions 11 and 12, the function $V(\cdot)$ is convex.*

Proof. Pick $\lambda_i \in \mathbb{R}^m_+$, $i = 1, 2$, and an $\delta \in [0, 1]$. By definition,

$$\begin{aligned} V(\delta\lambda_1 + (1-\delta)\lambda_2) &= \sup_x \{\, f(x) + (\delta\lambda_1 + (1-\delta)\lambda_2) \cdot g(x) \,\} \\ &= \sup_x \{\, \delta\, [f(x) + \lambda_1 \cdot g(x)] + (1-\delta)[f(x) + \lambda_2 \cdot g(x)] \,\} \\ &\leq \delta \,\{\, \sup_x [f(x) + \lambda_1 \cdot g(x)] \,\} + (1-\delta)\{\, \sup_x [f(x) + \lambda_2 \cdot g(x)] \,\} \\ &= \delta V(\lambda_1) + (1-\delta) V(\lambda_2). \end{aligned}$$

If the objective function is smooth (i.e., sufficiently differentiable), then there are several relatively simple algorithms to solve this type of problem. In general, the objective function may not be smooth. However, the past 20 years has seen a number of efficient algorithms developed to solve such nonsmooth convex minimization problems.

Remark E.19. From a practical perspective, it is often unnecessary to find the optimal solution, a near-optimal solution will suffice. Theorem E.16 then becomes an important tool: the ratio of $V(\lambda)$ to $f(x)$ for a feasible x and *any* choice of λ provides an *a posteriori* bound on how good x is. For example, if $V(\lambda)/f(x) = 1.02$, then $f(x^*) \leq 1.02 f(x)$, which implies that $f(x^*)$ is within 2% of the optimal objective function value.

E.4 Application of Duality to Economic Lot Sizes

We consider the following optimization problem:

$$(P): \quad \min\left\{ \sum_i \left(\frac{a_i}{x_i} + b_i x_i\right) : \sum_i r_i x_i \le R \right\}. \tag{E.13}$$

This problem has very special structure—an additively-separable objective function and constraint—we shall shortly exploit, and is common to many of the economic optimization problems we shall formulate in this book. The function $f_i(x) := a_i/x_i + b_i x_i$ arises in a variety of contexts, and is commonly referred to as an *EOQ form*. In an inventory context, the (positive) decision variable x_i refers to how much quantity to order of product i. Parameter a_i denotes the setup cost for making a single order, b_i denotes the holding cost penalty per unit time, r_i denotes the resource required by item i (typically cash or space), and R denotes the maximum resource available. For future reference, we note now that

- The **unconstrained minimum** of $f_i(x_i)$ is achieved at $x_i^* = \sqrt{a_i/b_i}$ and $f_i(x_i^*) = 2\sqrt{a_i b_i}$.

- The **aggregate demand for resource in the unconstrained problem** is $\sum_i r_i \sqrt{a_i/b_i}$.

Since there are possibly many items, but only one resource constraint, we shall tackle this problem via duality. First, we shall express this problem in canonical form (E.2) as

$$\max\left\{ -\sum_i \left(\frac{a_i}{x_i} + b_i x_i\right) : R - \sum_i r_i x_i \ge 0 \right\}. \tag{E.14}$$

The Lagrangian is

$$\begin{aligned} L(x,\lambda) &= -\sum_i \left(\frac{a_i}{x_i} + b_i x_i\right) + \lambda\left(R - \sum_i r_i x_i\right) \\ &= \lambda R - \sum_i \left[\frac{a_i}{x_i} + (b_i + \lambda r_i) x_i\right], \end{aligned} \tag{E.15}$$

and the dual objective function is

$$\begin{aligned} V(\lambda) &= \sup_x \left\{ \lambda R - \sum_i \left[\frac{a_i}{x_i} + (b_i + \lambda r_i) x_i\right] \right\} \\ &= \lambda R - \inf_x \left\{ \sum_i \left[\frac{a_i}{x_i} + (b_i + \lambda r_i) x_i\right] \right\} \\ &= \lambda R - 2\sum_i \sqrt{a_i (b_i + \lambda r_i)}. \end{aligned} \tag{E.16}$$

Since $-\sqrt{a_i(b_i + \lambda r_i)}$ is convex and the sum of convex functions is convex, $V(\cdot)$ is convex, too, as required. It is obviously differentiable.

We seek to minimize $V(\cdot)$ on $[0, \infty)$. Its derivative is

$$V'(\lambda) = R - \sum_i \frac{a_i r_i}{\sqrt{a_i(b_i + \lambda r_i)}}, \tag{E.17}$$

which is clearly a strictly increasing function of λ whose limiting value is R. If the derivative at zero is negative, then eventually the derivative must cross zero at a unique point, and this point will indeed be the optimal solution to the dual problem. It is possible, however, that the derivative never vanishes, which can only happen if the derivative at zero is nonnegative. From (E.17)

$$V'(0) = R - \sum_i r_i \sqrt{a_i / b_i}\ ,$$

which is precisely the difference between the supply of the resource, R, and the aggregate demand for the resource for the unconstrained problem. Thus, if the aggregate demand in the unconstrained problem does not violate the constraint, the optimal value of the dual variable is indeed zero, and the solution for the dual problem will coincide with the solution of the unconstrained problem. (Note that the optimum dual variable in this case must be zero by complementary slackness.)

Let us assume that resource is scarce so that $V'(0) < 0$. To find the point λ^* at which $V'(\lambda^*) = 0$, one merely performs a bisection search, informally expressed as: if $V'(\lambda) < 0$, increase the lower bound for λ; if $V'(\lambda) > 0$, decrease the upper bound for λ; otherwise, stop (when $V'(0)$ is sufficiently close to zero).

We close here by providing an interpretation to the bisection search from the primal problem's perspective. The aggregate demand for resource for a given value of λ is

$$x(\lambda) := \sum_i \frac{a_i r_i}{b_i + \lambda v_i}.$$

At the optimum value λ^*, we know complementary slackness holds, which means here that $\lambda^*[R - x(\lambda^*)] = 0$. Thus, the constraint will be tight when supply is scarce. Since $x(\cdot)$ is a decreasing function, there is obviously a unique value for λ^*, and it may be found by performing a bisection search. But note that $V'(\lambda)$ given in (E.17) is precisely the difference between the supply, R, and the aggregate demand, $x(\lambda)$, and so this is the same bisection search.

E.5 Application of Duality to Linear Programming

We apply Lagrangian duality to establish the familiar duality of linear programming. Consider the **linear programming problem** defined as

$$\max_{x \geq 0} \{c^T x : Ax \leq b\}, \tag{E.18}$$

where A denotes an m by n matrix (of full rank). We assume the linear program has an optimal solution. Then the dual problem (D) may be expressed as

$$\min_{y \geq 0} \{y^T b : y^T A \geq c^T\}, \tag{E.19}$$

and it has an optimal solution whose objective value coincides with that of the primal problem.

To derive the linear programming duality from Lagrangian duality, we begin by expressing the problem as an instance of problem (P), namely, as

$$\max\{c^T x : b - Ax \geq 0,\ x \geq 0\}.$$

Let y denote the dual variables associated with the constraints $b - Ax \geq 0$, let ν denote the dual variables associated with the nonnegativity constraints, which are essential to include, and let $\lambda = (y, \nu)$. The Lagrangian here is

$$L(x, \lambda) = c^T x + y^T(b - Ax) + \nu^T x = (c^T - y^T A + \nu^T)x + y^T b,$$

and the dual objective function $V(\lambda)$ is

$$V(\lambda) = \sup_x [\ (c^T - y^T A + \nu^T)x + y^T b\ ]. \tag{E.20}$$

Since a finite optimal solution is assumed to exist for problem (P), Theorem (E.17) ensures that the dual problem (D) also has a finite optimal solution. Since x may chosen arbitrarily in (E.20), a finite optimal solution for (D) can only exist if

$$c^T - y^T A + \nu^T = 0, \tag{E.21}$$

which implies that $V(\lambda) = y^T b$. Since both y and ν are nonnegative and since $\nu^T = y^T A - c^T$, the dual variables must satisfy $y^T A \geq c^T$. To conclude, when minimizing over the dual variables λ in problem (D), it is sufficient to restrict attention to the domain where $y^T A \geq c^T$, and on this domain $V(\lambda) = y^T b$. The linear programming dual problem (E.19) immediately follows from the definition of (D), and Theorem E.17 guarantees it has an optimal solution whose objective value coincides with that of the primal problem.

