Introduction to Mathematical Optimization

R. Clark Robinson

Department of Mathematics
Northwestern University
Evanston Illinois 60208 USA

Contents

Preface

This book has been used in an upper division undergraduate course about optimization given in the Mathematics Department at Northwestern University. Only deterministic problems with a continuous choice of options are considered, hence optimization of functions whose variables are (possibly) restricted to a subset of the real numbers or some Euclidean space. We treat the case of both linear and nonlinear functions. Optimization of linear functions with linear constraints is the topic of Chapter 1, linear programming. The optimization of nonlinear functions begins in Chapter 2 with a more complete treatment of maximization of unconstrained functions that is covered in calculus. Chapter 3 considers optimization with constraints. First, we treat equality constraints that includes the Implicit Function Theorem and the method of Lagrange multipliers. Then we treat inequality constraints, which is the covers Karush-Kuhn-Tucker Theory. Included is a consideration of convex and concave functions. Finally, Chapter 4 considers maximization over multiple time periods, or dynamic programming. Although only considering discrete time, we treat both finite and infinite number of time periods. In previous years, I have used the textbooks by Sundaram [**14**] and Walker [**16**]. Our presentation of linear programming is heavily influenced by [**16**] and the material on nonlinear functions by [**14**].

The prerequisites include courses on linear algebra and differentiation of functions of several variables. Knowledge of linear algebra is especially important. The simplex method to solve linear programs involves a special type of row reduction of matrices. We also use the concepts of the rank of a matrix and linear independence of a collection of vectors. In terms of differentiation, we assume that the reader knows about partial derivatives and the gradient of a real-valued function. With this background, we introduce the derivative of a function between Euclidean spaces as a matrix. This approach is used to treat implicit differentiation with several constraints and independence of constraints. In addition, we assume an exposure into formal presentation of mathematical definitions and theorems that our student get through a course on foundation to higher mathematics or a calculus course that introduces formal mathematical notation as our freshman MENU and MMSS courses do at Northwestern. Appendix A contains a brief summary of some of the mathematical language that is assumed from such a course. We do not assume the reader has had a course in real analysis. We do introduce a few concepts and terminology that is covered in depth in such a course, but merely require the reader to gain a working knowledge of this material so we can state results succinctly and accurately.

I would appreciate being informed by email of any errors or confusing statements, whether typographical, numerical, grammatical, or miscellaneous.

R. Clark Robinson
Northwestern University
Evanston Illinois
clark@math.northwestern.edu
May 2013

Chapter 1

Linear Programming

In this chapter, we begin our consideration of optimization by considering linear programming, maximization or minimization of linear functions over a region determined by linear inequalities. The word 'programming' does not refer to 'computer programming' but originally referred to the 'preparation of a schedule of activities'. Also, these problems can be solved by an algorithm that can be implemented on a computer, so linear programming is a widely used practical tool. We will only present the ideas and work low dimensional examples to gain an understanding of what the computer is doing when it solves the problem. However, the method scales well to examples with many variables and many constraints.

1.1. Basic Problem

For a real valued function f defined on a set $\mathscr{F}$, written $f : \mathscr{F} \subset \mathbb{R}^n \to \mathbb{R}$, $f(\mathscr{F}) = \{f(\mathbf{x}) : \mathbf{x} \in \mathscr{F}\}$ is the set of *attainable values* of f on $\mathscr{F}$, or the *image* of $\mathscr{F}$ by f.

We say that the real valued function f has a *maximum* on $\mathscr{F}$ at a point $\mathbf{x}_M \in \mathscr{F}$ provided that $f(\mathbf{x}_M) \geq f(\mathbf{x})$ for all $\mathbf{x} \in \mathscr{F}$. When $\mathbf{x}_M$ exists, the maximum value $f(\mathbf{x}_M)$ is written as

$$\max\{f(\mathbf{x}) : \mathbf{x} \in \mathscr{F}\}$$

and is called the *maximum or maximal value of f on $\mathscr{F}$*. We also say that $\mathbf{x}_M$ is a *maximizer* of f on $\mathscr{F}$, or that $f(\mathbf{x}_M)$ maximizes $f(\mathbf{x})$ subject to $\mathbf{x} \in \mathscr{F}$.

In the same way, we say that f has a *minimum* on $\mathscr{F}$ at a point $\mathbf{x}_m \in \mathscr{F}$ provided that $f(\mathbf{x}_m) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathscr{F}$. The value $f(\mathbf{x}_m)$ is written as $\min\{f(\mathbf{x}) : \mathbf{x} \in \mathscr{F}\}$ and is called the *minimum or minimal value of f on $\mathscr{F}$*. We also say that $\mathbf{x}_m$ is a *minimizer* of f on $\mathscr{F}$, or that $f(\mathbf{x}_M)$ minimizes $f(\mathbf{x})$ subject to $\mathbf{x} \in \mathscr{F}$.

The function f has an *extremum* at $\mathbf{x}_0 \in \mathscr{F}$ provided that $\mathbf{x}_0$ is either a maximizer or a minimizer. The point $\mathbf{x}_0$ is also called an *optimal solution*.

Optimization Questions.

- Does $f(\mathbf{x})$ attain a maximum (or minimum) on $\mathscr{F}$ for some $\mathbf{x} \in \mathscr{F}$?
- If so, what are the points at which $f(\mathbf{x})$ attains a maximum (or minimum) subject to $\mathbf{x} \in \mathscr{F}$, and what is the maximal value (or minimal value)?

There are problems that have no maximum nor minimum. For example, there is no maximizer or minimizer of $f(x) = 1/x$ on $(0, \infty)$.

Notations

We denote the *transpose* of a matrix (or a vector) $\mathbf{A}$ by $\mathbf{A}^\mathsf{T}$.

For two vectors $\mathbf{v}$ and $\mathbf{w}$ in $\mathbb{R}^n$, $\mathbf{v} \geq \mathbf{w}$ means that $v_i \geq w_i$ for all $1 \leq i \leq n$. In the same way, $\mathbf{v} \gg \mathbf{w}$ means that $v_i > w_i$ for all $1 \leq i \leq n$. We also consider two positive "quadrants" of $\mathbb{R}^n$,

$$\mathbb{R}^n_+ = \{\, \mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq \mathbf{0} \,\} = \{\, \mathbf{x} \in \mathbb{R}^n : x_i \geq 0 \text{ for } 1 \leq i \leq n \,\} \quad \text{and}$$
$$\mathbb{R}^n_{++} = \{\, \mathbf{x} \in \mathbb{R}^n : \mathbf{x} \gg \mathbf{0} \,\} = \{\, \mathbf{x} \in \mathbb{R}^n : x_i > 0 \text{ for } 1 \leq i \leq n \,\}.$$

Definition. The rank of an $m \times n$ matrix $\mathbf{A}$ is used at various times in this course. The *rank* of $\mathbf{A}$, $\operatorname{rank}(\mathbf{A})$, is the dimension of the column space of $\mathbf{A}$, i.e., the largest number of linearly independent columns of $\mathbf{A}$. This integer is the same as number of pivots in the row reduced echelon form of $\mathbf{A}$. (See Lay [**9**].)

It is possible to determine the rank by determinants of submatrices: $\operatorname{rank}(\mathbf{A}) = k$ iff $k \geq 0$ is the largest integer with $\det(\mathbf{A}_k) \neq 0$, where $\mathbf{A}_k$ is any $k \times k$ submatrix of $\mathbf{A}$ formed by selecting k columns and k rows. To see this is equivalent, let $\mathbf{A}'$ be submatrix of $\mathbf{A}$ with k columns that are linearly independent and span column space, so $\operatorname{rank}(\mathbf{A}') = k$. The dimension of the row space of $\mathbf{A}'$ is k, so there are k rows of $\mathbf{A}'$ forming the $k \times k$ submatrix $\mathbf{A}_k$ with $\operatorname{rank}(\mathbf{A}_k) = k$, so $\det(\mathbf{A}_k) \neq 0$. The submatrix $\mathbf{A}_k$ is also a submatrix of pivot columns and rows. Note in linear algebra, the pivot columns are usually chosen with the matrix in echelon form with zeroes in the row to the left of a pivot positions. We often choose other columns as will be since in the examples given in the rest of the chapter.

1.2. Graphical Solution

A linear programming problem with a few number of variables can be solved graphically by finding the vertices of the allowed values of the variables. We illustrate this solution method with an example.

Example 1.1 (Production). A company can make two products with x_1 and x_2 being the amount of each and with profit $f(x_1, x_2) = 8x_1 + 6x_2$. Because of limits to the amounts of three inputs for production there are three constraints on production. The first input I_1 restricts production by $x_1 + x_2 \leq 2$, the second input I_2 restricts $5x_1 + 10x_2 \leq 16$, and the third input I_3 restricts $2x_1 + x_2 \leq 3$. Therefore the problem is

$$\begin{array}{lll} \text{Maximize:} & 8x_1 + 6x_2 & \text{(profit)}, \\ \text{Subject to:} & x_1 + x_2 \leq 2, & (I_1) \\ & 5x_1 + 10x_2 \leq 16, & (I_2) \\ & 2x_1 + x_2 \leq 3, & (I_3) \\ & x_1 \geq 0,\ x_2 \geq 0. & \end{array}$$

The function $f(x_1, x_2)$ and all the constraint functions are linear so this is a linear programming problem. The *feasible set* $\mathscr{F}$ is the set of all the points satisfying the constraint equations and is the shaded region in Figure 1.2.1.

The vertices of the feasible region are $(0, 0)$, $(1.5, 0)$, $(1, 1)$, $(0.8, 1.2)$, and $(0, 1.6)$. Other points where two constraints are equal include the points $(\frac{14}{15}, \frac{17}{15})$, $(2, 0)$, $(3.2, 0)$, $(0, 2)$, and $(0, 3)$, each of which lies outside the feasible set (some other constraint function is negative).

Since the gradient $\nabla f(x_1, x_2) = (8, 6)^\mathsf{T} \neq (0, 0)^\mathsf{T}$, the maximum must be on the boundary of $\mathscr{F}$. Since the value of $f(x_1, x_2)$ along an edge is a linear combination of the values at the end points, if a point in the middle of one of the edges were a maximizer, then f would have the same value at the two end points of the edge. Therefore, a point that maximizes f can be found at one of the vertices of $\mathscr{F}$.

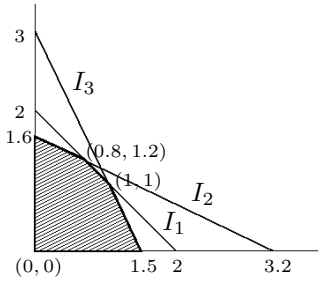


Figure 1.2.1. Feasible set for production problem

The values of the objective function at the vertices are $f(0,0) = 0$, $f(1.5,0) = 12$, $f(1,1) = 14$, $f(0.8,1.2) = 13.6$, and $f(0,1.6) = 9.6$. Therefore, the maximizer on $\mathscr{F}$ is $(1,1)$ and the maximal value is 14. ■

Definition. A *resource constraint* is one of the form $a_{i1}x_1 + \cdots + a_{in}x_n \le b_i$ with $b_i \ge 0$. The name refers to the upper limit on the combination of the variables.

A *standard maximization linear programming problem* has only resource constraints:

$$\begin{aligned} \text{Maximize:}\quad & f(\mathbf{x}) = \mathbf{c}\cdot\mathbf{x} = c_1x_1 + \cdots + c_nx_n, \\ \text{Subject to:}\quad & a_{11}x_1 + \cdots + a_{1n}x_n \le b_1, \\ & \quad\vdots \qquad\qquad \vdots\ \vdots \\ & a_{m1}x_1 + \cdots + a_{mn}x_n \le b_m, \\ & x_j \ge 0 \quad \text{for } 1 \le j \le n. \end{aligned}$$

The data specifying the problem includes an $m \times n$ matrix $\mathbf{A} = (a_{ij})$, $\mathbf{c} = (c_1, \ldots, c_n)^{\mathsf{T}} \in \mathbb{R}^n$, and $\mathbf{b} = (b_1, \ldots, b_m)^{\mathsf{T}} \in \mathbb{R}^m_+$ with all the $b_i \ge 0$.

Using matrix notation, the constraints can be written as $\mathbf{Ax} \le \mathbf{b}$ and $\mathbf{x} \ge 0$. The *feasible set* is the set $\mathscr{F} = \{\, \mathbf{x} \in \mathbb{R}^n_+ : \mathbf{Ax} \le \mathbf{b} \,\}$.

Since the objective is to maximize $f(\mathbf{x})$, the function $f(\mathbf{x})$ is called the *objective function*.

A nonstandard linear program allows other types of inequalities for the constraints. See the exercises of this section and Section 1.3.3 for examples of these other types of constraints.

Example 1.2 (Minimization).

$$\begin{aligned} \text{Minimize:}\quad & 3x_1 + 2x_2, \\ \text{Subject to:}\quad & 2x_1 + x_2 \ge 4, \\ & x_1 + x_2 \ge 3, \\ & x_1 + 2x_2 \ge 4, \\ & x_1 \ge 0, \text{ and } x_2 \ge 0. \end{aligned}$$

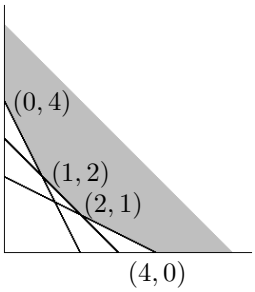


Figure 1.2.2. Feasible set for the minimization example

The feasible set that is shaded in Figure 1.2.2 is unbounded but $f(\mathbf{x}) \ge 0$ is bounded below, so $f(\mathbf{x})$ has a minimum. The vertices are $(4,0)$, $(2,1)$, $(1,2)$, and $(0,4)$, with values

$f(4,0) = 12,\ f(2,1) = 8,\ f(1,2) = 7,$ and $f(0,4) = 12$. Therefore, the minimum value is 7 which is attained at $(1,2)$. ■

Graphical Solution Method for a Linear Programming Problem
1. Determine or draw the feasible set $\mathscr{F}$. If $\mathscr{F} = \emptyset$, then the problem has no optimal solution, and it is said to be *infeasible*.
2. The problem is called *unbounded* and has no optimal solution provided that $\mathscr{F} \neq \emptyset$ and the objective function on $\mathscr{F}$ has **a.** arbitrarily large positive values for a maximization problem, or **b.** arbitrarily large negative values for a minimization problem.
3. A problem is called *bounded* provided that it is neither infeasible nor unbounded. In this case, an optimal solution exists. Determine all the vertices of $\mathscr{F}$ and values of the objective function at the vertices. Choose the vertex of $\mathscr{F}$ producing the maximal or minimal value of the objective function.

1.2. Exercises

1.2.1. Consider the following maximization linear programming problem.

$$\begin{aligned} \text{Maximize} \quad & 4x + 5y \\ \text{Subject to:} \quad & x + 2y \leq 12 \\ & x + y \leq 7 \\ & 3x + 2y \leq 18 \\ & 0 \leq x,\ 0 \leq y. \end{aligned}$$

a. Sketch the feasible set.
b. Why must a maximum exist?
c. Find the maximal feasible solution and the maximal value using the geometric method.

1.2.2. Consider the following linear programming problem:

$$\begin{aligned} \text{Minimize:} \quad & 3x + 2y \\ \text{Subject to:} \quad & 2x + y \geq 4 \\ & x + y \geq 3 \\ & x \geq 0, y \geq 0. \end{aligned}$$

1. Sketch the feasible set.
2. Why must a minimum exist?
3. Find the optimal feasible solution and the optimal value using the geometric method.

1.2.3. Consider the following minimization linear programming problem.

$$\begin{aligned} \text{Minimize} \quad & 3x + 2y \\ \text{Subject to:} \quad & 2x + y \geq 4 \\ & x + y \geq 3 \\ & 3x + 2y \leq 18 \\ & 0 \leq x, 0 \leq y. \end{aligned}$$

a. Sketch the feasible set.
b. Why must a minimum exist?
c. Find the minimal feasible solution and the minimal value using the geometric method.

1.2.4. Show graphically that the following linear program does not have a unique solution:

$$\begin{aligned} \text{Maximize} \quad & 30x + 40y \\ \text{Subject to:} \quad & 3x + 4y \leq 48 \\ & x + y \leq 14 \\ & 0 \leq x, 0 \leq y. \end{aligned}$$

1.2.5. Consider the following maximization linear programming problem.

$$\begin{aligned} \text{Maximize} \quad & 4x + 5y \\ \text{Subject to:} \quad & -x + 2y \geq 0 \\ & 2x - y \geq 3 \\ & 0 \leq x,\ 0 \leq y. \end{aligned}$$

a. Sketch the feasible set.
b. Explain why the problem is unbounded and a maximizer does not exists.

1.3. Simplex Method

The graphical solution method is not a practical algorithm for most problems because the number of vertices for a linear program grows very fast as the number of variables and constraints increase. Dantzig developed a practical algorithm based on row reduction from linear algebra. The first step is to add more variables into the standard maximization linear programming problem to make all the inequalities of the form $x_i \geq 0$ for some variables x_i.

1.3.1. Slack Variables

Definition. For a resource constraint, a *slack variable* can be set equal to the amount of unused resource: The inequality $a_{i1}x_1 + \cdots + a_{in}x_n \leq b_i$ with slack variable s_i becomes

$$a_{i1}x_1 + \cdots + a_{in}x_n + s_i = b_i \qquad \text{with } s_i \geq 0.$$

Example 1.3. The production linear program given earlier has only resource constraints:

$$\begin{aligned} \text{Maximize:} \quad & 8x_1 + 6x_2, \\ \text{Subject to:} \quad & x_1 + x_2 \leq 2, \\ & 5x_1 + 10x_2 \leq 16, \\ & 2x_1 + x_2 \leq 3, \\ & x_1 \geq 0, \quad x_2 \geq 0. \end{aligned}$$

When the slack variables are included the equations become

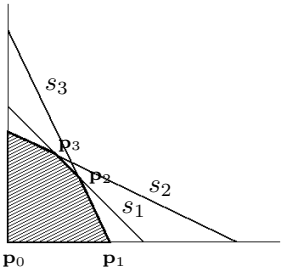
$$\begin{aligned} x_1 + x_2 + s_1 &= 2 \\ 5x_1 + 10x_2 + s_2 &= 16 \\ 2x_1 + x_2 + s_3 &= 3. \end{aligned}$$

Therefore, the problem is

$$\begin{aligned} \text{Maximize:} \quad & (8, 6, 0, 0, 0) \cdot (x_1, x_2, s_1, s_2, s_3) \\ \text{Subject to:} \quad & \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 5 & 10 & 0 & 1 & 0 \\ 2 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ s_1 \\ s_2 \\ s_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 16 \\ 3 \end{bmatrix} \qquad \text{and} \\ & x_1 \geq 0,\ x_2 \geq 0,\ s_1 \geq 0, \quad s_2 \geq 0, \quad s_3 \geq 0. \end{aligned}$$

The last three columns of the coefficient matrix for the slack variables can be considered as pivot columns and are linearly independent. Thus, the rank of the matrix is 3 and a solution to the nonhomogeneous system has two free variables. The starting solution $\mathbf{p}_0 = (0, 0, 2, 16, 3)$ in the feasible set is formed by setting the free variables x_1 and x_2 equal to zero and solving for the values of the three slack variables. The value is $f(\mathbf{p}_0) = 0$.

We proceed to change pivots to make a different pair the free variables equal to zero and a different triple of positive variables, in a process called "pivoting".



If we leave the vertex $(x_1, x_2) = (0, 0)$, or $\mathbf{p}_0 = (0, 0, 2, 16, 3)$, making $x_1 > 0$ the entering variable while keeping $x_2 = 0$, the first slack variable to become zero is s_3 when $x_1 = 1.5$. We have arrived at the vertex $(x_1, x_2) = (1.5, 0)$ of the feasible set by moving along one edge. Therefore we have a new solution with two zero variables and three nonzero variables, $\mathbf{p}_1 = (1.5, 0, 0.5, 8.5, 0)$ with $f(\mathbf{p}_1) = 8(1.5) = 12 > 0 = f(\mathbf{p}_0)$. The point $\mathbf{p}_1$ is a better feasible solution than $\mathbf{p}_0$. At this vertex, all of input one is used, $s_3 = 0$.

We repeat leaving $\mathbf{p}_1$ by making $x_2 > 0$ the entering variable while keeping $s_3 = 0$; the first other variable to become zero is s_1. We have arrived at $\mathbf{p}_2 = (1, 1, 0, 1, 0)$ with $f(\mathbf{p}_2) = 80(50) + 60(50) = 14 > 12 = f(\mathbf{p}_1)$. The point $\mathbf{p}_2$ is a better feasible solution than $\mathbf{p}_1$. In the $\mathbf{x}$-plane, we leave $(x_1, x_2) = (1.5, 0)$ and to arrive at $(x_1, x_2) = (1, 1)$. At this point, all of inputs one and three are used, $s_1 = 0$ and $s_3 = 0$, and $s_2 = 1$ units of input two remains.

If we were to leave $\mathbf{p}_2$ by making $s_3 > 0$ the entering variable while keeping $s_1 = 0$, the first variable to become zero is s_2, and we arrive at the point $\mathbf{p}_3 = (0.8, 1.2, 0, 0, 0.2)$. The objective function $f(\mathbf{p}_3) = 8(0.8) + 6(1.2) = 13.6 < 14 = f(\mathbf{p}_2)$, and $\mathbf{p}_3$ is a worse feasible solution than $\mathbf{p}_2$.

Let $\mathbf{z} \in \mathscr{F} \setminus \{\mathbf{p_2}\}$, $\mathbf{v} = \mathbf{z} - \mathbf{p}_2$, and $\mathbf{v}_j = \mathbf{p}_j - \mathbf{p}_2$ for $j = 1, 3$. Then

$$f(\mathbf{v}_j) = f(\mathbf{p}_j) - f(\mathbf{p}_2) < 0 \qquad \text{for } j = 1,\ 3.$$

The vectors $\{\, \mathbf{v}_1,\ \mathbf{v}_3 \,\}$ form a basis of $\mathbb{R}^2$, so $\mathbf{v} = y_1\, \mathbf{v}_1 + y_3\, \mathbf{v}_3$ for some y_j. The vector $\mathbf{v}$ points into the feasible set so $y_1, y_3 \geq 0$ and either $y_1 > 0$ or $y_2 > 0$. Therefore,

$$f(\mathbf{z}) = f(\mathbf{p}_2) + f(\mathbf{v}) = f(\mathbf{p}_2) + y_1\, f(\mathbf{v}_1) + y_3\, f(\mathbf{v}_3) < f(\mathbf{p}_2).$$

Since f cannot increase by moving along either edge going out from $\mathbf{p}_2$, $f(\mathbf{p}_2)$ is the maximum among the values at all the points of the feasible set and $\mathbf{p}_2$ is an optimal feasible solution.

In this example, the points $\mathbf{p}_0 = (0, 0, 2, 16, 3)$, $\mathbf{p}_1 = (1.5, 0, 0.5, 8.5, 0)$, $\mathbf{p}_2 = (1, 1, 0, 1, 0)$, $\mathbf{p}_3 = (0.8, 1.2, 0, 0, 10)$, and $\mathbf{p}_4 = (0, 1.6, 0.4, 0, 1.4)$ are called basic solutions since at most 3 variables are positive, where 3 is the rank of the coefficient matrix for the constraint equations. ■

Definition. Consider a standard maximization linear programming problem with only resource constraints: Maximize $f(\mathbf{x}) = \mathbf{c} \cdot \mathbf{x}$ subject to $\mathbf{A}\mathbf{x} \leq \mathbf{b}$ and $\mathbf{x} \geq 0$ for $\mathbf{c} = (c_1, \ldots, c_n)^\mathsf{T}$ in $\mathbb{R}^n$, $\mathbf{A} = (a_{ij})$ an $m \times n$ matrix, and $\mathbf{b} = (b_1, \ldots, b_m)^\mathsf{T}$ with $b_i \geq 0$ for $1 \leq i \leq m$. The corresponding *standard maximization linear programming problem with slack variables* $s_1, \ldots, s_m$ *added*, called *slack-variable form*, is the following:

$$\begin{array}{lll}
\text{Maximize:} & f(\mathbf{x}) = \mathbf{c}\cdot\mathbf{x} & \\
\text{Subject to:} & a_{11}x_1 + \cdots + a_{1n}x_n + s_1 & = b_1, \\
& \quad\vdots \qquad\qquad \vdots & \quad\vdots \\
& a_{m1}x_1 + \cdots + a_{mn}x_n + s_m & = b_m, \\
& x_j \geq 0 \text{ for } 1 \leq j \leq n, & \text{and} \\
& s_i \geq 0 \text{ for } 1 \leq i \leq m. &
\end{array}$$

Using matrix notation with $\mathbf{I}$ the $m \times m$ identity matrix, the problem is to maximize $f(\mathbf{x}) = \mathbf{c}\cdot\mathbf{x}$ subject to

$$[\,\mathbf{A}, \mathbf{I}\,]\begin{bmatrix}\mathbf{x}\\ \mathbf{s}\end{bmatrix} = \mathbf{b} \text{ with } \mathbf{x} \geq 0 \text{ and } \mathbf{s} \geq \mathbf{0}.$$

When solving this matrix equation by row reduction, we indicate the partition of the augmented matrix into the columns of (i) the original variables, (ii) the slack variables, and (iii) the constant vector $\mathbf{b}$ by inserting extra vertical lines between $\mathbf{A}$ and $\mathbf{I}$ and between $\mathbf{I}$ and $\mathbf{b}$, $[\,\mathbf{A}\,|\,\mathbf{I}\,|\,\mathbf{b}\,]$.

1.3.2. Simplex Method for Resource Requirements

Definition. Assume that $\bar{\mathbf{A}}\mathbf{x} = \mathbf{b}$ with $\mathbf{x} \geq \mathbf{0}$ is in slack-variable form, where $\bar{\mathbf{A}}$ is an $m \times (n+m)$ matrix with rank m and $\mathbf{b} \geq \mathbf{0}$. A general solution from a row reduced echelon form of the augmented matrix $[\bar{\mathbf{A}}|\mathbf{b}]$ writes the m *dependent variables* corresponding to the pivot columns in terms of n *free variables* corresponding to the non-pivot columns. The m dependent variables of a solution are called the *basic variables*. The n free variables of a solution are called the *non-basic variables*. In applications to linear programming, we usually take the solution formed by setting the free variables equal to zero, so there is at most m nonzero components corresponding to some or all of the basic variables. Since the basic variables correspond to pivot columns of some row reduced matrix, the columns of the basic variables are linear independent.

For a solution $\bar{\mathbf{A}}\mathbf{p} = \mathbf{b}$, if the set of columns of $\bar{\mathbf{A}}$ corresponding to $p_i \neq 0$, $\{\mathbf{C}_i : p_i \neq 0\}$, is linearly independent, then $\mathbf{p}$ is called a *basic solution*. If $\mathbf{p}$ is also feasible with $\mathbf{p} \geq 0$, then it is called a *basic feasible solution*. Since the $\text{rank}(\bar{\mathbf{A}}) = m$, there can be at most m nonzero components in a basic solution. A basic solution $\mathbf{p}$ is called *non-degenerate* provided that there are exactly m nonzero components. A basic solution $\mathbf{p}$ is called *degenerate* provided that there are fewer than m nonzero components. We find basic solutions by setting the n free variables equal zero and allow some basic variables to also be zero. If all the basic variables are nonzero then the solution is non-degenerate, otherwise it is degenerate.

Example 1.4 (Linear Algebra Solution of the Production Linear Program). The augmented matrix for the original system is

$$\left[\begin{array}{cc|ccc|c} x_1 & x_2 & s_1 & s_2 & s_3 & \\ \hline 1 & 1 & 1 & 0 & 0 & 2 \\ 5 & 10 & 0 & 1 & 0 & 16 \\ 2 & 1 & 0 & 0 & 1 & 3 \end{array}\right]$$

with free variables x_1 and x_2 and basic variables s_1, s_2, and s_3.

If we make $x_1 > 0$ while keeping $x_2 = 0$, x_1 becomes a new basic variable, called the *entering variable*.

(i) s_1 will become zero when $x_1 = \frac{2}{1} = 2$,
(ii) s_2 will become zero when $x_1 = \frac{16}{5} = 3.2$, and
(iii) s_3 will become zero when $x_1 = \frac{3}{2} = 1.5$.

Since s_3 becomes zero for the smallest value of x_1, s_3 is the *departing variable*. Since the third row determined the value of s_3, the entry in the first column third row is the new pivot.