Remark E.20. Using the fact that an equality constraint may be written as two linear inequalities, the dual linear program to the linear program defined by

$$\min\ \{c^T x :\ Ax = b\}$$

is

$$\min_{y \in \mathbb{R}^m} \{y^T b : y^T A \geq c^T\}. \tag{E.22}$$

That is, it is the same dual linear program as before, *except* that the dual variables are now *unconstrained*.

Remark E.21. Linear programming duality holds under much weaker conditions than what is assumed here. Lagrangian duality is, however, a convenient means for remembering the form of the dual linear program!

E.6 Bibliographical Notes

The classic reference on Lagrangian duality and convexity is Rockafellar [1970]. Bazarra et. al. [1993] provides a more accessible treatment, and also covers the basics of linear programming.

F

Envelope Theorem

Sensitivity analyzes play a major role in economics. In this chapter, for a general class of optimization problems we show how to obtain the sensitivity of the optimal value to a change in a problem parameter.

F.1 Statement and Proof

We consider the general optimization problem given by

$$M(a) \ := \ \max_{x \in \mathbb{R}^n} \{F(x,a) : \ g_k(x,a) \geq 0, \ k = 1, \ldots, K\}. \tag{F.1}$$

The vector a is viewed as representing the *parameters* of the optimization problem, and we shall be interested to see how $M(a)$ varies with changes in the coordinates of a.

Example F.1. The producer's cost minimization problem is, of course, a special case of (F.1): identify a with (u, p), M with Q, $F(x, a)$ with $-p \cdot x$, and $g_k(x, a) = x_k$ for $k = 1, 2, \ldots n$, and $g_{n+1}(x, a)$ with $\Phi(x) - u$.

For the statement and proof to follow, we make the following assumptions:

Assumption 13

- $\mathsf{A} \subset \mathbb{R}^m$ *is open.*
- $F(\cdot, \cdot)$ *and each* $g_k(\cdot, \cdot)$ *are differentiable on* $\mathbb{R}^n \times \mathsf{A}$.
- *For each* $a \in \mathsf{A}$*:*
  - *there is a unique solution* $x(a)$*;*
  - $x(\cdot)$ *is differentiable on* A*;*
  - *the first-order optimality (KKT) conditions are satisfied;*
  - *there is an open neighborhood* N_a *containing* a *for which the set of binding constraints* $I(a) := \{j : \ g_j(x(a), a) = 0\}$ *does not change.*

The Lagrangian for the optimization problem is

$$L(x, \lambda, a) \;=\; F(x, a) \;+\; \sum_k \lambda_k g_k(x, a), \tag{F.2}$$

where we have explicitly denoted its dependence on a. For each a, let $\lambda(a)$ denote a vector of Lagrange multipliers that satisfy the (KKT) conditions.

Theorem F.2. Envelope Theorem *Under Assumption 13,*

$$\frac{\partial M(a)}{\partial a_n} \;=\; \frac{\partial L(x(a), \lambda(a), a)}{\partial a_n}\,. \tag{F.3}$$

Proof. It is worthwhile to recall how Shephard's Lemma, Section 5.4.2, p. 77, was established. We began by using the chain rule, then used the facts that the gradient of the Lagrangian (with respect to x) must vanish, and that the constraint must be tight at the optimum. A similar approach will be used in this more general setting.

Fix $a \in \mathsf{A}$. From the chain rule[1]

$$\frac{\partial M}{\partial a_n} \;=\; \sum_i \frac{\partial F}{\partial x_i}\,\frac{\partial x_i}{\partial a_n} \;+\; \frac{\partial F}{\partial a_n}\,. \tag{F.4}$$

Since $x(a)$ and $\lambda(a)$ jointly satisfy the (KKT) conditions,

$$\frac{\partial F}{\partial x_i} \;=\; -\sum_k \lambda_k(a)\frac{\partial g_k}{\partial x_i} \;\text{ for all } i = 1, 2, \ldots K. \tag{F.5}$$

Substituting (F.5) into (F.4) and interchanging the order of summation, we obtain that

$$\frac{\partial M}{\partial a_n} \;=\; -\sum_k \lambda_k(a)\left[\sum_i \frac{\partial g_k}{\partial x_i}\,\frac{\partial x_i}{\partial a_n}\right] \;+\; \frac{\partial F}{\partial a_n}\,. \tag{F.6}$$

Since there is a neighborhood of a for which the set of tight constraints does not change,

$$\sum_i \frac{\partial g_k}{\partial x_i}\,\frac{\partial x_i}{\partial a_n} \;+\; \frac{\partial g_k}{\partial a_n} \;=\; 0 \;\text{ for all } k \in I(a), \tag{F.7}$$

which shows that the expression in the brackets for $k \in I(a)$ in (F.6) is identically $-\partial g_k/\partial a_n$. Since $\lambda_k(a) = 0$ when $k \notin I(a)$, we may substitute $-\partial g_k/\partial a_n$ for the expression in brackets for all k to obtain

$$\frac{\partial M}{\partial a_n} \;=\; \sum_k \lambda_k(a)\frac{\partial g_k}{\partial a_n} \;+\; \frac{\partial F}{\partial a_n} = \frac{\partial L}{\partial a_n}\,, \tag{F.8}$$

which completes the proof. □

[1] Keep in mind that the partial derivatives to follow are evaluated at the points a, $(x(a), a)$, and $(x(a), \lambda(a), a)$ where appropriate.

F.2 Application to Sensitivity Analysis of Cost

We consider the sensitivity analysis of the cost function $Q(u,p)$. Fix u and p and let $x^* = x(u,p)$ and $\lambda^* = \lambda(u,p)$. The Lagrangian of the cost minimization problem is

$$L(x,\lambda,p,u) \;=\; p\cdot x \;-\; \lambda(\Phi(x)-u). \tag{F.9}$$

As an immediate application of the Theorem F.2,

$$\frac{\partial Q(u,p)}{\partial p_i} \;=\; \frac{\partial L(x^*,\lambda^*,p,u)}{\partial p_i} \;=\; x_i^*, \tag{F.10}$$

and

$$\frac{\partial Q(u,p)}{\partial u} \;=\; \frac{\partial L(x^*,\lambda^*,p,u)}{\partial u} \;=\; \lambda^*\,. \tag{F.11}$$

F.3 A Monopoly Pricing Example

We consider a very special case of our general optimization problem in which the dimension of both x and a is one and there are no constraints. In such a case the Envelope theorem reduces to the statement that $M'(a) = \partial F/\partial a$. Here is an interpretation. When a changes, it *directly* affects $M(\cdot)$ via its direct effect on $F(\cdot,\cdot)$, but it also *indirectly* affects $M(\cdot)$ since the optimal choice $x(a)$ will change. Theorem F.2 says that since $x(a)$ is an optimal choice, *this indirect effect will be negligible.*

To make matters concrete, we shall consider a monopoly pricing problem and derive the Envelope theorem graphically. To this end, we consider a monopolist who faces the inverse demand curve

$$P \;=\; e - Q/4, \tag{F.12}$$

and whose marginal cost, c, is constant. The monopolist's profit function is $[e-Q/4]*Q \;-\; c*Q$, which may be equivalently expressed as

$$\pi(Q,a) :=\; a*Q \;-\; Q^2/4, \tag{F.13}$$

where we have set $a = e - c$. The monopolist, of course, will choose the value for output, $Q(a)$, to optimize $\pi(\cdot,\cdot)$, and we let $M(a) = \pi(Q(a),a)$ denote the corresponding optimal profit. We want to examine how $M(\cdot)$ varies with the parameter a. Note that an increase (decrease) in a due to an increase (decrease) in e shifts the inverse demand curve outward (inward), which will lead to an increase (decrease) in profit (and output) for the monopolist. The same effects will hold if there is a decrease (increase) in the unit marginal cost.

For each fixed value for Q, the profit function, viewed as a function $\pi_Q(\cdot)$ of a, is obviously *linear* and increasing in a. It intersects the y-axis at $-Q^2/4$,

the x-axis at $Q/4$, and the vertical line $x = a$ for any value of a. The optimal choice $Q(a)$ will be that value for Q whose line $y = \pi_Q(a)$ intersects the vertical line $x = a$ at the highest point.