Row reducing to make a pivot in the first column third row, we get

$$\left[\begin{array}{cc|ccc|c} x_1 & x_2 & s_1 & s_2 & s_3 & \\ \hline 1 & 1 & 1 & 0 & 0 & 2 \\ 5 & 10 & 0 & 1 & 0 & 16 \\ 2 & 1 & 0 & 0 & 1 & 3 \end{array}\right] \sim \left[\begin{array}{cc|ccc|c} x_1 & x_2 & s_1 & s_2 & s_3 & \\ \hline 0 & .5 & 1 & 0 & -.5 & 0.5 \\ 0 & 7.5 & 0 & 1 & -2.5 & 8.5 \\ 1 & .5 & 0 & 0 & .5 & 1.5 \end{array}\right].$$

Setting the free variables x_2 and s_3 equal zero, gives $\mathbf{p}_1 = (75, 0, 25, 425, 0)^{\mathsf{T}}$ as the new basic solution. The entries in the furthest right (augmented) column give the values of the basic variables and are all positive.

In order to keep track of the value of the objective function (or variable), $f = 80x_1 + 60x_2$, or $-80x_1 - 60x_2 + f = 0$, we include in the the augmented matrix a row for this equation and a column for the variable f,

$$\left[\begin{array}{cc|ccc|c|c} x_1 & x_2 & s_1 & s_2 & s_3 & f & \\ \hline 1 & 1 & 1 & 0 & 0 & 0 & 2 \\ 5 & 10 & 0 & 1 & 0 & 0 & 16 \\ 2 & 1 & 0 & 0 & 1 & 0 & 3 \\ \hline -8 & -6 & 0 & 0 & 0 & 1 & 0 \end{array}\right]$$

The entry in the column for f is one in the last objective function row and zero elsewhere. The entries for the x_i in this objective function row often start out negative as is the case in this example.

Performing the row reduction on this this further augmented matrix, we get

$$\left[\begin{array}{cc|ccc|c|c} x_1 & x_2 & s_1 & s_2 & s_3 & f & \\ \hline 1 & 1 & 1 & 0 & 0 & 0 & 2 \\ 5 & 10 & 0 & 1 & 0 & 0 & 16 \\ 2 & 1 & 0 & 0 & 1 & 0 & 3 \\ \hline -8 & -6 & 0 & 0 & 0 & 1 & 0 \end{array}\right] \sim \left[\begin{array}{cc|ccc|c|c} x_1 & x_2 & s_1 & s_2 & s_3 & f & \\ \hline 0 & .5 & 1 & 0 & -.5 & 0 & 0.5 \\ 0 & 7.5 & 0 & 1 & -2.5 & 0 & 8.5 \\ 1 & .5 & 0 & 0 & .5 & 0 & 1.5 \\ \hline 0 & -2 & 0 & 0 & 4 & 1 & 12 \end{array}\right].$$

Since f is the pivot for the objective function row, the bottom right entry gives that the new value $f = 12$ for $x_2 = s_3 = 0$, $x_1 = 1.5 > 0$, $s_1 = 0.5 > 0$, and $s_2 = 8.5 > 0$.

If we pivot back to make $s_3 > 0$, the value of f becomes smaller, so we select x_2 as the next entering variable, keeping $s_3 = 0$.

(i) x_1 becomes zero when $x_2 = \frac{1.5}{0.5} = 3$,
(ii) s_1 becomes zero when $x_2 = \frac{0.5}{0.5} = 1$, and
(iii) s_2 becomes zero when $x_2 = \frac{8.5}{7.5} \approx 1.13$.

Since the smallest positive value of x_1 comes from s_1, s_1 is the departing variable and we make the entry in the first row second column the new pivot.

$$\left[\begin{array}{cc|ccc|c|c} x_1 & x_2 & s_1 & s_2 & s_3 & f & \\ \hline 0 & .5 & 1 & 0 & -.5 & 0 & 0.5 \\ 0 & 7.5 & 0 & 1 & -2.5 & 0 & 8.5 \\ 1 & .5 & 0 & 0 & .5 & 0 & 1.5 \\ \hline 0 & -2 & 0 & 0 & 4 & 1 & 12 \end{array}\right] \sim \left[\begin{array}{cc|ccc|c|c} x_1 & x_2 & s_1 & s_2 & s_3 & f & \\ \hline 0 & 1 & 2 & 0 & -1 & 0 & 1 \\ 0 & 0 & -15 & 1 & 5 & 0 & 1 \\ 1 & 0 & -1 & 0 & 1 & 0 & 1 \\ \hline 0 & 0 & 4 & 0 & 2 & 1 & 14 \end{array}\right].$$

The value of the objective function is now $f = 14$.

Why does the objective function decrease when moving along the third edge making $s_3 > 0$ an entering variable, keeping $s_1 = 0$?

(i) x_1 becomes zero when $s_3 = \frac{1}{1} = 1$,

(ii) x_2 becomes zero when $s_3 = \frac{1}{-1} = -1$, and

(iii) s_2 becomes zero when $s_3 = \frac{1}{5} = 0.2$.

The smallest positive value of s_3 comes from s_2, and we pivot on the second row fifth column.

$$\left[\begin{array}{cc|ccc|c|c} x_1 & x_2 & s_1 & s_2 & s_3 & f & \\ \hline 0 & 1 & 2 & 0 & -1 & 0 & 1 \\ 0 & 0 & -15 & 1 & 5 & 0 & 1 \\ 1 & 0 & -1 & 0 & 1 & 0 & 1 \\ \hline 0 & 0 & 4 & 0 & 2 & 1 & 14 \end{array}\right] \sim \left[\begin{array}{cc|ccc|c|c} x_1 & x_2 & s_1 & s_2 & s_3 & f & \\ \hline 0 & 1 & -1 & 0 & 0 & 0 & 1.2 \\ 0 & 0 & -3 & 0.2 & 1 & 0 & 0.2 \\ 1 & 0 & 2 & -0.2 & 0 & 0 & 0.8 \\ \hline 0 & 0 & 10 & -0.4 & 0 & 1 & 13.6 \end{array}\right].$$

The value of the objective function decreased to 13.6. Since the value in the fifth column of the row for the objective function is already positive before we pivot, the value decreases.

Therefore, the values $(x_1^*, x_2^*, s_1^*, s_2^*, s_3^*) = (1, 1, 0, 1, 0)$ give the maximal value for f with $f(1, 1, 0, 1, 0) = 14$. ■

In the pivoting process used in the example, the variable f remains the pivot variable for the objective function row. This explains why the value in the bottom right of the augment matrix is the value of the objective function at every step. However, since it does not play a role in the row reduction, we drop this column for the variable f from the augmented matrices.

Definition. The augmented matrix with a row augmented for the objective function but without a column for the objective function variable is called the *tableau*. The row in the tableau for the objective function is called the *objective function row*.

Steps in the Simplex Method with only Resource Constraints

1. Add a slack variable for each resource inequality and set up the tableau. An initial feasible solution is determined by setting all the original variables equal to zero and solving for the slack variables.

2. Choose as entering variable any non-basic variable with a negative entry in the objection function row. We usually use the most negative entry. (The entry must be negative in order for the result of the pivoting to be an increase in the objective function.)

3. From the column selected in the previous step, select as a new pivot the row for which the ratio of entry in the augmented column divided by the entry in the column selected is the smallest nonnegative value. If such a pivot position exists, then row reduce the matrix making this selected entry a new pivot position and all the other entries in this column zero. (The pivot position must have a positive entry with the smallest ratio so that this variable becomes zero for the smallest value of the entering variable.)

One pivoting step interchanges one free variable with one basic variable. The variable for the column with the new pivot position is the entering variable that changes from a free variable equal to zero to a basic variable which is nonnegative and usually positive. The departing variable is the old basic variable for the row of the new pivot position which becomes a free variable with value zero.

Steps in the Simplex Method with only Resource Constraints, continued	
3′.	If there is a column with a negative coefficient in the objective function row and only nonpositive entries in the column above, then the objective function has no upper bound and the problem is unbounded. We illustrate this case with an exercise.
4.	The solution is optimal when all entries in the objective function row are non-negative. This tableau is called the *optimal tableau*.
5.	For the optimal tableau, if there is a zero entry in the objective function row for a non-basic variable with a positive entry in the column, then a different set of basic variables is possible. If, in addition, all the basic variables are positive, then the optimal solution is not unique.

1.3.3. General Constraints

We proceed to consider non-resource constraints. We continue to take all the constants $b_i \geq 0$ by multiplying one of the inequalities by -1 if necessary.

Requirement Constraints. A *requirement constraint* is given by $a_{i1}x_1 + \cdots + a_{in}x_n \geq b_i$ with $b_i > 0$. Such inequalities occur especially for a minimization problem. A solution of a requirement constraint can have a surplus of quantity, so instead of adding on the unused resource with a slack variable, we need to subtract off the excess resource by means of a *surplus variable* $s_i \geq 0$, $a_{i1}x_1 + \cdots + a_{in}x_n - s_i = b_i$. In order to find an initial feasible solution, for each requirement constraint, we also add an *artificial variable* $r_i \geq 0$, resulting in the equation

$$a_{i1}x_1 + \cdots + a_{in}x_n - s_i + r_i = b_i,$$

An initial solution is formed by setting the artificial variable $r_i = b_i > 0$, while setting all the surplus variables $s_i = 0$ and all the $x_j = 0$.

Equality Constraints. For an *equality constraint* $a_{i1}x_1 + \cdots + a_{in}x_n = b_i$ with $b_i \geq 0$, we only add an artificial variable $r_i \geq 0$, resulting in the equation

$$a_{i1}x_1 + \cdots + a_{in}x_n + r_i = b_i.$$

An initial solution is $r_i = b_i$ while all the $x_j = 0$. Since initially, $r_i = b_i \geq 0$, r_i will remain non-negative throughout the row reduction. (Alternatively, we could replace the equality by two inequalities $a_{i1}x_1 + \cdots + a_{in}x_n \geq b_i$ and $a_{i1}x_1 + \cdots + a_{in}x_n \leq b_i$, but this involves more equations and variables.)

If either requirement constraints or equality constraints are present, the initial solution has all the $x_i = 0$ and all the surplus variables equal zero while the slack and artificial variables are greater than or equal to zero. This initial solution is not feasible if an artificial variables is positive for a requirement constraint or an equality constraint.

Example 1.5 (Minimization Example). Assume that two foods are consumed in amounts x_1 and x_2 with costs of 15 and 7 per unit and yield $(5, 3, 5)$ and $(2, 2, 1)$ units of three vitamins respectively.

The problem is to minimize the cost $15\,x_1 + 7\,x_2$ or maximize $-15\,x_1 - 7\,x_2$ with the constraints:

$$\begin{aligned} \text{Maximize:} \quad & -15\,x_1 - 7\,x_2 \\ \text{Subject to:} \quad & 5\,x_1 + 2\,x_2 \geq 60, \\ & 3\,x_1 + 2\,x_2 \geq 40, \\ & 5\,x_1 + 1\,x_2 \geq 35, \\ & x_1 \geq 0, \text{ and } x_2 \geq 0. \end{aligned}$$

The original tableau is

$$\left[\begin{array}{cc|ccc|ccc|c} x_1 & x_2 & s_1 & s_2 & s_3 & r_1 & r_2 & r_3 & \\ \hline 5 & 2 & -1 & 0 & 0 & 1 & 0 & 0 & 60 \\ 3 & 2 & 0 & -1 & 0 & 0 & 1 & 0 & 40 \\ 5 & 1 & 0 & 0 & -1 & 0 & 0 & 1 & 35 \\ \hline 15 & 7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

For the original problem, the solution involves the artificial variables and does not give a feasible solution for $\mathbf{x}$. To eliminate the artificial variables, preliminary steps are added to the simplex algorithm to force all the artificial variables to be zero. The artificial variables are forced to zero by means of an *artificial objective function* that is the negative sum of all the equations that contain artificial variables,

$$-13x_1 - 5x_2 + s_1 + s_2 + s_3 + (-r_1 - r_2 - r_3) = -135.$$

We think of $R = -r_1 - r_2 - r_3$ as a new variable. It is always less than or equal to zero and so has a maximum less than or equal to zero. If the artificial objective function can be made equal zero, then this gives an initial feasible basic solution using only the original and slack variables without using the artificial variables, and the artificial variables can be dropped and proceed as before.

The tableau with the artificial objective function included is

$$\left[\begin{array}{cc|ccc|ccc|c} x_1 & x_2 & s_1 & s_2 & s_3 & r_1 & r_2 & r_3 & \\ \hline 5 & 2 & -1 & 0 & 0 & 1 & 0 & 0 & 60 \\ 3 & 2 & 0 & -1 & 0 & 0 & 1 & 0 & 40 \\ 5 & 1 & 0 & 0 & -1 & 0 & 0 & 1 & 35 \\ \hline 15 & 7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -13 & -5 & 1 & 1 & 1 & 0 & 0 & 0 & -135 \end{array}\right].$$

Note that there are zeroes in the columns for the r_i in artificial function row. The most negative coefficient in the artificial objective function (outside the artificial variables) is -13. For this first column, the entry that has the smallest positive ratio of $\frac{b_i}{a_{i1}}$, is the third row. Pivoting first on a_{31}, then a_{22}, and finally on a_{14} yields a feasible initial basic solution without any artificial variables.

$$\sim \left[\begin{array}{cc|ccc|ccc|c} x_1 & x_2 & s_1 & s_2 & s_3 & r_1 & r_2 & r_3 & \\ \hline 0 & 1 & -1 & 0 & 1 & 1 & 0 & -1 & 25 \\ 0 & \frac{7}{5} & 0 & -1 & \frac{3}{5} & 0 & 1 & -\frac{3}{5} & 19 \\ 1 & \frac{1}{5} & 0 & 0 & -\frac{1}{5} & 0 & 0 & \frac{1}{5} & 7 \\ \hline 0 & 4 & 0 & 0 & 3 & 0 & 0 & -3 & -105 \\ 0 & -\frac{12}{5} & 1 & 1 & -\frac{8}{5} & 0 & 0 & \frac{13}{5} & -44 \end{array}\right]$$

$$\sim \left[\begin{array}{cc|ccc|ccc|c} x_1 & x_2 & s_1 & s_2 & s_3 & r_1 & r_2 & r_3 & \\ \hline 0 & 0 & -1 & \frac{5}{7} & \frac{4}{7} & 1 & -\frac{5}{7} & -\frac{4}{7} & \frac{80}{7} \\ 0 & 1 & 0 & -\frac{5}{7} & \frac{3}{7} & 0 & \frac{5}{7} & -\frac{3}{7} & \frac{95}{7} \\ 1 & 0 & 0 & \frac{1}{7} & -\frac{2}{7} & 0 & -\frac{1}{7} & \frac{2}{7} & \frac{30}{7} \\ \hline 0 & 0 & 0 & \frac{20}{7} & \frac{9}{7} & 0 & -\frac{20}{7} & -\frac{9}{7} & -\frac{1115}{7} \\ 0 & 0 & 1 & -\frac{5}{7} & -\frac{4}{7} & 0 & \frac{12}{7} & \frac{11}{7} & -\frac{80}{7} \end{array}\right]$$

$$\sim \left[\begin{array}{cc|ccc|ccc|c} x_1 & x_2 & s_1 & s_2 & s_3 & r_1 & r_2 & r_3 & \\ \hline 0 & 0 & -\frac{7}{5} & 1 & \frac{4}{5} & \frac{7}{5} & -1 & -\frac{4}{5} & 16 \\ 0 & 1 & -1 & 0 & 1 & 1 & 0 & -1 & 25 \\ 1 & 0 & \frac{1}{5} & 0 & -\frac{2}{5} & -\frac{1}{5} & 0 & \frac{2}{5} & 2 \\ \hline 0 & 0 & 4 & 0 & -1 & -4 & 0 & 1 & -205 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{array}\right]$$

After these three steps, the artificial variables are no longer pivots and are zero, so they can be dropped. The values $(x_1, x_2, s_1, s_2, s_3) = (2, 25, 0, 16, 0)$ form an initial feasible basic solution. Finally, pivoting on a_{15} yields the optimal solution:

$$\left[\begin{array}{cc|ccc|c} x_1 & x_2 & s_1 & s_2 & s_3 & \\ \hline 0 & 0 & -\frac{7}{5} & 1 & \frac{4}{5} & 16 \\ 0 & 1 & -1 & 0 & 1 & 25 \\ 1 & 0 & \frac{1}{5} & 0 & -\frac{2}{5} & 2 \\ \hline 0 & 0 & 4 & 0 & -1 & -205 \end{array}\right] \sim \left[\begin{array}{cc|ccc|c} x_1 & x_2 & s_1 & s_2 & s_3 & \\ \hline 0 & 0 & -\frac{7}{4} & \frac{5}{4} & 1 & 20 \\ 0 & 1 & \frac{3}{4} & -\frac{5}{4} & 0 & 5 \\ 1 & 0 & -\frac{1}{2} & \frac{1}{2} & 0 & 10 \\ \hline 0 & 0 & \frac{9}{4} & \frac{5}{4} & 0 & -185 \end{array}\right].$$

At this point, all the entries in the objective function are positive, so this is the optimal solution: $(x_1, x_2, s_1, s_2, s_3) = (10, 5, 0, 0, 20)$ with an value of -185 for f.

For the original problem, the minimal solution has a value of 185. ■

Steps in the Simplex Method with any Type of Constraints

1. Make all the constraints on the right hand side of any inequality or equation positive, $b_i \geq 0$, by multiplying by -1 if necessary.

2. Add a slack variable for each resource inequality, add a surplus variable and an artificial variable for each requirement constraint, and add an artificial variable for each equality constraint.

3. If either a requirement constraint or an equality constraint is present, then form the *artificial objective function* by taking the negative sum of all the equations that contain artificial variables, dropping the terms involving the artificial variables. Set up the tableau. (The row for the artificial objective function has zeroes in the columns of the artificial variables.) An initial solution of the equation including the artificial variables is determined by setting all the original variables $x_j = 0$, all the surplus variables $s_i = 0$, all the slack variables $s_i = b_i$, and all the artificial variables $a_i = b_i$.

4. Apply the simplex algorithm using the artificial objective function.

a. If it is not possible to make the artificial objective function equal to zero i.e., if there is a positive artificial variable in the optimal solution of the artificial objective function), then there is no feasible solution and stop.

b. If the value of the artificial objective function can be made equal to zero (when the artificial variables are not pivot columns), then all the artificial variables have been made equal to zero. (This is true even if some of the entries in the artificial objective function row are nonzero and possibly even negative.) At this point, drop the artificial variables and artificial objective function from the tableau and continue using the initial feasible basic solution constructed.

5. Apply the simplex algorithm to the actual objective function. The solution is optimal when all entries in the objective function row are nonnegative.

Example 1.6. Consider the problem of

$$\begin{aligned} \text{Maximize:} \quad & 3\,x_1 + 4\,x_2 \\ \text{Subject to:} \quad & -2\,x_1 + x_2 \leq 6, \\ & 2\,x_1 + 2\,x_2 \geq 24, \\ & x_1 = 8, \\ & x_1 \geq 0,\ x_2 \geq 0. \end{aligned}$$

With slack, surplus, and artificial variables added the problem becomes

$$\begin{aligned} \text{Maximize:} \quad & 3\,x_1 + 4\,x_2 \\ \text{Subject to:} \quad & -2\,x_1 + x_2 + s_1 = 6, \\ & 2\,x_1 + 2\,x_2 - s_2 + r_2 = 24, \\ & x_1 + r_3 = 8. \end{aligned}$$

The negative sum of the second and third equation, gives the artificial objective function $-3\,x_1 - 2\,x_2 + s_2 - r_2 - r_3 = -32$.

The tableau with variables is

$$\left[\begin{array}{cc|cc|cc|c} x_1 & x_2 & s_1 & s_2 & r_2 & r_3 & \\ \hline -2 & 1 & 1 & 0 & 0 & 0 & 6 \\ 2 & 2 & 0 & -1 & 1 & 0 & 24 \\ 1 & 0 & 0 & 0 & 0 & 1 & 8 \\ \hline -3 & -4 & 0 & 0 & 0 & 0 & 0 \\ -3 & -2 & 0 & 1 & 0 & 0 & -32 \end{array}\right]$$

Pivoting on a_{31} and then a_{22},

$$\sim \left[\begin{array}{cc|cc|cc|c} x_1 & x_2 & s_1 & s_2 & r_2 & r_3 & \\ \hline 0 & 1 & 1 & 0 & 0 & 2 & 22 \\ 0 & 2 & 0 & -1 & 1 & -2 & 8 \\ 1 & 0 & 0 & 0 & 0 & 1 & 8 \\ \hline 0 & -4 & 0 & 0 & 0 & 3 & 24 \\ 0 & -2 & 0 & 1 & 0 & 3 & -8 \end{array}\right] \sim \left[\begin{array}{cc|cc|cc|c} x_1 & x_2 & s_1 & s_2 & r_2 & r_3 & \\ \hline 0 & 0 & 1 & \frac{1}{2} & -\frac{1}{2} & 3 & 18 \\ 0 & 1 & 0 & -\frac{1}{2} & \frac{1}{2} & -1 & 4 \\ 1 & 0 & 0 & 0 & 0 & 1 & 8 \\ \hline 0 & 0 & 0 & -2 & 2 & -1 & 40 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{array}\right]$$

We have attained a feasible solution of $(x_1, x_2, s_1, s_2) = (8, 4, 18, 0)$. We can now drop the artificial objective function and artificial variables. Pivoting on $a_{1,4}$

$$\left[\begin{array}{cc|cc|c} x_1 & x_2 & s_1 & s_2 & \\ \hline 0 & 0 & 1 & \frac{1}{2} & 18 \\ 0 & 1 & 0 & -\frac{1}{2} & 4 \\ 1 & 0 & 0 & 0 & 8 \\ \hline 0 & 0 & 0 & -2 & 40 \end{array}\right] \sim \left[\begin{array}{cc|cc|c} x_1 & x_2 & s_1 & s_2 & \\ \hline 0 & 0 & 2 & 1 & 36 \\ 0 & 1 & 1 & 0 & 22 \\ 1 & 0 & 0 & 0 & 8 \\ \hline 0 & 0 & 4 & 0 & 112 \end{array}\right]$$

This gives the optimal solution of $f = 112$ for $(x_1, x_2, s_1, s_2) = (8, 22, 0, 36)$. ■

Example 1.7 (Unbounded problem). Consider the problem of

$$\begin{aligned} \text{Maximize:} \quad & 5\,x_1 + 3\,x_2 \\ \text{Subject to:} \quad & 2\,x_1 + x_2 \geq 4, \\ & x_1 + 2\,x_2 \geq 4, \\ & x_1 + x_2 \geq 3, \\ & x_1 \geq 0,\ x_2 \geq 0 \end{aligned}$$

The tableau with surplus and artificial variables added is

$$\left[\begin{array}{cc|ccc|ccc|c} x_1 & x_2 & s_1 & s_2 & s_3 & r_1 & r_2 & r_3 & \\ \hline 2 & 1 & -1 & 0 & 0 & 1 & 0 & 0 & 4 \\ 1 & 2 & 0 & -1 & 0 & 0 & 1 & 0 & 4 \\ 1 & 1 & 0 & 0 & -1 & 0 & 0 & 1 & 3 \\ \hline -5 & -3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -4 & -4 & 1 & 1 & 1 & 0 & 0 & 0 & -11 \end{array}\right].$$

Applying the algorithm to the artificial objective function, we get the following:

$$\sim \left[\begin{array}{cc|ccc|ccc|c} x_1 & x_2 & s_1 & s_2 & s_3 & r_1 & r_2 & r_3 & \\ \hline 1 & \frac{1}{2} & -\frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 2 \\ 0 & \frac{3}{2} & \frac{1}{2} & -1 & 0 & -\frac{1}{2} & 1 & 0 & 2 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & -1 & -\frac{1}{2} & 0 & 1 & 1 \\ \hline 0 & -\frac{1}{2} & -\frac{5}{2} & 0 & 0 & \frac{5}{2} & 0 & 0 & 10 \\ 0 & -2 & -1 & 1 & 1 & 2 & 0 & 0 & -3 \end{array}\right]$$

$$\sim \left[\begin{array}{cc|ccc|ccc|c} x_1 & x_2 & s_1 & s_2 & s_3 & r_1 & r_2 & r_3 & \\ \hline 1 & 1 & 0 & 0 & -1 & 0 & 0 & 1 & 3 \\ 0 & 1 & 0 & -1 & 1 & 0 & 1 & -1 & 1 \\ 0 & 1 & 1 & 0 & -2 & -1 & 0 & 2 & 2 \\ \hline 0 & 2 & 0 & 0 & -5 & 0 & 0 & 5 & 15 \\ 0 & -1 & 0 & 1 & -1 & 1 & 0 & 2 & -1 \end{array}\right]$$

$$\sim \left[\begin{array}{cc|ccc|ccc|c} x_1 & x_2 & s_1 & s_2 & s_3 & r_1 & r_2 & r_3 & \\ \hline 1 & 0 & 0 & 1 & -2 & 0 & -1 & 2 & 2 \\ 0 & 1 & 0 & -1 & 1 & 0 & 1 & -1 & 1 \\ 0 & 0 & 1 & 1 & -3 & -1 & -1 & 3 & 1 \\ \hline 0 & 0 & 0 & 2 & -7 & 0 & -2 & 7 & 13 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{array}\right]$$

$$\sim \left[\begin{array}{cc|ccc|c} x_1 & x_2 & s_1 & s_2 & s_3 & \\ \hline 1 & 2 & 0 & -1 & 0 & 4 \\ 0 & 1 & 0 & -1 & 1 & 1 \\ 0 & 3 & 1 & -2 & 0 & 4 \\ \hline 0 & 7 & 0 & -5 & 0 & 20 \end{array}\right].$$

In the second pivot, we used the third column rather than the second because it eliminates fractions. In the optimal tableau, the column for s_2 has a negative value in the objection function row and negatives above it. If we keep $0 = x_2$ but use s_2 as a free variable, we get the solution

$$x_1 - s_2 = 4, \quad x_2 = 0, \quad s_1 - 2\,s_2 = 4, \quad s_3 - s_2 = 1, \qquad \text{or}$$
$$x_1 = 4 + s_2, \quad x_2 = 0, \quad s_1 = 4 + 2\,s_2, \quad s_3 = 1 + s_2, \quad f = 5\,x_1 + 3\,x_2 = 20 + 5\,s_2.$$

Thus, as $s_2 \geq 0$ increases, we keep a feasible solution and the objective function is unbounded.

This example indicates why a maximization linear program with a column of negative entries in its tableau above a negative entry in the objective function row is unbounded. ■

1.3. Exercises

1.3.1. Solve the following problem (a) graphically and (b) by the simplex method:

$$\begin{aligned} \text{Maximize:} \quad & 3\,x_1 + 2\,x_2 \\ \text{Subject to:} \quad & x_1 + 2\,x_2 \leq 70 \\ & x_1 + x_2 \leq 40 \\ & 3\,x_1 + x_2 \leq 90 \\ & x_1 \geq 0,\ x_2 \geq 0. \end{aligned}$$

1.3.2. Solve the following problem by the simplex method:

$$\begin{aligned} \text{Maximize:} \quad & 2\,x_1 + 3\,x_2 + 5\,x_3 \\ \text{Subject to:} \quad & x_1 + 4\,x_2 - 2\,x_3 \leq 30 \\ & -x_1 + 2\,x_2 + 5\,x_3 \leq 9 \\ & x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0. \end{aligned}$$

Give the values of the optimal solution, including the x_i values, the values of the slack variables, and the optimal value of the objective function.