Now fix values for a and Q, say at $\bar{a}$ and $\bar{Q}$. When will it be the case that $\bar{Q}$ equals $Q(\bar{a})$? For any Q the two lines $y = \pi_Q(a)$ and $y = \pi_{\bar{Q}}(a)$ intersect at the point $a = (Q + \bar{Q})/4$. Consequently,

$$\bar{Q} = Q(a) \implies \begin{cases} (Q + \bar{Q})/4 \leq \bar{a} & \text{if } Q \leq \bar{Q}, \\ (Q + \bar{Q})/4 \geq \bar{a} & \text{if } Q \geq \bar{Q}\ . \end{cases} \tag{F.14}$$

Since one choice for Q is $\bar{Q}$, it then follows from (F.14) that

$$Q(\bar{a}) = 2\bar{a} \ \text{ and } \ M(a) = a^2. \tag{F.15}$$

Obviously, $M'(a) = 2a$. More importantly, we see from (F.15) that $M'(a)$ also equals $Q(a)$, which just happens to equal the slope of the line $\pi_{Q(a)}(a)$. In other words,

$$M'(a) = \frac{\partial \pi(Q(a), a)}{\partial a}\ , \tag{F.16}$$

exactly as predicted by Theorem (F.2)!

F.4 Bibliographical Notes

Consult the graduate microeconomic textbooks previously mentioned.

G

Correspondence Theory

The cost function, $Q(u, p)$, and indirect production function, $\Gamma(p)$, define two *parametric* optimization problems. In the first case, the parameters are given by the required output, u, and prices, p, whereas in the second case the parameters are given simply by the prices p. It is often desirable to know how such functions vary with respect to their parameters. For example, are the cost and indirect production functions continuous in prices? It is also desirable to know how the optimal solutions of these optimization problems vary with respect to the parameters. For example, how do the cost-minimizing inputs or the output-maximizing inputs vary with respect to output and/or prices? In general, there can be several optimal solutions. Consequently, these questions cannot be answered by applying the usual notion of continuity of functions. For such analyses, a function must be replaced with the concept of a correspondence, and continuity of a function must be replaced with the concept of upper or lower hemicontinuity of a correspondence. Correspondences can be viewed as a point to set mapping. The input and output possibility sets of a technology are examples of correspondences.

In what follows, sets S and T will denote metric spaces. If it is useful, think of $S \subset \mathbb{R}^n$ and $T \subset \mathbb{R}^m$.

G.1 Core Concepts

Definition G.1. *A* **relation** ϕ *of a set* S *to* T *is a subset of* $S \times T$*. The* **domain** *of the relation* ϕ *is the set*

$$dom(\phi) := \{x \in S : \quad \text{there exists a } y \in T \text{ with } (x, y) \in \phi\}.$$

A relation ϕ *of* S *into* T *is a* **correspondence** *if the domain of* ϕ *is* S*. We shall identify the correspondence* ϕ *with a* **point to set** *mapping* $f_\phi : S \longrightarrow 2^T$ *defined by*[1]

[1] Here 2^Y denotes the power set of the set Y, namely, the collection of all subsets of Y.

$$f_\phi(x) := \{y \in T : (x, y) \in \phi\},$$

and with slight abuse of notation simply refer to f_ϕ *as* ϕ*. Thus,* $\phi(x) \subset T$ *for each* $x \in S$*.*

Definition G.2. *Let* $\phi(\cdot)$ *be a correspondence from* S *to* T*. Let* $A \subset S$ *and* $B \subset T$*.*

- *The* **image of** A **under** ϕ *is*

$$\phi(A) := \cup_{x \in S}\, \phi(x).$$

- *The* **image of a point** $x \in S$ **under** ϕ *is* $\phi(\{x\})$ *and will be denoted by* $\phi(x)$*.*
- ϕ *is* **single-valued** *if* $\phi(x)$ *is single-valued for each* $x \in S$*. A single-valued correspondence* ϕ *can be identified with a function from* S *into* T*, and with slight abuse of notation we shall denote this function by* ϕ*, too.*
- *The* **graph of** ϕ *is*

$$Gr(\phi) := \{(x, y) \in S \times T :\ y \in \phi(x)\}.$$

- *The* **upper or strong inverse** *of* B *is*

$$\phi^+(B) := \{x \in S : \phi(x) \subset B\}.$$

- *The* **lower or weak inverse** *of* B *is*

$$\phi^-(B) := \{x \in S : \phi(x) \cap B \text{ is not empty.}\}$$

Definition G.3. *A correspondence* ϕ *from* S *to* T *is* **upper hemicontinuous (u.h.c.) at** $x \in S$ *if* $\phi(x)$ *is nonempty, and for every open neighborhood* U *of* $\phi(x)$ *there exists a neighborhood* V *of* x *for which* $\phi(z) \subset U$ *for all* $z \in V$*. The correspondence* ϕ *is* **u.h.c.** *if it is u.h.c. at every* $x \in S$*.*

Theorem G.4. Characterization of u.h.c. *Let* ϕ *be a correspondence from* S *into* T*. The following assertions are equivalent:*

a) ϕ *is u.h.c.*
b) $\phi^+(G)$ *is open for every open set* $G \subset T$*.*
c) $\phi^-(F)$ *is closed for every closed set* $F \subset T$*.*

Proof. $(a) \Longrightarrow (b)$. Pick an open $G \subset T$ and an $x \in \phi^+(G)$. Since G is an open neighborhood of $\phi(x)$, by (a) we know there exists an open neighborhood V of x for which $\phi(V) \subset G$, which implies in particular that $V \subset \phi^+(G)$. Thus, $\phi^+(G)$ is open.

$(b) \Longrightarrow (a)$. Pick an $x \in S$ and let U be an open neighborhood of $\phi(x)$. By (b) the set $V := \phi^+(U)$ is an open neighborhood of x for which $\phi(z) \subset U$ for all $z \in V$, which implies ϕ is u.h.c. at x.

$(b) \Longleftrightarrow (c)$. The definitions of ϕ^+ and ϕ^- imply that

$$\phi^+(G) = (\phi^-(G^c))^c$$

holds for all $G \subset T$. (Here, the symbol 'c' denotes complement of the respective set in either S or T.) That is, $\phi(x) \subset G$ if and only if $\phi(x) \cap G^c = \emptyset$ if and only if $x \notin \phi^-(G^c)$. Now take G to be open and recall the complement of an open (closed) set is closed (open). □

Definition G.5. *Let $\phi(\cdot)$ be a correspondence from S to T.*

- *ϕ is* **closed at** *x if for every sequence $(x_n, y_n) \in S \times T$ for which $y_n \in \phi(x_n)$, $x_n \longrightarrow x$ and $y_n \longrightarrow y$, it follows that $y \in \phi(x)$.*
- *ϕ* **is closed** *if it is closed at every $x \in S$.*
- *ϕ* **is closed-valued** *if $\phi(x)$ is closed for every $x \in S$.*
- *ϕ* **is compact-valued** *if $\phi(x)$ is compact for every $x \in S$.*

A correspondence ϕ is closed if and only if its graph $Gr(\phi)$ is closed in $S \times T$. In particular, a closed correspondence is closed-valued. The converse is in general not true. If, on the other hand, the correspondence is u.h.c., then a closed-valued correspondence must also be closed, as the following proposition shows.

Proposition G.6. *Let $\phi(\cdot)$ be a correspondence from S to T.*

a) Suppose ϕ is closed-valued. If ϕ is u.h.c., then ϕ is closed.
b) Suppose T is compact. If ϕ is closed, then ϕ is u.h.c.

Proof. Part (a). We shall show that the complement of $Gr(\phi)$ (in S) is open. Pick a point $(x, y) \notin Gr(\phi)$. We need to find an open neighborhood N of (x, y) such that if $(x', y') \in N$, then $y' \notin \phi(x')$.

Since $y \notin \phi(x)$, a closed set, there exists an open sets A and B for which $y \in A \subset B$ and $B \cap \phi(x) = \emptyset$. Let U denote the complement of B (in T). Since U is open and ϕ is u.h.c. we know $V := \phi^+(U)$ is open, too. By construction, $\phi(z) \cap A = \emptyset$ for each $z \in V$. Thus, we have found an open neighborhood $N := V \times A$ that fulfills the requisite properties.

Part (b). If ϕ is not u.h.c., then we can find an $x \in S$, an open neighborhood U containing $\phi(x)$, and a sequence $x_n \longrightarrow x$ such that for each n a $y_n \in \phi(x_n)$ exists for which $y_n \notin U$. The complement of U is compact since T is compact, and so we may extract a convergent subsequence of the $\{y_n\}$'s whose limit point y obviously does not belong to U. Thus, ϕ is not closed. □

The following corollary summarizes the conditions under which u.h.c. is equivalent to closure of the graph.

Corollary G.7. *Let $\phi(\cdot)$ be a closed-valued correspondence from S to T and suppose T is compact. Then ϕ is u.h.c. if and only if ϕ is closed.*

Continuous functions preserve the topological property of compactness; that is, the continuous image of a compact set is compact. The following proposition shows that the same can be said for upper hemicontinuous correspondences.