1.3.3. Solve the following problem by the simplex method:

$$\begin{aligned} \text{Maximize:} \quad & 2\,x_1 + 5\,x_2 + 3\,x_3 \\ \text{Subject to:} \quad & x_1 + 2\,x_2 \leq 28 \\ & 2\,x_1 + 4\,x_3 \leq 16 \\ & x_2 + x_3 \leq 12 \\ & x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0. \end{aligned}$$

1.3.4. Solve the following problem by the simplex method:

$$\begin{aligned} \text{Maximize:} \quad & 3\,x_1 + 4\,x_2 + 2\,x_3 \\ \text{Subject to:} \quad & 3\,x_1 + 2\,x_2 + 4\,x_3 \leq 45 \\ & x_1 + 2\,x_2 + 3\,x_3 \leq 21 \\ & 4\,x_1 + 2\,x_2 + 2\,x_3 \leq 36 \\ & x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0. \end{aligned}$$

1.3.5. Solve the linear program by the simplex algorithm with artificial variables:

$$\begin{aligned} \text{Maximize:} \quad & x_1 + 2x_2 + 3x_3 \\ \text{Subject to:} \quad & x_1 + 2x_2 + x_3 = 36 \\ & 2x_1 + x_2 + 4x_3 \geq 12 \\ & 0 \leq x_1,\ 0 \leq x_2,\ 0 \leq x_3. \end{aligned}$$

1.3.6. Solve the linear program by the simplex algorithm with artificial variables:

$$\begin{aligned} \text{Maximize:} \quad & 4x_1 + 5x_2 + 3x_3 \\ \text{Subject to:} \quad & x_1 + 2x_2 + 3x_3 \leq 20 \\ & x_1 + 2x_2 \geq 2 \\ & x_1 - x_2 + x_3 = 7 \\ & x_1,\ x_2,\ x_3 \geq 0. \end{aligned}$$

1.3.7. Solve the linear program by the simplex algorithm with artificial variables:

$$\begin{aligned} \text{Maximize:} \quad & 4x_1 + 5x_2 + 3x_3 \\ \text{Subject to:} \quad & x_1 + 2x_2 + 3x_3 \leq 20 \\ & x_1 + 2x_2 \geq 2 \\ & x_1 - x_2 + x_3 = 7 \\ & x_1,\, x_2,\, x_3 \geq 0. \end{aligned}$$

1.3.8. Solve the linear program by the simplex algorithm with artificial variables:

$$\begin{aligned} \text{Maximize:} \quad & -x_1 - 2\,x_2 + 2x_3 \\ \text{Subject to:} \quad & x_1 + 2\,x_3 \leq 12 \\ & 2\,x_1 + 3\,x_2 + x_3 \geq 4 \\ & x_1 + 2\,x_2 - x_3 \geq 6 \\ & x_1,\, x_2,\, x_3 \geq 0. \end{aligned}$$

1.3.9. Solve the linear program by the simplex algorithm with artificial variables:

$$\begin{aligned} \text{Maximize:} \quad & x_1 + x_2 + x_3 \\ \text{Subject to:} \quad & x_1 + x_2 \geq 3 \\ & x_1 + 2\,x_2 + x_3 \geq 4 \\ & 2\,x_1 + x_2 + x_3 \leq 2 \\ & x_1,\, x_2,\, x_3 \geq 0. \end{aligned}$$

1.3.10. Use artificial variables to determine whether there are any vectors satisfying

$$\begin{aligned} & x_1 + x_2 + \leq 40 \\ & 2\,x_1 + x_2 \geq 70 \\ & x_1 + 3\,x_2 \leq 90 \\ & x_1,\, x_2,\, x_3 \geq 0. \end{aligned}$$

1.3.11. Show that the following problem is unbounded.

$$\begin{aligned} \text{Maximize} \quad & 2x + y \\ \text{Subject to:} \quad & x - y \leq 3 \\ & -3x + y \leq 1 \\ & 0 \leq x, 0 \leq y. \end{aligned}$$

1.3.12. Consider the following maximization linear program:

$$\begin{aligned} \text{Maximize} \quad & 2\,x_1 - 5\,x_2 \\ \text{Subject to:} \quad & x_1 + 2\,x_2 \leq 2 \\ & x_1 - 4\,x_2 \geq 4 \\ & 0 \leq x_1,\ 0 \leq x_2. \end{aligned}$$

a. Use the simplex method with artificial variables to show that there are no feasible solutions.

b. Plot the constraints and argue why the feasible set is empty.

1.4. Duality

A linear program can be associated with a two-person zero-sum game. A Nash equilibrium in mixed strategies for this game gives an optimal solution to not only the original game but also to an associated dual linear programming problem. If the original problem is a maximization problem, then the dual problem is a minimization problem and vice versa. We do not discuss this zero-sum game, but introduce this dual minimization problem using an example.

Example 1.8. Consider the production maximization linear programming problem that was given previously:

$$\begin{array}{lllll} \text{MLP:} & \text{Maximize:} & z = 8x_1 + 6x_2 & & \text{(profit)} \\ & \text{subject to:} & x_1 + x_2 \leq 2, & & (I_1) \\ & & 5x_1 + 10x_2 \leq 16, & & (I_2) \\ & & 2x_1 + x_2 \leq 3, & & (I_3), \\ & & x_1, x_2 \geq 0 & & \end{array}$$

Assume that excess inputs can be sold and shortfalls can be purchased for prices of y_1, y_2, and y_3, each of which is nonnegative, $y_j \geq 0$ for $j = 1, 2, 3$. The prices y_j are called *shadow prices* or *imputed values* of the resources or inputs and are set by the market (the potential for competing firms). With either purchase or sale of the input resource allowed, the profit for the firm is

$$P = 8x_1 + 6x_2 + (2 - x_1 - x_2)y_1 + (16 - 5x_1 - 10x_2)y_2 + (3 - 2x_1 - x_2)y_3.$$

The potential for outside competitors controls the imputed prices of the resources. If additional profit could be made by purchasing a resource to raise production (the constraint is violated), then outside competitors would bid for the resource and force the imputed price to rise until it was no longer profitable to purchase that resource. On the other hand, if there were a surplus of a resource, then outside competitors would not be willing to buy it and the imputed price would fall to zero. Therefore, either $2 - x_1 - x_2 = 0$ or $y_1 = 0$, and there are break even imputed prices for inputs at the margin. Similar results hold for the other resources, so

$$\begin{aligned} 0 &= (2 - x_1 - x_2)y_1, \\ 0 &= (16 - 5x_1 - 10x_2)y_2, \qquad \text{and} \\ 0 &= (3 - 2x_1 - x_2)y_3. \end{aligned} \tag{1}$$

For such shadow prices of inputs, the firm's profit is $P = 8x_1 + 6x_2$ and the situation from the firm's perspective yields the original maximization problem.

For a feasible choice of (x_1, x_2), the coefficient of each y_j in P is nonnegative, $2 - x_1 - x_2 \geq 0$, $16 - 5x_1 - 10x_2 \geq 0$, $3 - 2x_1 - x_2 \geq 0$, and the competitive market is forcing to zero the shadow price of any resource with a surplus; therefore, the market is minimizing P as a function of (y_1, y_2, y_3).

Regrouping the profit function P from the markets perspective yields

$$P = (8 - y_1 - 5y_2 - 2y_3)x_1 + (6 - y_1 - 10y_2 - y_3)x_2 + 2y_1 + 16y_2 + 3y_3.$$

One unit of first product (x_1) cost $1y_1$ for the first input, $5y_2$ for the second input, and $2y_3$ for the third input, so the cost of producing a unit of this product by a competitor is $y_1 + 5y_2 + 2y_3$. The market (potential competitors) forces this cost to be greater than or equal to its value of 8, $y_1 + 5y_2 + 2y_3 \geq 8$, since if the cost were less than 8 then other competitors would enter the market. In other words, the net profit of selling a unit of first product is $8 - y_1 - 5y_2 - 2y_3 \leq 0$. Similarly, the potential for competition for the second product forces $y_1 + 10y_2 + y_3 \geq 6$. These

are the inequalities of the dual problem,

$$y_1 + 5y_2 + 2y_3 \geq 8 \qquad \text{and}$$
$$y_1 + 10y_2 + y_3 \geq 6.$$

Also, if the net profit is negative, then the firm would not produce that output, so

$$0 = (8 - y_1 - 5y_2 - 2y_3)x_1 \qquad \text{and} \tag{2}$$
$$0 = (6 - y_1 - 10y_2 - y_3)x_2.$$

Therefore, at the optimal production levels, the profit for the firm is equal to the imputed value of the inputs for the two products, $P = 2y_1 + 16y_2 + 3y_3$. Since the the competitive market is minimizing P (see above), the market is minimizing $P = 2y_1 + 16y_2 + 3y_3$ subject to the shadow prices satisfying the dual constraints. So, from the market's perspective, we get the dual minimization problem:

mLP Minimize: $w = 2y_1 + 16y_2 + 3y_3$
Subject to: $y_1 + 5y_2 + 2y_3 \geq 8$,
$y_1 + 10y_2 + y_3 \geq 6$, and
$y_1, y_2, y_3 \geq 0$.

Relationship of dual minimization and primal maximization linear problem
1. The coefficient matrices of the x_i and y_i are transposes of each other.
2. The coefficients in the objective function for the MLP become the constants of the inequalities for the mLP.
3. The constants of the inequalities of the MLP become coefficients in the objective function for the mLP.

For the production linear program.

$$\left[\begin{array}{cc|ccc|c} x_1 & x_2 & s_1 & s_2 & s_3 & \\ \hline 1 & 1 & 1 & 0 & 0 & 2 \\ 5 & 10 & 0 & 1 & 0 & 16 \\ 2 & 1 & 0 & 0 & 1 & 3 \\ \hline -8 & -6 & 0 & 0 & 0 & 0 \end{array}\right] \sim \left[\begin{array}{cc|ccc|c} x_1 & x_2 & s_1 & s_2 & s_3 & \\ \hline 0 & 1 & 2 & 0 & -1 & 1 \\ 0 & 0 & -15 & 1 & 5 & 1 \\ 1 & 0 & -1 & 0 & 1 & 1 \\ \hline 0 & 0 & 4 & 0 & 2 & 14 \end{array}\right]$$

The optimal solution the the maximization problem is $x_1 = 1$ and $x_2 = 1$ with a payoff of 14.

As we state in Theorem 1.16 given subsequently, dual minimization problem must have a solution of $y_1 = 4$, $y_2 = 0$, and $y_3 = 2$ with the same payoff, where 4, 0, and 2 are the entries in the bottom row of the optimal tableau in the columns associated with the slack variables. Each value of a y_i corresponds to the marginal value of each addition unit of the corresponding input.

The sets of equations (1) and (2) are called *complementary slackness*. Note that the optimal solutions for of the maximization and dual minimization problems satisfy complementary slackness. ■

Example 1.9. A manufacturing company produces x_1 amount of the regular type and x_2 amount of the super type with a profit of $P = 12x_1 + 15x_2$. The assembly time is 20 minutes per regular unit and 30 minutes per super unit with a total of 2400 minutes available, so $2\,x_1 + 3\,x_2 \leq 240$. The painting time is 15 minutes per regular unit and 40 minutes per super unit with a total of 3000 minutes available, so $15\,x_1 + 40\,x_2 \leq 3000$ or $3\,x_1 + 8\,x_2 \leq 600$. Finally, the inspection time is 12 minutes per unit of each type with 1200 minutes available. Thus, we have the following maximization problem.

$$\begin{array}{lll} \text{MLP} \quad \text{Maximize:} & P = 12x_1 + 15x_2 & \text{(profit)} \\ \text{Subject to:} & 2\,x_1 + 3\,x_2 \le 240 & \text{(assembly time per 10 minutes),} \\ & 3\,x_1 + 8\,x_2 \le 600 & \text{(painting time per 5 minutes),} \\ & x_1 + x_2 \le 100 & \text{(inspection time per 12 minutes),} \\ & x_1 \ge 0 \;\; x_2 \ge 0. & \end{array}$$

The dual problem is

$$\begin{array}{lll} \text{mLP} \quad \text{Minimize:} & P = 240\,y_1 + 600\,y_2 + 100\,y_3, & \\ \text{Subject to:} & 2\,y_1 + 3\,y_2 + y_3 \ge 12, & \\ & 3\,y_1 + 8\,y_2 + y_3 \ge 15, & \\ & y_1 \ge 0,\; y_2 \ge 0,\; y_3 \ge 0. & \end{array}$$

The tableau for the MLP is

$$\left[\begin{array}{ccc|cc|c} x_1 & x_2 & s_1 & s_2 & s_3 & \\ \hline 2 & 3 & 1 & 0 & 0 & 240 \\ 3 & 8 & 0 & 1 & 0 & 600 \\ 1 & 1 & 0 & 0 & 1 & 100 \\ \hline -12 & -15 & 0 & 0 & 0 & 0 \end{array}\right] \sim \left[\begin{array}{ccc|cc|c} x_1 & x_2 & s_1 & s_2 & s_3 & \\ \hline 0 & 1 & 1 & 0 & -2 & 40 \\ 0 & 5 & 0 & 1 & -3 & 300 \\ 1 & 1 & 0 & 0 & 1 & 100 \\ \hline 0 & -3 & 0 & 0 & 12 & 1200 \end{array}\right]$$

$$\left[\begin{array}{ccc|cc|c} x_1 & x_2 & s_1 & s_2 & s_3 & \\ \hline 0 & 1 & 1 & 0 & -2 & 40 \\ 0 & 5 & 0 & 1 & -3 & 300 \\ 1 & 1 & 0 & 0 & 1 & 100 \\ \hline 0 & -3 & 0 & 0 & 12 & 1200 \end{array}\right] \sim \left[\begin{array}{ccc|cc|c} x_1 & x_2 & s_1 & s_2 & s_3 & \\ \hline 0 & 1 & 1 & 0 & -2 & 40 \\ 0 & 0 & -5 & 1 & 7 & 100 \\ 1 & 0 & -1 & 0 & 3 & 60 \\ \hline 0 & 0 & 3 & 0 & 6 & 1320 \end{array}\right].$$

The optimal solutions has $x_1 = 60$ regular type, $x_2 = 40$ super type, with a profit of \$1320.

The optimal values of the y_i can be read off in the columns of the slack variables in the objective function row: $y_1 = 3$ profit per 10 minutes of assembly time or \$0.30 per minute, $y_2 = 0$ profit per 5 minutes of painting time, and $y_3 = 2$ profit per 12 minutes of inspection or \$0.16 per minute . The values of the y_i correspond to the *marginal value* of each additional unit of assembly, painting, and inspection time. Additional units of the exhausted resources, assembly and inspection times, contribute to the profit but not from painting time, which is slack. ■

1.4.1. Duality for Non-standard Linear Programs

We have discussed the dual of a standard MLP. The proof of the Duality Theorem 1.15 given later in the section shows how to form the dual linear program in general. The following table indicates the rules for forming a dual linear program for non-standard conditions on variables and constraint inequalities. Starting on either side, the corresponding condition on the other side of the same row gives the condition on the dual linear program.