Proposition G.8. *Let $\phi(\cdot)$ be a correspondence from S to T. If ϕ is u.h.c., then $\phi(E)$ is compact for every compact $E \subset S$.*

Proof. Let $E \subset S$ be compact and let $\{U_\alpha\}_{\alpha \in \mathcal{A}}$ be an open cover of $\phi(E)$, which is also an open cover of $\phi(x)$ for each $x \in S$. Since $\phi(x)$ is compact we may extract a finite subcover, and let U_x denote the union of the open sets in this finite subcover. The u.h.c. of ϕ ensures each $V_x := \phi^+(U_x)$ is open, and so $\{V_x\}$ as x ranges over E is an open cover of E. Since E is compact, we may extract a finite subcover, say $\{V_{x_i}\}_{i \in \mathcal{I}}$. Clearly,

$$\phi(E) \subset \phi(\bigcup_{i \in \mathcal{I}} V_{x_i}) = \bigcup_{i \in \mathcal{I}} \phi(V_{x_i}) \subset \bigcup_{i \in \mathcal{I}} U_{x_i}.$$

Since each U_{x_i} is a union of a finite number of the U_α's, we have shown a finite subset of the U_α's exists that covers $\phi(E)$, as required. □

G.2 Characterization by Sequences

Theorem G.9. Characterization of u.h.c. by sequences *The compact-valued correspondence ϕ from S to T is u.h.c. at x if and only if $\Phi(x)$ is nonempty, and for every sequence $x_n \longrightarrow x$ and every sequence $y_n \in \phi(x_n)$, there is a convergent subsequence of the y_n's whose limit y belongs to $\phi(x)$.*

Proof. ($\Longrightarrow$) Let E denote the collection of the x_n's plus x. E is compact since $x_n \longrightarrow x$, and so $\phi(E)$ is compact by Proposition G.8. Thus, we can extract a convergent subsequence of the y_n's with limit point y. Since a compact set is closed, ϕ is closed by Proposition G.6(a), which immediately implies $y \in \phi(x)$, as required.

($\Longleftarrow$) If the conclusion is false, then we can find an $x \in S$, an open neighborhood U containing $\phi(x)$, and a sequence $x_n \longrightarrow x$ such that for each n a $y_n \in \phi(x_n)$ exists for which $y_n \notin U$. By assumption, there exists a convergent subsequence of the y_n's with limit point $y \in \phi(x)$. However, this is not possible, since obviously $y \notin U$ and so $y \notin \phi(x)$. A contradiction has been reached. □

Definition G.10. *A correspondence ϕ from S to T is* **lower hemicontinuous (l.h.c.) at** *$x \in S$ if $\Phi(x)$ is nonempty, and for every open neighborhood U that meets $\phi(x)$ there exists a neighborhood V of x for which $\phi(z)$ meets U for all $z \in V$. The correspondence ϕ is* **l.h.c.** *if it is l.h.c. at every $x \in S$.*

The following theorem characterizes l.h.c. in a manner analogous to Theorem G.4. Its proof is left as an exercise.

Theorem G.11. Characterization of l.h.c. *Let ϕ be a correspondence from S into T. The following assertions are equivalent:*

a) ϕ is l.h.c.

b) $\phi^+(F)$ *is closed for every closed set* $F \subset T$.
c) $\phi^-(G)$ *is open for every open set* $G \subset T$.

Theorem G.12. Characterization of l.h.c. by sequences *The correspondence* ϕ *from* S *to* T *is l.h.c. at* x *if and only if* $\Phi(x)$ *is nonempty, and for every sequence* $x_n \longrightarrow x$ *and* $y \in \phi(x)$*, there exists a sequence* $y_n \in \phi(x_n)$ *for which* $y_n \longrightarrow y$.

Proof. ($\Longrightarrow$) Pick an $x \in S$, $y \in \phi(x)$ and a sequence $x_n \longrightarrow x$. Let U_k denote the open ball centered at y with radius of $1/k$ and let $V_k := \phi^-(U_k)$, $k = 1, 2, \ldots$. For each k an n_k exists for which $x_n \in U_k$ for all $n \geq n_k$. We may pick the n_k's to form an increasing sequence. For each integer $n < n_1$ pick $y_n \in \phi(x_n)$ arbitrarily, and use the l.h.c. of ϕ to pick a $y_n \in \phi(x_n) \cap V_k$ for each integer $n \in [n_k, n_{k+1})$. By construction, $y_n \in \phi(x_n)$ and obviously $y_n \longrightarrow y$.

($\Longleftarrow$) If the conclusion is false, then we can find an $x \in S$, an open neighborhood U that meets $\phi(x)$ and a sequence $x_n \longrightarrow x$ such that $\phi(x_n) \cap U = \emptyset$ for each n. Pick a $y \in \phi(x) \cap U$. The limit point of any convergent sequence of y_n's for which $y_n \in \phi(x_n)$ necessarily lies in the complement of U, and so cannot converge to y. A contradiction has been reached. □

Proposition G.13. *A single-valued correspondence* ϕ *from* S *to* T *is continuous if it is u.h.c. or l.h.c.*

Proof. A function is continuous if the inverse image of an open set is always open. The result is a direct consequence of Theorem G.4(b), Theorem G.11(c), and the fact that ϕ is single-valued. □

G.3 Bibliographical Notes

Hildebrand [1974] is the definitive reference concerning correspondence theory. Border [1985] provides a list of the main facts with some proofs.

H

Theorem of the Maximum

The following lemma we state without proof.

Lemma H.1. *Let $\{a_n\}$ be an infinite sequence of real numbers. If every subsequence of $\{a_n\}$ contains a subsequence that converges to the real number a, then $a_n \longrightarrow a$.*

Theorem H.2. Theorem of the Maximum *Let $f : S \times T \to \mathbb{R}$ be continuous, let γ be a compact-valued, continuous correspondence from T to S, and define*

$$m(\pi) := \max\{f(x, \pi) : \ x \in \gamma(\pi)\},$$
$$\mu(\pi) := \{x \in \gamma(\pi) : \ f(x, \pi) = m(\pi)\}.$$

a) The function $m : T \to \mathbb{R}$ is continuous.
b) $\mu(\cdot)$ is a compact-valued, u.h.c. correspondence from T to S.

Proof. Part (a). First note that $m(\cdot)$ is well-defined since $\gamma(\cdot)$ is compact-valued. Let $\pi_n \to \pi$. We shall show that $m(\pi_n) \to m(\pi)$.

Since $\gamma(\cdot)$ is u.h.c. we may extract a convergent subsequence $\{x_{n_k}\}$ whose limit point $x \in \gamma(\pi)$. Since $f(\cdot, \cdot)$ is continuous and x is feasible for the maximum problem defined by π, we have

$$m(\pi_{n_k}) = f(x_{n_k}, \pi_{n_k}) \to f(x, \pi) \leq m(\pi). \tag{H.1}$$

Pick a $z \in \gamma(\pi)$ and use the l.h.c. of $\gamma(\cdot)$ to find a sequence $z_n \to z$ for which $z_n \in \gamma(\pi_n)$. Since z_{n_k} is feasible for the maximum problem defined by π_{n_k},

$$f(z_{n_k}, \pi_{n_k}) \leq m(\pi_{n_k}). \tag{H.2}$$

Since the left-hand side of (H.2) converges to $f(z, \pi)$ and z was chosen arbitrarily, it follows from (H.1) that

$$m(\pi) = \max\{f(z, \pi) : \ z \in \gamma(\pi)\} \leq f(x, \pi) \leq m(\pi),$$

and thus

$$m(\pi_{n_k}) = f(x_{n_k}, \pi_{n_k}) \to f(x, \pi) = m(\pi).$$

We have shown the sequence of real numbers $\{m(\pi_n)\}$ contains a subsequence that converges to $m(\pi)$. The arguments above apply to any subsequence of the $\{\pi_n\}$'s, and so the result now follows from Lemma H.1.

Part (b). $\mu(\cdot)$ is a correspondence since $\gamma(\cdot)$ is compact-valued. It is closed-valued since $f(\cdot, \cdot)$ is continuous, and thus compact-valued, since closed subsets of compact sets are compact. We shall use Theorem G.12, p. 499, to prove u.h.c. of $\mu(\cdot)$.