Rules for Forming the Dual LP	
Maximization Problem, MLP	**Minimization Problem, mLP**
i^{th} constraint $\sum_j a_{ij}x_j \le b_i$	i^{th} variable $0 \le y_i$
i^{th} constraint $\sum_j a_{ij}x_j \ge b_i$	i^{th} variable $0 \ge y_i$
i^{th} constraint $\sum_j a_{ij}x_j = b_i$	i^{th} variable y_i is unrestricted
j^{th} variable $0 \le x_j$	j^{th} constraint $\sum_i a_{ij}y_i \ge c_j$
j^{th} variable $0 \ge x_j$	j^{th} constraint $\sum_i a_{ij}y_i \le c_j$
j^{th} variable x_j is unrestricted	j^{th} constraint $\sum_i a_{ij}y_i = c_j$

Note that the standard conditions for the MLP, $\sum_j a_{ij}x_j \le b_i$ or $0 \le x_j$, corresponds to the standard conditions for the mLP, $0 \le y_i$ or $\sum_i a_{ij}x_i \ge c_j$. Nonstandard conditions for MLP, $\ge b_i$ or $0 \ge x_j$, corresponds to nonstandard conditions for mLP, $0 \ge y_i$ or $\le c_j$. Finally, an equality constraint for either optimization problem corresponds to an unrestricted variable in its dual problem.

Example 1.10.

$$\begin{aligned} \text{Minimize :} \quad & 10\,y_1 + 4\,y_2 + 8\,y_3 \\ \text{Subject to :} \quad & 2\,y_1 - 3\,y_2 + 4\,y_3 \ge 2 \\ & 3\,y_1 + 5\,y_2 + 2\,y_3 \le 15 \\ & 2\,y_1 + 4\,y_2 + 6\,y_3 = 4 \\ & y_1 \ge 0,\, y_2 \ge 0,\; y_3 \le 0 \end{aligned}$$

Using the table on the relations starting from the minimization, we get the dual maximization problem as follows:

New objective function	$2\,x_1 + 15\,x_2 + 4\,x_3$.
$2\,y_1 - 3\,y_2 + 4\,y_3 \ge 2$	implies $0 \le x_1$.
$3\,y_1 + 5\,y_2 + 2\,y_3 \le 15$	implies $0 \ge x_2$.
$2\,y_1 + 4\,y_2 + 6y_3 = 4$	implies x_3 unrestricted.
$0 \le y_1$	implies $2\,x_1 + 3\,x_2 + 2\,x_3 \le 10$.
$0 \le y_2$	implies $-3\,x_1 + 5\,x_2 + 4\,x_3 \le 4$.
$0 \ge y_3$	implies $4\,x_1 + 2\,x_2 + 6\,x_3 \ge 8$

By making the change of variables $x_2 = -z_2$ and $x_3 = z_3 - v_3$, all the variables are now restricted to be greater than or equal to zero. Summarizing the new maximization problem:

$$\begin{aligned} \text{Maximize :} \quad & 2\,x_1 - 15\,z_2 + 4\,z_3 - 4\,v_3 \\ \text{Subject to :} \quad & 2\,x_1 - 3\,z_2 + 2\,z_3 - 2\,v_3 \le 10 \\ & -3\,x_1 - 5\,z_2 + 4\,z_3 - 4\,v_3 \le 4 \\ & 4\,x_1 - 2\,z_2 + 6\,z_3 - 6\,v_3 \ge 8 \\ & x_1 \ge 0,\, z_2 \ge 0,\, z_3 \ge 0,\, v_3 \ge 0. \end{aligned}$$

The tableau for the maximization problem with variables x_1, z_2, z_3, v_3, with slack variables s_1 and s_2, and with surplus and artificial variables s_3 and r_3 is

$$\left[\begin{array}{cccc|cccc|c} x_1 & z_2 & z_3 & v_3 & s_1 & s_2 & s_3 & r_3 & \\ \hline 2 & -3 & 2 & -2 & 1 & 0 & 0 & 0 & 10 \\ -3 & -5 & 4 & -4 & 0 & 1 & 0 & 0 & 4 \\ 4 & -2 & 6 & -6 & 0 & 0 & -1 & 1 & 8 \\ \hline -2 & 15 & -4 & 4 & 0 & 0 & 0 & 0 & 0 \\ -4 & 2 & -6 & 6 & 0 & 0 & 1 & 0 & -8 \end{array}\right]$$

$$\sim \left[\begin{array}{cccc|cccc|c} x_1 & z_2 & z_3 & v_3 & s_1 & s_2 & s_3 & r_3 & \\ \hline 0 & -2 & -1 & 1 & 1 & 0 & \frac{1}{2} & -\frac{1}{2} & 6 \\ 0 & -\frac{13}{2} & \frac{17}{2} & -\frac{17}{2} & 0 & 1 & -\frac{3}{4} & \frac{3}{4} & 10 \\ 1 & -\frac{1}{2} & \frac{3}{2} & -\frac{3}{2} & 0 & 0 & -\frac{1}{4} & \frac{1}{4} & 2 \\ \hline 0 & 14 & -1 & 1 & 0 & 0 & -\frac{1}{2} & \frac{1}{2} & 4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right].$$

The artificial objective function is no longer needed but we keep the artificial variable to determine the value of its dual variable.

$$\sim \left[\begin{array}{cccc|cccc|c} x_1 & z_2 & z_3 & v_3 & s_1 & s_2 & s_3 & r_3 & \\ \hline 0 & -\frac{47}{17} & 0 & 0 & 1 & \frac{2}{17} & \frac{7}{17} & -\frac{7}{17} & \frac{122}{17} \\ 0 & -\frac{13}{17} & 1 & -1 & 0 & \frac{2}{17} & -\frac{3}{34} & \frac{3}{34} & \frac{20}{17} \\ 1 & \frac{11}{17} & 0 & 0 & 0 & -\frac{3}{17} & -\frac{2}{17} & \frac{2}{17} & \frac{4}{17} \\ \hline 0 & \frac{225}{17} & 0 & 0 & 0 & \frac{2}{17} & -\frac{10}{17} & \frac{10}{17} & \frac{88}{17} \end{array}\right]$$

The solution of the maximization problem is $x_1 = \frac{4}{17}$, $x_2 = -z_2 = 0$, $x_3 = z_3 - v_3 = \frac{20}{17} - 0 = \frac{20}{17}$, $s_1 = \frac{122}{17}$, and $s_2 = s_3 = r_3 = 0$, with a maximal value of $\frac{88}{17}$.

According to Theorem 1.16, the optimal solution for the original minimization problem can be also read off from the slack and artificial columns of the final tableau, $y_1 = 0$, $y_2 = \frac{2}{17}$, $y_3 = \frac{10}{17}$, and optimal value $10\,(0) + 4\left(\frac{2}{17}\right) + 8\left(\frac{10}{17}\right) = \frac{88}{17}$.

Alternatively, we could first write the minimization problem in standard form by setting $y_3 = -u_3$,

$$\begin{array}{ll} \text{Minimize :} & 10y_1 + 4y_2 - 8y_3 \\ \text{Subject to :} & 2y_1 - 3y_2 - 4u_3 \geq 2 \\ & 3y_1 + 5y_2 - 2u_3 \leq 15 \\ & 2y_1 + 4y_2 - 6u_3 = 4 \\ & y_1 \geq 0, y_2 \geq 0, u_3 \geq 0. \end{array}$$

The dual maximization problem will now have a different tableau than before but the same solution for the original problem. ■

1.4.2. Duality Theorems

We present a sequence of duality results giving the relationship between the solutions of dual linear programs and then summarize these results.

Notation for Duality Theorems

The *(primal) maximization linear programming problem MLP* with feasible set $\mathscr{F}_M$ is

Maximize: $f(\mathbf{x}) = \mathbf{c} \cdot \mathbf{x}$
Subject to: $\sum_j a_{ij}\, x_j \le b_i,\ \ge b_i,$ or $= b_i$ for $1 \le i \le m$ and
$x_j \ge 0,\ \le 0,$ or unrestricted for $1 \le j \le n$.

The *(dual) minimization linear programming problem mLP* with feasible set $\mathscr{F}_m$ is

Minimize: $g(\mathbf{y}) = \mathbf{b} \cdot \mathbf{y}$
Subject to: $\sum_i a_{ij}\, y_i \ge c_j,\ \le c_j,$ or $= c_j$ for $1 \le j \le n$ and
$y_i \ge 0,\ \le 0,$ or unrestricted for $1 \le i \le m$.

Theorem 1.11 (Weak Duality Theorem). *Assume that $\mathbf{x} \in \mathscr{F}_M$ is a feasible solution for a primal maximization linear programming problem MLP and $\mathbf{y} \in \mathscr{F}_m$ is a feasible solution for its dual minimization linear programming problem mLP.*

a. *Then $\mathbf{c} \cdot \mathbf{x} \le \mathbf{b} \cdot \mathbf{y}$. Thus, the optimal value M to either problem must satisfy $\mathbf{c} \cdot \mathbf{x} \le M \le \mathbf{y} \cdot \mathbf{b}$.*

b. *Further, $\mathbf{c} \cdot \mathbf{x} = \mathbf{b} \cdot \mathbf{y}$ iff*

$$0 = \mathbf{y} \cdot (\mathbf{b} - \mathbf{A}\mathbf{x}) \qquad \textit{and}$$
$$0 = \mathbf{x} \cdot \left(\mathbf{A}^{\mathsf{T}}\mathbf{y} - \mathbf{c}\right).$$

Remark. The equations of part (b) of the theorem are known as *complementary slackness*. They imply that for $1 \le j \le m$ either

$$y_j = 0 \quad \text{or} \quad 0 = b_j - a_{j1}x_1 - \cdots - a_{jn}x_n;$$

similarly for $1 \le i \le n$, either

$$x_i = 0 \quad \text{or} \quad 0 = a_{1i}y_1 + \cdots + a_{mi}y_m - c_i.$$

In the material on nonlinear equations, we have a similar result that gives the necessary conditions for a maximum, called the Karush-Kuhn-Tucker equations. We usually solve linear programming problems by the simplex method, i.e., by row reduction. For nonlinear programming with inequalities, we often solve them using the Karush-Kuhn-Tucker equations.

Proof. (**1**) If $\sum_j a_{ij}x_j \le b_i$ then $y_i \ge 0$, so

$$y_i\,(\mathbf{A}\mathbf{x})_i = y_i \sum_j a_{ij}x_j \le y_i\, b_i.$$

If $\sum_j a_{ij}x_j \ge b_i$ then $y_i \le 0$, so the same inequality holds. If $\sum_j a_{ij}x_j = b_i$ then y_i is arbitrary, so

$$y_i\,(\mathbf{A}\mathbf{x})_i = y_i \sum_j a_{ij}x_j = y_i\, b_i.$$

Summing over i,

$$\mathbf{y} \cdot \mathbf{A}\mathbf{x} = \sum_i y_i (\mathbf{A}\mathbf{x})_i \le \sum_i y_i\, b_i = \mathbf{y} \cdot \mathbf{b} \qquad \text{or}$$
$$\mathbf{y}^{\mathsf{T}}\,(\mathbf{b} - \mathbf{A}\mathbf{x}) \ge 0.$$

(**2**) By same type of argument as (1),

$$\mathbf{c} \cdot \mathbf{x} \le \mathbf{x} \cdot \left(\mathbf{A}^{\mathsf{T}}\mathbf{y}\right) = \left(\mathbf{A}^{\mathsf{T}}\mathbf{y}\right)^{\mathsf{T}}\mathbf{x} = \mathbf{y}^{\mathsf{T}}\,(\mathbf{A}\mathbf{x}) = \mathbf{y} \cdot \mathbf{A}\mathbf{x} \qquad \text{or}$$
$$\left(\mathbf{A}^{\mathsf{T}}\mathbf{y} - \mathbf{c}\right)^{\mathsf{T}}\mathbf{x} \ge 0.$$

Combining gives part (a),

$$\mathbf{c} \cdot \mathbf{x} \le \mathbf{y} \cdot \mathbf{A}\mathbf{x} \le \mathbf{y} \cdot \mathbf{b}.$$

Also,

$$\mathbf{y}\cdot\mathbf{b}-\mathbf{c}\cdot\mathbf{x}=\mathbf{y}^{\mathsf{T}}(\mathbf{b}-\mathbf{A}\mathbf{x})+\left(\mathbf{A}^{\mathsf{T}}\mathbf{y}-\mathbf{c}\right)^{\mathsf{T}}\mathbf{x}=0 \qquad \text{iff}$$
$$0=\left(\mathbf{A}^{\mathsf{T}}\mathbf{y}-\mathbf{c}\right)\cdot\mathbf{x} \quad \text{and} \quad 0=\mathbf{y}\cdot(\mathbf{b}-\mathbf{A}\mathbf{x}).$$

This proves part (b). □

Corollary 1.12 (**Feasibility/Boundedness).** *Assume that MLP and mLP both have feasible solutions. Then, MLP is bounded above and has an optimal solution. Also, mLP is bounded below and has an optimal solution.*

Proof. If mLP has a feasible solution $\mathbf{y}_0\in\boldsymbol{\mathscr{F}}_m$, then for any feasible solution $\mathbf{x}\in\boldsymbol{\mathscr{F}}_M$ of MLP, $f(\mathbf{x})=\mathbf{c}\cdot\mathbf{x}\le\mathbf{b}\cdot\mathbf{y}_0$ so f is bounded above and has an optimal solution.

Similarly, if MLP has a feasible solution $\mathbf{x}_0\in\boldsymbol{\mathscr{F}}_M$, then for any feasible solution $\mathbf{y}\in\boldsymbol{\mathscr{F}}_m$ of mLP, $g(\mathbf{y})=\mathbf{b}\cdot\mathbf{y}\ge\mathbf{c}\cdot\mathbf{x}_0$ so g is bounded below and has an optimal solution. □

Proposition 1.13 (**Necessary Conditions).** *If $\bar{\mathbf{x}}$ is an optimal solution for MLP, then there is a feasible solution $\bar{\mathbf{y}}\in\boldsymbol{\mathscr{F}}_m$ of the dual mLP that satisfies complementary slackness equations,*

$$0=\bar{\mathbf{y}}\cdot(\mathbf{b}-\mathbf{A}\bar{\mathbf{x}}) \qquad \textit{and}$$
$$0=\left(\mathbf{A}^{\mathsf{T}}\bar{\mathbf{y}}-\mathbf{c}\right)\cdot\bar{\mathbf{x}}.$$

Similarly, if $\bar{\mathbf{y}}$ is an optimal solution for mLP, then there is a feasible solution $\bar{\mathbf{x}}\in\boldsymbol{\mathscr{F}}_M$ of the Dual MLP that that satisfies complementary slackness equations,

Proof. Let $\mathbf{E}$ be the set of i such that $b_i\ge\sum_j a_{ij}\bar{x}_j$ is tight or effective. The i^{th}-row of $\mathbf{A}$, $\mathbf{R}_i^{\mathsf{T}}$, is the gradient of this constraint. Let $\mathbf{E}'$ the set of i such that $x_i=0$, so this constraint is tight. The gradient of this constraint is the standard unit vector $\mathbf{e}_i=(0,\dots,1,\dots 0)^{\mathsf{T}}$, with a 1 only in the i^{th}-coordinate.