To this end, let $\pi_n \to \pi$ and $x_n \in \mu(\pi_n)$. Using the l.h.c. of $\gamma(\cdot)$ we may extract a subsequence $\{x_{n_k}\}$ for which $x_{n_k} \to x \in \gamma(\pi)$. Since both $f(\cdot, \cdot)$ and $m(\cdot)$ are continuous (by part a),

$$m(\pi_{n_k}) = f(x_{n_k}, \pi_{n_k}) \to f(x, \pi) = m(\pi),$$

which shows that $x \in \mu(\pi)$, as required by Theorem G.12. □

H.1 Application to the Indirect Production Function

The indirect production function is defined as

$$\Gamma(p) := \max\{\Phi(x) : \ p \cdot x \leq 1\}, \tag{H.3}$$

and measures the maximum output (utility) that can be achieved when there is a budget constraint. In (H.3), we assume that all prices are positive and that $\Phi(\cdot)$ is continuous.

Definition H.3. *The* **budget correspondence** $\mathcal{B}$ *from* $\mathbb{R}^k_{++}$ *into* $\mathbb{R}^k_+$ *is defined as*

$$\mathcal{B}(p) := \{x \in \mathbb{R}^k_+ : \ p \cdot x \leq 1\}. \tag{H.4}$$

Definition H.4. *The* **Marshallian demand correspondence** D^M *from* $\mathbb{R}^k_{++}$ *into* $\mathbb{R}^k_+$ *is defined as*

$$D^M(p) := \{z \in \mathcal{B}(p) : \ \Gamma(p) = \Phi(z)\}.$$

If $\Phi(\cdot)$ is also strictly quasiconcave, namely,

$$\Phi((\lambda x + (1 - \lambda)y) > \min\{\Phi(x), \Phi(y)\}$$

for each $x, \ y \in \mathbb{R}^k_+$ and $\lambda \in (0, 1)$, then there must be a *unique* maximum in (H.3), which would imply that $D^M(\cdot)$ is single-valued. (Otherwise, a non-trivial convex combination of two distinct maximizers would be budget feasible and would yield a higher output.) If it can be shown that $\mathcal{B}(\cdot)$ is continuous, then as a direct application of the Theorem of the Maximum H.2 it follows that $\Gamma(\cdot)$ is continuous and $D^M(\cdot)$ is u.h.c. and thus continuous by (G.13).

Proposition H.5. *The correspondence $\mathcal{B}(\cdot)$ in (H.4) is continuous.*

Proof. Clearly, $\mathcal{B}(\cdot)$ is compact-valued, and so we shall use Theorems G.9 and G.12 to establish the u.h.c. and l.h.c., respectively.

To establish u.h.c., let $p_n \to p$ and $x_n \in \mathcal{B}(p_n)$. $\mathcal{B}(\cdot)$ is closed since the dot product is a continuous function of its arguments. By Theorem G.9, it is sufficient to show the x_n's are bounded. Since the coordinates of p are positive and p_n converges to p, the coordinates of each p_n are eventually bounded below by a positive number. Since $p_n \cdot x_n \leq 1$ for all n, it easily follows the x_n's are bounded, as required.

To establish l.h.c., let $p_n \to p$ and pick $x \in \mathcal{B}(p)$. Without loss of generality we shall assume $p \cdot x = 1$. (Divide each p_n and p by $p \cdot x$.) For each n and coordinate i, $1 \leq i \leq k$, define $x_n^i = (p^i/p_n^i)x^i$. By construction,

$$p_n \cdot x_n = \sum_i p_n^i x_n^i = p \cdot x = 1,$$

and so $x_n \in \mathcal{B}(p_n)$. Clearly, $x_n \to x$, as required by Theorem G.12. □

H.2 Application to the Cost Function

Let $\mathcal{F} = \{L_\Phi(u) : \ u \geq 0\}$ be a well-behaved technology. In addition to the usual properties on $\Phi(\cdot)$ that hold for a well-behaved technology, in this section we make the following assumption:

Assumption 14 *$\Phi(\cdot)$ is continuous, strictly quasiconcave and increasing.*

The cost function is defined as

$$Q(u,p) = \min\{p \cdot x : \ \Phi(x) \geq u\}. \tag{H.5}$$

If all prices are positive, the cost function has a minimum. This is because it is possible to constrain the feasible region of (H.5) to the compact domain

$$\{x \in \mathbb{R}_+^k : \ p \cdot x \leq p \cdot e(u)\},$$

where $e \in \mathbb{R}^k$ denotes the vector whose coordinates are identically 1 and where $e(u) := e/\mathcal{D}(e,u)$ for all $u > 0$. (Here $\mathcal{D}(\cdot,\cdot)$ is the input distance function for $\mathcal{F}$.) If some prices are zero, however, then an additional property must be assumed to ensure that (H.5) has a minimum. In this case, we assume that the Efficient Frontier is bounded, see Axiom **A.5**, p. 38.

Definition H.6. *The* **cost feasible correspondence** γ *from* $\mathbb{R}_+ \times \mathbb{R}_{++}^k$ *into* $\mathbb{R}_+^k$ *defined as*

$$\gamma(u,p) := \{x \in \mathbb{R}_+^k : \ \Phi(x) \geq u \text{ and } p \cdot x \leq p \cdot e(u)\}.$$

Definition H.7. *The* **Hicksian demand correspondence** D^H *from* $\mathbb{R}^k_{++}$ *into* $\mathbb{R}^k_+$ *is defined as*

$$D^H(p) := \{z \in \gamma(u,p) : \ Q(u,p) = p \cdot z\}.$$

Since $\Phi(\cdot)$ is strictly quasiconcave and increasing, there must be a *unique* maximum in (H.5), which would imply that $D^H(\cdot)$ is single-valued.[1] If it can be shown that $\gamma(\cdot,\cdot)$ is continuous, then as a direct application of the Theorem of the Maximum (see H.2, p. 501), it follows that $Q(\cdot,\cdot)$ is jointly continuous and $D^H(\cdot)$ is u.h.c. and thus continuous by (G.13).

Under the conditions herein the input distance function is continuous in u. We state the following Lemma without proof.

Lemma H.8. *Under Assumption 14, for each* $x \in \mathbb{R}^k_{++}$ *the function* $f(u) := \mathcal{D}(x,u)^{-1}$ *is continuous in* u.

Proposition H.9. *The correspondence* γ *from* $\mathbb{R}_+ \times \mathbb{R}^k_{++}$ *into* $\mathbb{R}^k_+$ *defined by*

$$\gamma(u,p) := \{x \in \mathbb{R}^k_+ : \ \Phi(x) \geq u \text{ and } p \cdot x \leq p \cdot e(u)\}$$

is continuous.

Proof. Clearly, $\gamma(\cdot,\cdot)$ is compact-valued, and so we shall use Theorems G.9 and G.12 to establish the u.h.c. and l.h.c., respectively. Let $p_n \to p$, $u_n \to u$, and define $b_n := p_n \cdot e(u_n)$ for each n and $b := p \cdot e(u)$. Note that $b_n \to b$ by Lemma H.8. Let $\hat{p}_n = p_n/b_n$ and $\hat{p} = p/b$. Of course $\hat{p_n} \to \hat{p}$. With this notation,

$$\gamma(u_n, p_n) = L(u_n) \cap \mathcal{B}(\hat{p}_n).$$

To establish u.h.c., let $x_n \in \gamma(u_n, p_n)$. Since the correspondence $\mathcal{B}(\cdot)$ in Proposition H.5 is u.h.c. we may extract a subsequence $\{x_{n_k}\}$ whose limit point $x \in \gamma(\hat{p})$. The continuity of $\Phi(\cdot)$ ensures $x \in L(u)$ since $\Phi(x_{n_k}) \geq u_{n_k}$ for each k. Thus, a convergent subsequence of the x_n's has been found whose limit point lies in $\gamma(u,p)$, as required by Theorem G.9.

To establish l.h.c., pick $x \in \gamma(u,p)$. Since $\gamma(\cdot,\cdot)$ is compact-valued, for each n a vector z_n exists that minimizes the distance from x to $\gamma(u_n,p_n)$. Since $z_n \in \gamma(u_n,p_n)$, and we have proved that $\gamma(\cdot,\cdot)$ is u.h.c., it follows that we may extract a convergent subsequence $\{z_{n_k}\}$ whose limit point $z \in \gamma(u,p)$. The result will follow if we can show that $z = x$.

If this were not so, then $y := (x+z)/2 \neq x$ and $y \in \gamma(u,p)$ since $\gamma(u,p)$ is convex. For $\delta > 0$ let $y_\delta := y/(1+\delta)$. The value $\Phi(y)$ exceeds u since $\Phi(\cdot)$ is strictly quasiconcave. Let $\epsilon := (\Phi(y) - u)/2$. There exists an N_ϵ for which

$$\Phi(y) > u + \epsilon \geq u_n, \text{ for all } n \geq N_\epsilon. \tag{H.6}$$

[1] Otherwise, it would be possible to scale down a non-trivial convex combination of two distinct minimizers to still achieve the required output but at a lower cost.

Since $\Phi(\cdot)$ is increasing and continuous,

$$\Phi(y_\delta) \geq u + \epsilon \tag{H.7}$$

for δ sufficiently small. Using the triangle inequality, we have

$$||y_\delta - x|| \leq \frac{1}{1+\delta}(\ 1/2\ ||z - x|| + \delta|\ x||\ ). \tag{H.8}$$

It follows from (H.8) that we may pick a δ sufficiently small for which

$$||y_\delta - x|| < (2/3) \cdot ||z - x|| \tag{H.9}$$

and (H.7) hold, and we now assume δ has been so chosen. It follows from (H.9) (and the continuity of the norm) that an M exists for which

$$||y_\delta - x|| < (3/4) \cdot ||z_n - x|| \text{ for all } n \geq M. \tag{H.10}$$

Since $p_n \to p$ an N_δ exists for which

$$\hat{p}_n \cdot x \leq \hat{p} \cdot x + \delta \text{ for all } n \geq N_\delta,$$
$$\hat{p}_n \cdot z \leq \hat{p} \cdot z + \delta \text{ for all } n \geq N_\delta.$$

It is clear that $\hat{p}_n y_\delta \leq 1$ for all $n \geq N_\delta$. Let $N = \max\{N_\epsilon, N_\delta, M\}$. From (H.6) and (H.10), we have found a point $y_\delta \in \gamma(u_N, p_N)$ that is closer to x than z_N is to x, an obvious contradiction. □

H.3 Bibliographical Notes

Hildebrand [1974] contains a thorough and abstract development of correspondence theory and the Theorem of the Maximum. Border [1985] provides a somewhat condensed, less abstract treatment. Starr's [1997] presentation is more focused but very accessible.

I

Probability Basics

In this chapter, we review only the basic material on probability that is required for understanding the models presented in Chapter 20.

I.1 Binomial Random Variables

Suppose n independent experiments or trials are performed, each of which results in a "success" or "failure." The probability of success is a constant $p \in (0, 1)$ and the probability of failure is $1 - p$. Let X represent the number of successes that occur in these n experiments. The random variable X is said to have the **binomial distribution with parameters** n **and** p. Its probability mass function is

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k},\ k = 0, 1, 2, \ldots, n\ .$$

Here

$$\binom{n}{k} := \frac{n!}{k!(n-k)!}$$

counts the number of unordered groups of k objects that can be selected from a set of n objects.

For a binomially distributed random variable X with parameters n and p, its mean, $E[X]$, and variance, $Var[X] = E[(X - E[X])^2]$, are

$$E[X] = np, \tag{I.1}$$

$$Var[X] = np(1-p). \tag{I.2}$$

I.2 Poisson Random Variables

A **Poisson random variable X with mean** $\lambda > 0$ has the probability mass function

$$P(X = i) = e^{-\lambda}\frac{\lambda^i}{i!},\ i = 0, 1, 2, \ldots.$$

The mean, $E[X]$, and variance, $Var[X]$, both equal λ. It turns out that the Poisson probability mass function closely approximates a binomial probability mass function under the following conditions. Let X_1 denote a binomial random variable with parameters n and p such that n is "large," p is "small" and their product $\lambda = np$ is of "moderate" magnitude, and let X_2 denote a Poisson random variable with parameter λ. Then $P(X_1 = k) \approx P(X_2 = k)$.

I.3 Poisson Processes

A stochastic process $\{N(t),\ t \geq 0\}$ is a **counting process** if $N(t)$ records the total number of occurrences of an event up to time t (e.g., customer orders, arrivals to a queue). Let $T_1 < T_2 < \ldots$, denote the event occurrence times, often called the arrival times. (Assume the arrival times are distinct.) $N(t)$ simply counts the number of T_i in the time interval $[0, t]$, and $N(t) - N(s)$ counts the number of arrivals in the interval $(s, t]$. A counting process has **independent increments** if the number of arrivals in disjoint time intervals are independent random variables.

A **Poisson process** $N(t)$ is a counting process that has independent increments and $N(t) - N(s)$, $0 \leq s < t$, is a Poisson random variable with mean $\Lambda(t) - \Lambda(s)$, where $\Lambda(t) = E[N(t)]$. We will assume that

$$\Lambda(t) = \int_0^t \lambda(\tau)d\tau,$$

where $\lambda(\cdot)$ is the **intensity function** of the Poisson process. The Poisson process is **time-homogenous with rate** λ if the intensity function $\lambda(\cdot) = \lambda$ is constant. In such a case $\Lambda(t) = \lambda t$ and $E[N(t)] = \Lambda(t) = \lambda t$.

A Poisson process $N(t)$ has the following powerful property that actually uniquely characterizes it. Conditioned on $N(t) - N(s) = n$ arrivals in the interval $(s, t]$, the *unordered* set of arrival times has the same distribution as n independent and identically distributed random variables having cumulative distribution function

$$G(\tau) := \begin{cases} \frac{\Lambda(\tau)-\Lambda(s)}{\Lambda(t)-\Lambda(s)}, & s \leq \tau \leq t, \\ 1, & \tau > t. \end{cases} \tag{I.3}$$

(When $s = 0$, $\Lambda(s) = 0$, too.) Here is an interpretation. Imagine a system observer tasked with observing the arrival times. He faithfully record the times on separate index cards, but not the event index. (The observer keeps track of the event index.) He informs us there were n index cards obtained in the interval $[0, t]$, and he hands us a sealed envelope containing an index card randomly chosen from the set of n index cards. Let $\tilde{T}$ denote the time listed

on this first index card. The property above implies that $P(\tilde{T} \leq \tau) = G(\tau)$. If we were handed a second index card and then a third, and so on, these times would have the same distribution as $G(\tau)$.

I.4 Moment Generating Functions

The **moment generating function**

$$\phi_X(u) := E[e^{uX}]$$

(for u in some neighborhood of 0) of a random variable X is so named because all of the *moments* of X, namely, $E[X^k]$, $k = 1, 2, \ldots$, can be obtained by successively differentiating $\phi_X(t)$ and evaluating it at zero. For example, $E[X] = \phi'_X(0)$ and $E[X^2] = \phi''_X(0)$. Since the variance, $Var[X]$, of X equals $E[X^2] - (E[X])^2$, it too can be obtained from the moment generating function.

It turns out that the moment generating function of a random variable X *uniquely* determines its distribution. That is, if $\phi_X(\cdot)$ matches a known functional form for a certain type of distribution, then X must have this type of distribution with the parameters identified through its moment generating function.

The discrete random variable X has the binomial distribution with parameters n and p if and only if its moment generating function is

$$\phi_X(u) = [1 + p\,(e^u - 1)]^n. \tag{I.4}$$

A random variable X has the Poisson distribution with mean $\lambda > 0$ if and only if its moment generating function is

$$\phi_X(u) = e^{\lambda(e^u - 1)}. \tag{I.5}$$

I.5 Conditional Expectation and Variance

Let $E[X \mid Y]$ denote the function of the random variable Y whose value at $Y = y$ is $E[X \mid Y = y]$. Note that $E[X \mid Y]$ is itself a random variable. A fundamental property of conditional expectation that we repeatedly exploit is that for all random variables X and Y

$$E[X] = E\left(E[X \mid Y]\right). \tag{I.6}$$

If Y is a discrete random variable, then (I.6) states that

$$E[X] = \sum_y E[X \mid Y = y]\, P(Y = y),$$

whereas if Y is a continuous random variable with probability density function $f_Y(y)$, then (I.6) states that

$$E[X] = \int_{-\infty}^{\infty} E[X \mid Y = y]\ f_Y(y)dy.$$

The conditional variance of X given the random variable Y is defined by

$$Var[X \mid Y] := E\left[(X - E[X \mid Y])^2 \mid Y\right].$$

It can be shown that the *unconditioned* variance of X can be expressed as

$$Var[X] = E\left(Var[X \mid Y]\right) \ + \ Var\left(E[X \mid Y]\right). \tag{I.7}$$

I.6 Bibliographical Notes

For additional background material on stochastic processes, consult Ross [1985]. For a definitive reference on point processes, consult Serfozo [1990].