We assume that this solution is nondegenerate so that the gradients of the constraints $\{\mathbf{R}_i^{\mathsf{T}}\}_{i\in\mathbf{E}}\cup\{-\mathbf{e}^i\}_{i\in\mathbf{E}'}=\{\mathbf{w}_j\}_{j\in\mathbf{E}''}$ are linearly independent. (Otherwise we have to take an appropriate subset in the following argument.)

The objective function f has a local maximum on the level set for the tight constraints at $\bar{\mathbf{x}}$. By Lagrange multipliers, $\nabla(f)=\mathbf{c}$ is a linear combination of the gradients of the tight constraints,

$$\mathbf{c}=\sum_{i\in\mathbf{E}}\bar{y}_i\mathbf{R}_i^{\mathsf{T}}-\sum_{i\in\mathbf{E}'}\bar{z}_i\mathbf{e}_i.$$

By setting $\bar{y}_i=0$ for $i\notin\mathbf{E}$ and $\bar{z}_i=0$ for $i\notin\mathbf{E}'$, the sum can be extended to all the appropriate i,

$$\mathbf{c}=\sum_{1\le i\le m}\bar{y}_i\mathbf{R}_i^{\mathsf{T}}-\sum_{1\le i\le n}\bar{z}_i\mathbf{e}_i=\mathbf{A}^{\mathsf{T}}\bar{\mathbf{y}}-\bar{\mathbf{z}}. \qquad (*)$$

Since $\bar{y}_i=0$ for $b_i-\sum_j a_{ij}\bar{x}_j\ne 0$ and $\bar{z}_i=0$ for $x_i\ne 0$,

$$0=\bar{y}_i\left(b_i-\sum_j a_{ij}\bar{x}_j\right) \qquad \text{for } 1\le i\le m \quad \text{and}$$
$$0=\bar{z}_i\bar{x}_i \qquad \text{for } 1\le i\le n,$$

or in vector form using (*)

$$0=\bar{\mathbf{y}}\cdot(\mathbf{b}-\mathbf{A}\bar{\mathbf{x}}) \qquad \text{and}$$
$$0=\bar{\mathbf{z}}\cdot\bar{\mathbf{x}}=\bar{\mathbf{x}}\cdot\left(\mathbf{A}^{\mathsf{T}}\bar{\mathbf{y}}-\mathbf{c}\right).$$

To finish the proof, we need to show that (i) $\bar{y}_i \geq 0$ for a resource constraint ($\leq b_i$), (ii) $\bar{y}_i \leq 0$ for a requirement constraint ($\geq b_i$), (iii) $\bar{y}_i$ is unrestricted for an equality constraint ($= b_i$), (iv) $\bar{z}_j = \sum_i a_{ij}\bar{y}_i - c_j \geq 0$ for $x_j \geq 0$, (iv) $\bar{z}_j = \sum_i a_{ij}\bar{y}_i - c_j \leq 0$ for $x_j \leq 0$, (iv) $\bar{z}_j = \sum_i a_{ij}\bar{y}_i - c_j = 0$ for x_j unrestricted,

The set of vectors $\{\mathbf{R}_i^\mathsf{T}\}_{i\in\mathbf{E}} \cup \{\text{-}\mathbf{e}^i\}_{i\in\mathbf{E}'} = \{\mathbf{w}_j\}_{j\in\mathbf{E}''} = \{\mathbf{w}_k\}_{k\in\mathbf{E}''}$ are linearly independent, so we can complete it to a basis $\{\mathbf{w}_k$ of $\mathbb{R}^n$ using vectors perpendicular to these first vectors. Let $\mathbf{W}$ be the $n \times n$ matrix with these $\mathbf{w}_j$ as columns. Then $\mathbf{W}^\mathsf{T}$ is invertible because its rows are linearly independent, so there is an $n \times n$ matrix $\mathbf{V}$ such that $\mathbf{W}^\mathsf{T}\mathbf{V} = \mathbf{I}$. For each k, the k^{th} column of $\mathbf{V}$, $\mathbf{v}_k$, is perpendicular to all the $\mathbf{w}_i$ except $\mathbf{w}_k$. Remember that

$$\mathbf{c} = \sum_{1\leq i\leq m} \bar{y}_i \mathbf{R}_i^\mathsf{T} - \sum_{1\leq i\leq n} \bar{z}_i \mathbf{e}_i = \sum_j p_j \mathbf{w}_j$$

where p_j is corresponding $\bar{y}_i$, $\bar{z}_i$, or 0. Then

$$\mathbf{c}\cdot\mathbf{v}_k = \left(\sum_j p_j \mathbf{w}_j\right)\cdot\mathbf{v}_k = p_k.$$

Take $i \in \mathbf{E}$. The gradient of this constraint is $\mathbf{R}_i^\mathsf{T} = \mathbf{w}_k$ for some $k \in \mathbf{E}''$. Set $\delta = \text{-}1$ for a resource constraint and $\delta = 1$ for a requirement constraint. The vector $\delta\mathbf{R}_i^\mathsf{T}$ points into $\mathscr{F}$ in both cases (except equality constraint). For small $t \geq 0$, $\bar{\mathbf{x}} + t\,\mathbf{v}_k \in \mathscr{F}$, so

$$0 \leq f(\bar{\mathbf{x}}) - f(\bar{\mathbf{x}} + t\,\mathbf{v}_k) = \text{-}\,t\,\delta\,\mathbf{c}\cdot\mathbf{v}_k = \text{-}\,t\,p_k.$$

Therefore, $\delta\,\bar{y}_i = \delta\,p_k \leq 0$, and $\bar{y}_i \geq 0$ for a resource constraint and ≤ 0 for a requirement constraint. For an equality constraint, we are not allowed to move off in either direction so $\bar{y}_i$ is unrestricted.

Next take $i \in \mathbf{E}'$, with $\text{-}\mathbf{e}_i = w_k$ for some $k \in \mathbf{E}''$. Set $\delta = \text{-}1$ if $x_i \geq 0$ and $\delta = 1$ if $x_i \leq 0$. Then $\delta\,\mathbf{w}_k = \text{-}\delta\,\mathbf{e}_i$ points into $\mathscr{F}$ (unless x_i is unrestricted). By an argument as before, $\text{-}\delta\,p_k = \text{-}\delta\,\bar{z}_i \geq 0$. Therefore, $\bar{z}_i \geq 0$ if $x_i \geq 0$ and $\bar{z}_i \leq 0$ if $x_i \leq 0$. If x_i is unrestricted, then the equation is not tight and $\bar{z}_i = 0$.

We have shown that $\bar{y} \in \mathscr{F}_m$ and satisfies complementary slackness.

The proof in the case of an optimal solution of mLP is similar. □

Corollary 1.14 (Optimality and Complementary Slackness). *Assume that $\bar{\mathbf{x}} \in \mathscr{F}_M$ is a feasible solution for a primal MLP and $\bar{\mathbf{y}} \in \mathscr{F}_m$ is a feasible solution for the dual mLP. Then the following are equivalent.*

a. $\bar{\mathbf{x}}$ *is an optimal solution of MLP and* $\bar{\mathbf{y}}$ *is an optimal solution of mLP.*

b. $\mathbf{c}\cdot\bar{\mathbf{x}} = \mathbf{b}\cdot\bar{\mathbf{y}}$.

c. $\mathbf{0} = \bar{\mathbf{x}}\cdot(\mathbf{c} - \mathbf{A}^\mathsf{T}\bar{\mathbf{y}})$ *and* $\mathbf{0} = (\mathbf{b} - \mathbf{A}\bar{\mathbf{x}})\cdot\bar{\mathbf{y}}$.

Proof. (**b** $\Leftrightarrow$ **c**) This is a restatement of the Weak Duality Theorem 1.11(b).

(**a** $\Rightarrow$ **c**) By Proposition 1.13, there exists $\bar{\mathbf{y}}'$ of mLP that satisfies the complementary slackness equations. By the Weak Duality Theorem 1.11, $\mathbf{c}\cdot\bar{\mathbf{x}} = \mathbf{b}\cdot\bar{\mathbf{y}}'$, $\mathbf{c}\cdot\bar{\mathbf{x}} \leq \mathbf{b}\cdot\bar{\mathbf{y}} \leq \mathbf{b}\cdot\bar{\mathbf{y}}' = \mathbf{c}\cdot\bar{\mathbf{x}}$, and so $\mathbf{c}\cdot\bar{\mathbf{x}} = \mathbf{b}\cdot\bar{\mathbf{y}}$. Again by the Weak Duality Theorem 1.11, $\bar{\mathbf{y}}$ must satisfy the complementary slackness equations.

(**b** $\Rightarrow$ **a**) Assume $\bar{\mathbf{y}}$ and $\bar{\mathbf{x}}$ satisfy $\mathbf{c}\cdot\bar{\mathbf{x}} = \mathbf{b}\cdot\bar{\mathbf{y}}$. Then, for any other feasible solutions $\mathbf{x} \in \mathscr{F}_M$ and $\mathbf{y} \in \mathscr{F}_m$, $\mathbf{c}\cdot\mathbf{x} \leq \mathbf{b}\cdot\bar{\mathbf{y}} = \mathbf{c}\cdot\bar{\mathbf{x}} \leq \mathbf{b}\cdot\mathbf{y}$, so $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ must be optimal solutions. □

Theorem 1.15 (**Duality Theorem**). *Consider dual problems MLP and mLP. Then, MLP has an optimal solution iff the dual mLP has an optimal solution.*

Proof. If MLP has an optimal solution $\bar{\mathbf{x}}$, then mLP has a feasible solution $\bar{\mathbf{y}}$ that satisfies the complementary slackness equations by the necessary conditions of Proposition 1.13. By Corollary 1.14, $\bar{\mathbf{y}}$ is an optimal solution of mLP. □

Summary of Duality Results:

1. If $\mathbf{x} \in \mathscr{F}_{\boldsymbol{M}}$ and $\mathbf{y} \in \mathscr{F}_{\boldsymbol{m}}$, then $\mathbf{c} \cdot \mathbf{x} \leq \mathbf{b} \cdot \mathbf{y}$. (Weak Duality Theorem 1.11)
2. MLP has an optimal solution iff mLP has an optimal solution. (Duality Theorem 1.15)
3. If $\bar{\mathbf{x}} \in \mathscr{F}_{\boldsymbol{M}} \neq \emptyset$ and $\bar{\mathbf{y}} \in \mathscr{F}_{\boldsymbol{m}} \neq \emptyset$, then MLP and mLP each has an optimal solution. (Corollary 1.12)
4. If $\bar{\mathbf{x}} \in \mathscr{F}_{\boldsymbol{M}}$ and $\bar{\mathbf{y}} \in \mathscr{F}_{\boldsymbol{m}}$, then the three conditions **a**–**c** of Corollary 1.14 are equivalent.

Theorem 1.16 (**Duality and Tableau**). *If either MLP or mLP is solved for an optimal solution by the simplex method, then the solution of its dual is displayed in the bottom row of the optimal tableau in the columns associated with the slack and artificial variables (not those for the surplus variables).*

Proof. To use the tableau to solve MLP, we need all the $x_i \geq 0$, so all the constraints of the dual mLP will be requirement constraints. We group the equations into resource, requirement, and equality constraints and get the tableau for MLP

$$\left[\begin{array}{c|cc|cc|c} \mathbf{A}_1 & \mathbf{I}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{b}_1 \\ \mathbf{A}_2 & \mathbf{0} & -\mathbf{I}_2 & \mathbf{I}_2 & \mathbf{0} & \mathbf{b}_2 \\ \mathbf{A}_3 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_3 & \mathbf{b}_3 \\ \hline -\mathbf{c}^{\mathsf{T}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & 0 \end{array}\right].$$

Row operations to optimal tableau are realized by multiplying on the left by a matrix. The last row is not added to the other rows so the row reduction to the optimal tableau is as follows:

$$\left[\begin{array}{ccc|c} \mathbf{M}_1 & \mathbf{M}_2 & \mathbf{M}_3 & \mathbf{0} \\ \hline \bar{\mathbf{y}}_1^{\mathsf{T}} & \bar{\mathbf{y}}_2^{\mathsf{T}} & \bar{\mathbf{y}}_2^{\mathsf{T}} & 1 \end{array}\right] \left[\begin{array}{c|cc|cc|c} \mathbf{A}_1 & \mathbf{I}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{b}_1 \\ \mathbf{A}_2 & \mathbf{0} & -\mathbf{I}_2 & \mathbf{I}_2 & \mathbf{0} & \mathbf{b}_2 \\ \mathbf{A}_3 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_3 & \mathbf{b}_3 \\ \hline -\mathbf{c}^{\mathsf{T}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & 0 \end{array}\right]$$

$$= \left[\begin{array}{c|cc|cc|c} \mathbf{M}_1\mathbf{A}_1 + \mathbf{M}_2\mathbf{A}_2 + \mathbf{M}_3\mathbf{A}_3 & \mathbf{M}_1 & -\mathbf{M}_2 & \mathbf{M}_2 & \mathbf{M}_3 & \mathbf{Mb} \\ \hline \bar{\mathbf{y}}_1^{\mathsf{T}}\mathbf{A}_1 + \bar{\mathbf{y}}_2^{\mathsf{T}}\mathbf{A}_2 + \bar{\mathbf{y}}_3^{\mathsf{T}}\mathbf{A}_3 - \mathbf{c}^{\mathsf{T}} & \bar{\mathbf{y}}_1^{\mathsf{T}} & -\bar{\mathbf{y}}_2^{\mathsf{T}} & \bar{\mathbf{y}}_2^{\mathsf{T}} & \bar{\mathbf{y}}_3^{\mathsf{T}} & \bar{\mathbf{y}}^{\mathsf{T}}\mathbf{b} \end{array}\right].$$

In the optimal tableau, the entries in the objective function row are nonnegative (except for artificial variable columns), so $\bar{\mathbf{y}}^{\mathsf{T}}\mathbf{A} - \mathbf{c}^{\mathsf{T}} \geq \mathbf{0}$, $\bar{\mathbf{y}}_1^{\mathsf{T}} \geq \mathbf{0}$, $\bar{\mathbf{y}}_2^{\mathsf{T}} \leq \mathbf{0}$. Thus, $\bar{\mathbf{y}}$ is a feasible solution of the mLP. If $\bar{\mathbf{x}}$ is a maximizer for the MLP, the value $\bar{\mathbf{y}}^{\mathsf{T}}\mathbf{b} = \mathbf{b} \cdot \bar{\mathbf{y}}$ in the lower right position of the optimal tableau is the optimal value of the MLP, so equals $\mathbf{c} \cdot \bar{\mathbf{x}}$. By the Optimality Corollary 1.14, $\bar{\mathbf{y}}$ is an optimal solution of mLP.