References

1. S.N. Afriat. The construction of utility functions from expenditure data. *International Economic Review*, 8:67–77, 1967.
2. S.N. Afriat. Efficiency estimation of production functions. *International Economic Review*, 13:568–598, 1967.
3. G. Alon, M. Beenstock, S.T. Hackman, U. Passy, and A. Shapiro. Nonparametric estimation of concave production technologies by entropic methods. *Journal of Applied Econometrics*, 22:795–816, 2007.
4. J. Asmundsson, R.L. Rardin, and R. Uzsoy. Tractable nonlinear production planning models for semiconductor wafer fabrication facilities. *IEEE Transactions on Semiconductor Manufacturing*, 19:95–111, 2006.
5. R.D. Banker. Estimating most productive scale size using data envelopment analysis. *European Journal of Operational Research*, 17:35–44, 1984.
6. R.D. Banker, A. Charnes, and W.W. Cooper. Models for estimation of technical and scale inefficiencies in data envelopment analysis. *Management Science*, 30:1078–1092, 1984.
7. R.D. Banker and A. Maindiratta. Nonparametric analysis of technical and allocative efficiencies in production. *Econometrica*, 56:1315–1332, 1988.
8. R.D. Banker and R. Morey. Efficiency analysis for exogenously fixed inputs and outputs. *Operations Research*, 34:512–521, 1984.
9. R.D. Banker and R. Morey. The use of categorical variables in data envelopment analysis. *Management Science*, 32:1613–1627, 1986.
10. J.J. Bartholdi and S.T. Hackman. *Warehouse and Distribution Science*. 2007. Available on line at http://www.warehouse-science.com.
11. R.G. Bartle. *The Elements of Real Analysis, 2nd Edition*. John Wiley and Sons, New York, NY, 1976.
12. M.J. Bazarra, H.D. Sherali, and C.M. Shetty. *Nonlinear Programming: Theory and Algorithms, 2nd Edition*. John Wiley and Sons, New York, NY, 1993.
13. C. Blackorby, D. Primont, and R.R. Russell. *Duality, Separability and Functional Structure: Theory and Economic Applications*. North-Holland, New York, NY, 1978.
14. K.C. Border. *Fixed Point Theorems with Applications to Economics and Game Theory*. Cambridge University Press, Cambridge, Great Britain, 1985.
15. S. Bouhnik, B. Golany, S.T. Hackman, U. Passy, and D.A. Vlatsa. Lower bound restrictions on intensities in data envelopment analysis. *Journal of Productivity Analysis*, 16:241–261, 2001.

16. M. Bunge. *Metascientific Queries*. Charles C. Thomas Publishers, Springfield, IL, 1959.
17. M. Bunge. *Method, Model and Matter*. D. Reidel Publishing Co., Boston, MA, 1973.
18. D.W. Caves, L.R. Christensen, and W.E. Diewert. The economic theory of index numbers of the measurement of input, output and productivity. *Econometrica*, 50:1393–1414, 1982.
19. R.G. Chambers. *Applied Production Analysis*. Cambridge University Press, New York, NY, 1988.
20. A. Charnes, W. Cooper, A.Y. Lewin, and L.M. Seiford, editors. *Data Envelopment Analysis; Theory, Methodology and Applications*. Kluwer Academic Publishers, Norwell, MA, 1996.
21. A. Charnes and W.W. Cooper. *Management Models and Industrial Applications*. John Wiley and Sons, New York, NY, 1961.
22. A. Charnes, W.W. Cooper, and E. Rhodes. Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2:429–444, 1984.
23. W.F. Christopher and C.G. Thor, eds. *Handbook for Productivity Measurement and Improvement*. Productivity Press, Portland, Ore, 1993.
24. W.W. Cooper, L.M. Seiford, and K. Tone. *DEA: A Comprehensive Text with Models, Applications, References and DEA-Solver Software*. Springer-Verlag, New York, NY, 2007.
25. T.J. Coelli, D.S. Prasada Rao, C.J. O'Donnell, and G.E. Battese. *An Introduction to Efficiency and Productivity, second edition*. Springer-Verlag, New York, NY, 2005.
26. G.B. Dantzig. *Linear Programming and Its Extensions*. Princeton University Press, Princeton, NJ, 1963.
27. F. de Mateo, T. Coelli and C. O'Donnell. Optimal paths and costs of adjustment in dynamic DEA models: with application to chilean department stores. *Annals of Operations Research*, 145: 211–227, 2006.
28. G. Debreu. The coefficient of resource utilization. *Econometrica*, 19:273–292, 1951.
29. G. Debreu. *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*. Cowles Foundation, Yale University Press, New Haven, CT, 1959.
30. A.G. de Kok and S.C. Graves. *Handbooks in Operations Research and Management Science on Supply Chain Management: Design, Coordination and Operation, Vol. 11*. Elsevier, 2003.
31. W.E. Diewert. Duality approaches to microeconomic theory. In K.J. Arrow and M.J. Intrilligator, editors, *Handbook of Mathematical Economics, Volume II*, pp. 535–599. North-Holland, New York, NY, 1982.
32. W.E. Diewert and C. Montmarquette, editors. *Price Level Measurement: Proceedings from a Conference Sponsored by Statistics Canada*, Ottawa, 1983. Statistics, Canada.
33. W.E. Diewert and A.O. Nakamura, editors. *Essays in Index Number Theory, Volume I*. North-Holland, Amsterdam, 1993.
34. W.E. Diewert and C. Parkan. Linear programming tests of regularity conditions for production functions. In W. Eichorn, R. Heen, K. Neumann, and R.W. Shephard, editors, *Quantitative Studies on Production and Prices*, pp. 131–158. Physica-Verlag, Vienna, 1983.
35. A.K. Dixit. *Optimization in Economic Theory*. Oxford University Press, New York, NY, 1976.

36. R. Fare. *Fundamentals of Production Theory*. Springer-Verlag, New York, NY, 1988.
37. R. Fare and S. Grosskopf. *Intertemporal Production Frontiers: With Dynamic DEA*. Kluwer Academic Publishers, Norwell, MA, 1996.
38. R. Fare, S. Grosskopf, and C.A.K. Lovell. *The Measurement of Effiency of Production*. Kluwer-Nijhoff Publishing, Boston, MA, 1985.
39. R. Fare, S. Grosskopf, and C.A.K. Lovell. *Production Frontiers*. Cambridge University Press, Cambridge, Great Britain, 1994.
40. R. Fare, S. Grosskopf, M. Norris, and Z. Zhang. Productivity growth, technical progress, and efficiency change in industrialized countries. *American Economic Review*, 84:66–83, 1994.
41. R. Fare and C.A.K. Lovell. Measuring the technical efficiency of production. *Journal of Economic Theory*, 19:250–262, 1978.
42. M.J. Farrell. The measurement of productive efficiency. *Journal of the Royal Statistical Society Series A, General*, 120:253–281, 1957.
43. M.J. Farrell and M. Fieldhouse. Estimating efficient production functions under increasing returns to scale. *Journal of the Royal Statistical Society Series A, General*, 125:252–267, 1962.
44. Z. First, S.T. Hackman, and U. Passy. Matrix-criteria for the pseudo p-convexity of a quadratic form. *Linear Algebra and Its Applications*, 136:235–255, 1990.
45. Z. First, S.T. Hackman, and U. Passy. Local-global properties of bi-functions. *Journal of Optimization Theory and Applications*, 73:279–297, 1992.
46. Z. First, S.T. Hackman, and U. Passy. Efficiency estimation and duality theory for nonconvex technologies. *Journal of Mathematical Economics*, 22:295–307, 1993.
47. E.H. Frazelle. *World-Class Warehousing and Materials Handling*. McGraw-Hill, New York, NY, 2001.
48. E.H. Frazelle and S.T. Hackman. The warehouse performance index: A single-point metric for benchmarking warehouse performance. Technical Report TR-93-14, Georgia Institute of Technology, 1993.
49. H. O. Fried, C.A.K. Lovell, and S. S. Schmidt, editors. *The Measurement of Productive Efficiency: Techniques and Applications*. Oxford University Press, Oxford, Great Britain, 1993.
50. M. Fuss and D. McFadden, editors. *Production Economics: A Dual Approach to Theory and Applications, Volumes I and II*. North-Holland, Amsterdam, 1978.
51. I.M. Gelfand and S.V. Fomin. *Calculus of Variations*. Dover Publications, New York, NY, 1963.
52. R. Gibbons. *Game Theory for Applied Economists*. Princeton University Press, Princeton, NJ, 1992.
53. B. Golany, S.T. Hackman, and U. Passy. An efficiency measurement framework for multi-stage production systems. *Annals of Operations Research*, 145:51–68, 2006.
54. S.C. Graves. A tactical planning model for a job shop. *Operations Research*, 34:552–533, 1986.
55. S.C. Graves, A.H.G. Rinnoy Kan, and P.H. Zipkin, editors. *Handbooks in Operations Research and Management Science on Logistics of Production and Inventory, Vol. 4*. North-Holland, Amsterdam, 1993.

56. S. Grosskopf. Efficiency and productivity. In H. O. Fried, C.A.K. Lovell, and S. S. Schmidt, editors, *The Measurement of Productive Efficiency: Techniques and Applications*, pp. 160–194. Oxford University Press, Oxford, Great Britain, 1993.
57. S.T. Hackman. A new, geometric proof of Shephard's duality theorem. *Journal of Economics*, 46:299–304, 1986.
58. S.T. Hackman. An axiomatic framework of dynamic production. *Journal of Productivity Analysis*, 1:309–324, 1990.
59. S.T. Hackman, E.H. Frazelle, P. Griffin, S.O. Griffin, and D.A. Vlatsa. Benchmarking warehousing and distribution operations: An input-output approach. *Journal of Productivity Analysis*, 16:79–100, 2001.
60. S.T. Hackman and R.C. Leachman. An aggregate model of project-oriented production. *IEEE Transactions on Systems, Man and Cybernetics*, 19:220–231, 1989.
61. S.T. Hackman and R.C. Leachman. A general framework for modeling production. *Management Science*, 35:478–495, 1989.
62. S.T. Hackman and U. Passy. Projectively-convex sets and functions. *Journal of Mathematical Economics*, 17:55–68, 1988.
63. S.T. Hackman and U. Passy. Maximizing a linear fractional function over the efficient frontier. *Journal of Optimization Theory and Applications*, 113:83–103, 2002.
64. S.T. Hackman, L.K. Platzman, and U. Passy. Explicit representation of a two-dimensional section of a production possibility set. *Journal of Productivity Analysis*, 5:161–170, 1994.
65. S.T. Hackman and R.R. Russell. Duality and continuity. *Journal of Productivity Analysis*, 6:99–116, 1995.
66. G. Hanoch and M. Rothschild. Testing the assumptions of production theory: a nonparametric approach. *Journal of Political Economy*, 80:256–275, 1972.
67. P.T. Harker, ed. *The Service Productivity and Quality Challenge.* Kluwer Academic Publishers, Norwell, MA, 1995.
68. W. Hildenbrand. *Core and Equilibria of a Large Economy.* Princeton University Press, Princeton, NJ, 1974.
69. G.A. Jehle and P.J. Reny. *Advanced Microeconomic Theory, 2nd Edition.* Addison-Wesley, New York, NY, 2001.
70. C.H. Jenkins. *Complete Guide to Modern Warehouse Management.* Prentice-Hall, New York, 1990.
71. J.L. Kelly. *General Topology.* Van Nostrand, New York, NY, 1955.
72. J. Kendrick. *Improving Company Productivity: Handbook with Case Studies.* The Johns Hopkins University Press, Baltimore, MD, 1984.
73. A.N. Kolmogorov and S.V. Fomin. *Introductory Real Analysis.* Dover Publications, New York, NY, 1970.
74. A.A. Konus. The problem of the true index of the cost of living. *Econometrica*, 7:10–29, 1939.
75. T.C. Koopmans. An analysis of production as an efficient combination of activities. In T.C. Koopmans, editor, *Activity Analysis of Production and Allocation.* Cowles Commission for Research in Economics, Monograph No. 13, New York, NY, 1951.
76. W.W. Leontief. *The Structure of the American Economy.* Oxford University Press, New York, NY, 1953.

77. S. Malmquist. Index numbers and indifference surfaces. *Trabajos de Estatistica*, 4:209–242, 1953.
78. A. Mas-Colel, M.D. Whinston, and J.R. Green. *Microeconomic Theory*. Oxford University Press, Oxford, Great Britain, 1995.
79. L.F. McGinnis, A. Johnson, and M. Villareal. Benchmarking Warehouse Performance Study. W.M. Keck Virtual Factory Lab, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332-0205.
80. D.E. Mulcahy. *Warehouse Distribution and Operations Handbook*. McGraw-Hill, New York, NY, 1993.
81. J. Pahl, S. Vos, and D.L. Woodruff. Production planning with load dependent lead times: an update of research. *Annals of Operations Research*, 153:297–345, 2007.
82. N.C. Petersen. Data envelopment analysis on a relaxed set of assumptions. *Management Sciences*, 36:305–314, 1990.
83. A. Rapoport. *N-Person Game Theory*. University of Michigan Press, Ann Arbor, MI, 1970.
84. M.J. Realff, J.C. Ammons, and D.J. Newton. Robust reverse production system design for carpet recycling. *IIE Transactions,* 36:767–776, 2004.
85. S.A. Reveliotis. *Real-Time Management of Resource Allocation Systems: A Discrete-Event Systems Approach*. International Series in Operations Research and Management Science. Springer, New York, NY, 2004.
86. G. Riano. *Transient Behavior of Stochastic Networks: Applications to Production Planning with Load-Dependent Lead Times*. PhD thesis, Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, 2002.
87. G. Riano, R.F. Serfozo, and S.T. Hackman. Transient behavior of queuing networks. Working paper, 2007.
88. R.T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.
89. S. Ross. *Introduction to Probability Models*. Academic Press, New York, NY, 1985.
90. H.L. Royden. *Real Analysis, 2nd Edition*. MacMillan Publishing Co., New York, NY, 1985.
91. R.R. Russell. Measures of technical efficiency. *Journal of Economic Theory*, 35:109–126, 1985.
92. M. Schefczyk. Industial benchmarking: A case-study of performance analysis techniques. *International Journal of Production Economics*, 32:1–11, 1993.
93. R. Schmalensee and R. Willig, editors. *Handbook of Industial Organization, Volume I*. North-Holland, Amsterdam, 1989.
94. J.K. Sengupta. *Efficiency Analysis By Production Frontiers: The Nonparametric Approach*. Kluwer Academic Publishers, Norwell, MA, 1989.
95. R.F. Serfozo. Point processes. In D.P. Heyman and M.J. Sobel, editors, *Stochastic Models*, pp. 1–93. Elsevier Science/North-Holland, Amsterdam, 1990.
96. R.W. Shephard. *Cost and Production Functions*. Princeton University Press, Princeton, NJ, 1953.
97. R.W. Shephard. *Theory of Cost and Production Functions*. Princeton University Press, Princeton, NJ, 1970.

98. R.W. Shephard, R. Al-Ayat, and R.C. Leachman. Shipbuiding production function: an example of a dynamic production function. In *Quantitative Wirtschaftsforschung Festschrift Volume (Wilhelm Krelle)*, H. Alback (ed.), J.B.C. Mohr, Tubingen, 1977.
99. R.W. Shephard and R. Fare. *Dynamic Theory of Production Correspondences.* Oelgeschlager, Gunn and Hain, Cambridge, MA, 1980.
100. O. Shy. *Industrial Organization.* M.I.T. Press, Cambridge, MA, 1995.
101. D. Simchi-Levi, P. Kaminsky, and E. Simchi-Levi. *Designing and Managing the Supply Chain, Second Edition.* McGraw-Hill, Boston, MA, 2002.
102. R. Solow. Technical change and the aggregate production function. *Review of Economic and Statistics*, 39:312–320, 1985.
103. R. Starr. *General Equilibrium Theory: An Introduction.* Cambridge University Press, Cambridge, Great Britain, 1997.
104. G.J. Stigler. The Xistence of X-efficiency. *American Economic Review*, 66:213–216, 1976.
105. J. Tirole. *The Theory of Industrial Organization.* M.I.T. Press, Cambridge, MA, 1988.
106. J.A. Tompins and J. Smith. *Warehouse Management Handbook, second edition.* Tompkins Press, 1998.
107. L. Tornqvist. The bank of finland's consumption price index. *Bank of Finland Monthly Bulletin*, 10:1–8, 1936.
108. H. Tulkens. On FDH efficiency analysis: Some methodological issues and applications to retail banking, courts and urban transit. *Journal of Productivity Analysis*, 4:183–210, 1993.
109. H.R. Varian. The nonparametric approach to production analysis. *Econometrica*, 52:579–597, 1984.
110. H.R. Varian. *Microeconomic Analysis, 3rd Edition.* W.W. Norton and Co., New York, NY, 1992.
111. D.A. Vlatsa. *Data Envelopment Analysis with Intensity Restriction.* PhD thesis, Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, 1995.
112. S. Voss and D.L. Woodruff. *Introduction to Computational Optimization Models for Production Planning in a Supply Chain, Second Edition.* Springer-Verlag, Berlin, 2006.
113. L.A. Wolsey and G.L. Nemhauser. *Integer and Combinatorial Optimization.* Wiley-Interscience, New York, NY, 1999.

Index