Note that $\left(\mathbf{A}^{\mathsf{T}}\bar{\mathbf{y}}\right)_i = \mathbf{L}_i \cdot \bar{\mathbf{y}}$ where $\mathbf{L}_i$ is the i^{th}-column of $\mathbf{A}$.

In the case when $x_i \leq 0$, we start by setting $\xi_i = -x_i \geq 0$. The column in the tableau is -1 time original column, and the new object function coefficient is $-c_i$. By the earlier argument for $\xi_i \geq 0$, we get $0 \leq (-\mathbf{L}_i) \cdot \bar{\mathbf{y}} - (-c_i)$, or

$$\mathbf{L}_i \cdot \bar{\mathbf{y}} \leq c_i,$$

which is a resource constraint for dual the problem as claimed.

In the case when x_i is arbitrary, we start by setting $x_i = \xi_i - \eta_i$. By the previous two cases, we get both

$$\begin{aligned} \mathbf{L}_i \cdot \bar{\mathbf{y}} &\geq c_i \qquad \text{and} \\ \mathbf{L}_i \cdot \bar{\mathbf{y}} &\leq c_i, \qquad \text{or} \\ \mathbf{L}_i \cdot \bar{\mathbf{y}} &= c_i, \end{aligned}$$

which is an equality constraint for the dual problem as claimed. □

1.4. Exercises

1.4.1. Determine the dual of the linear program:

$$\begin{aligned} \text{Minimize:} \quad & 4y_1 + 3y_2 + 8y_3 \\ \text{Subject to:} \quad & y_1 + y_2 + y_3 \geq 12 \\ & 5y_1 - 2y_2 + 4y_3 \leq 20 \\ & 2y_1 + 3y_2 - y_3 = 12 \\ & 0 \leq y_1,\ 0 \leq y_2,\ y_3 \text{ unrestricted.} \end{aligned}$$

1.4.2. Consider the following minimization linear programming problem, mLP:

$$\begin{aligned} \text{Minimize:} \quad & 8y_1 + 6y_2 \\ \text{Subject to:} \quad & 2y_1 + y_2 \geq 3 \\ & y_1 + y_2 \geq 2 \\ & y_1 \geq 0, y_2 \geq 0. \end{aligned}$$

a. Form the dual problem maximization MLP linear problem.
b. Solve the dual problem MLP by the simplex method. Give the optimal solution of the dual maximization problem and the maximal value.
c. Give the optimal solution of the original minimization problem mLP and the minimal value.

1.4.3. Consider the linear programming problem

$$\begin{aligned} \text{Maximize :} \quad & f(x,y) = 2\,x_1 + 3\,x_2 + 5\,x_3, \\ \text{Subject to :} \quad & 3\,x_1 + 4\,x_2 - 2\,x_3 \leq 10, \\ & -x_1 + 2\,x_2 + x_3 \leq 3, \\ & x_1 \geq 0,\ x_2 \geq 0.\ x_3 \geq 0. \end{aligned}$$

a. Using the simplex method, find an optimal solution. Give all the values (i) of the variables x_1, x_2, and x_3, (ii) of the the slack variables, and (iii) the maximal value of the objective function.
b. State the dual linear programming problem.
c. What is the optimal solution of the dual linear programming problem?

1.4.4. Use the dual problem MLP to discover the nature of the solution to the following minimization linear programming problem, mLP:

$$\begin{aligned} \text{Minimize:} \quad & -y_1 + 2\,y_2 \\ \text{Subject to:} \quad & -5y_1 + y_2 \geq 2 \\ & 4y_1 - y_2 \geq 3 \\ & y_1 \geq 0, y_2 \geq 0. \end{aligned}$$

1.5. Sensitivity Analysis

In this section, we investigate for what range of change in b_i the optimal value y_i^* of the dual problem correctly gives the change in the optimal value of the objective function. We also consider what marginal effect a change in the coefficients c_i has on the optimal value.

Example 1.17. A company produces two products, 1 and 2, with profits per item of \$40 and \$10 respectively. In the short run in stock, there are only 1020 unit of paint, 400 fasteners, and 420 hours of labor. Each unit of each product requires 15 and 10 units of paint respectively, 10 and 2 fasteners, and 3 and 5 hours of labor. The tableau for the maximization problem is

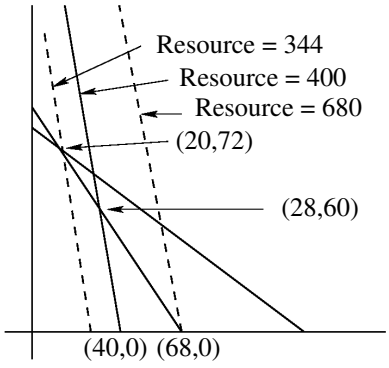
$$\left[\begin{array}{cc|ccc|c} x_1 & x_2 & s_1 & s_2 & s_3 & \\ \hline 15 & 10 & 1 & 0 & 0 & 1020 \\ 10 & 2 & 0 & 1 & 0 & 400 \\ 3 & 5 & 0 & 0 & 1 & 420 \\ \hline -40 & -10 & 0 & 0 & 0 & 0 \end{array}\right] \sim \left[\begin{array}{cc|ccc|c} x_1 & x_2 & s_1 & s_2 & s_3 & \\ \hline 0 & 7 & 1 & -1.5 & 0 & 420 \\ 1 & .2 & 0 & .1 & 0 & 40 \\ 0 & 4.4 & 0 & -.3 & 1 & 300 \\ \hline 0 & -2 & 0 & 4 & 0 & 1600 \end{array}\right]$$

$$\sim \left[\begin{array}{cc|ccc|c} x_1 & x_2 & s_1 & s_2 & s_3 & \\ \hline 0 & 1 & \frac{1}{7} & -\frac{3}{14} & 0 & 60 \\ 1 & 0 & -\frac{1}{35} & \frac{1}{7} & 0 & 28 \\ 0 & 0 & -\frac{22}{35} & \frac{9}{14} & 1 & 36 \\ \hline 0 & 0 & \frac{2}{7} & \frac{25}{7} & 0 & 1720 \end{array}\right].$$

The optimal solution is

$$x_1 = 28,\ \ x_2 = 60,\ \ s_1 = 0,\ \ s_2 = 0,\ \ s_3 = 36,$$

with optimal value of 1720. The values of an increase of the constrained quantities are $y_1 = \frac{2}{7}$ and $y_2 = \frac{25}{7}$, and $y_3 = 0$ for the quantity that is not tight.



We begin discussing the affect of its change to the limitation on fasteners, $b_2 + \delta_2$, while keeping the same basic variables and free variables $s_1 = s_2 = 0$. The starting form of the constraint is

$$10x_1 + 2x_2 + s_2 = 400 + \delta_2.$$

The slack variable s_2 and δ_2 play similar roles (and have similar units). So in the new optimal tableau, δ_2 times the s_2-column is added to the right hand column, the column for the constraint

constants. Since we continue to need $x_1, x_2, s_3 \geq 0$,

$$\begin{aligned}
0 \leq x_2 = 60 - \frac{3\delta_2}{14} \qquad &\text{or} \qquad \delta_2 \leq 60 \cdot \frac{14}{3} = 280, \\
0 \leq x_1 = 28 + \frac{\delta_2}{7} \qquad &\text{or} \qquad \delta_2 \geq -28 \cdot \frac{7}{1} = -196, \\
0 \leq s_3 = 36 + \frac{9\delta_2}{14} \qquad &\text{or} \qquad \delta_2 \geq -36 \cdot \frac{14}{9} = -56.
\end{aligned}$$

In order to keeps x_1, x_2, and s_3 as basic variables and $s_1 = 0 = s_2$ as the non-pivot variables, the resource can be increased at most 280 units and decreased at most 56 units, or

$$344 = 400 - 56 \leq b_2 \leq 400 + 280 = 680.$$

For this range of b_2, $^{25}/_{7}$ is the marginal value of the fasteners. For $\delta_2 = 280$ and $b_2 = 680$, we have

$$\begin{aligned}
x_1 &= 28 + 280 \cdot \frac{1}{7} = 68, \\
x_2 &= 60 - 280 \cdot \frac{3}{14} = 0, \\
s_3 &= 36 + 280 \cdot \frac{9}{14} = 216, \qquad \text{and} \\
z &= 1720 + 280 \cdot \frac{25}{7} = 2720 \qquad \text{as the optimal value.}
\end{aligned}$$

There is a similar calculation for $\delta_2 = -56$ has optimal value $z = 1720 - 56 \cdot \frac{25}{7} = 1520$.

A similar consideration can be applied to the supply of paint, the first constraint. The inequalities for the optimal tableau become

$$\begin{aligned}
0 \leq x_2 = 60 + \frac{\delta_1}{7} \qquad &\text{or} \qquad \delta_1 \geq -60 \cdot \frac{7}{1} = -420, \\
0 \leq x_1 = 28 - \frac{\delta_1}{35} \qquad &\text{or} \qquad \delta_1 \leq 28 \cdot \frac{35}{1} = 980, \\
0 \leq s_3 = 36 - \frac{22\delta_1}{35} \qquad &\text{or} \qquad \delta_1 \leq 36 \cdot \frac{35}{2} 2 \approx 57.27.
\end{aligned}$$

Therefore

$$0 = 420 - 420 \leq b_1 \leq 420 + 57.27 = 477.27,$$

with change in optimal value

$$1600 = 1720 - \tfrac{2}{7} \cdot 420 \leq f \leq 1720 - \tfrac{2}{7} \cdot 57.27 = 1736.36. \qquad \blacksquare$$

We include a sensitivity analysis for the general change in a constraint to indicate how the method is applied. These formula are hard to remember, so it is probably easier to work a given problem as we did in the preceding example. For all the general statements for sensitivity analysis, let denote the entries of the optimal tableau as follows: b'_i denotes the entry for the i^{th} constraint in the right hand column, $c'_j \geq 0$ denotes an objective row entry, a'_{ij} denotes the entry in the i^{th} row and j^{th} column (excluding the right hand constants and any artificial variable columns), $\mathbf{C}'_j$ denotes the j^{th}-column of $\mathbf{A}'$, and $\mathbf{R}_i$ denotes the i^{th}-row of $\mathbf{A}'$.

Changes in a tight resource constraint

Consider a change in the constant for a tight resource constraint, $b_r + \delta_r$. Let $k = k_r$ be the column for be the slack variable s_r of the r^{th} constraint that is not a pivot column. The

initial tableau row reduces to the optimal tableau as follows:

$$\left[\begin{array}{c|c|c} & s_r & \\ \hline \mathbf{A} & \mathbf{e}^r & \mathbf{b}+\delta_r\mathbf{e}^r \\ \hline \text{-}\mathbf{c}^\mathsf{T} & 0 & 0 \end{array}\right] \sim \left[\begin{array}{c|c|c} & s_r & \\ \hline \mathbf{A}' & \mathbf{C}'_k & \mathbf{b}'+\delta_r\,\mathbf{C}'_k \\ \hline \mathbf{c}'^\mathsf{T} & c'_k & M+\delta_r\,c'_k \end{array}\right].$$

Let z_i be the basic variable with a pivot in the i^{th}-row. To keep the same basic variables, we need $0 \le z_i = b'_i + \delta_r\, a'_{ik}$ for all i. For $a'_{ik} < 0$, we need $\text{-}\delta_r\, a'_{ik} \le b'_i$; for $a'_{ik} > 0$, we need $\text{-}b'_i \le \delta_r\, a'_{ik}$. Range of change in b_r, δ_r, with the same set of basic variables satisfies

$$\text{-}\min_i \left\{ \frac{b'_i}{a'_{ik}} : a'_{ij} > 0 \right\} \le \delta_r \le \min_i \left\{ \frac{b'_i}{\text{-}a'_{ik}} : a'_{ij} < 0 \right\},$$

where $k = k_r$ is the column for be the slack variable s_r. The change in the optimal value for δ_r in allowable range is given by $\delta_r\, c'_{k_r}$. The point of restricting to this range is that the change in the optimal value and the optimizing basic solution changes in a simple fashion.

Changes in a slack resource constraint

For a slack resource constraint with s_r in a pivot column, to keep the s_r-column a pivot column, we need $b'_r + \delta_r \ge 0$. Thus, $\delta_r \ge \text{-}b'_r$ gives the amount that b_r can be decreased before the set of basic variables changes. The change δ_r can be arbitrarily large. For δ_r in this range, the optimal value is unchanged.

1.5.1. Changes in Objective Function Coefficients

Example 1.18. We consider the same problem as in Example 1.17. Changes in the coefficients c_k correspond to changes in the profits of the products. This change could be accomplished by changing the price charged for each item produced.

For a change from c_1 to $c_1 + \Delta_1$, the change in the initial and optimal tableau are as follows:

$$\left[\begin{array}{rrc|ccc|c} x_1 & x_2 & & s_1 & s_2 & s_3 & \\ \hline 15 & 10 & & 1 & 0 & 0 & 1020 \\ 10 & 2 & & 0 & 1 & 0 & 400 \\ 3 & 5 & & 0 & 0 & 1 & 420 \\ \hline \text{-}40\text{-}\Delta_1 & \text{-}10 & & 0 & 0 & 0 & 0 \end{array}\right] \sim \left[\begin{array}{rrc|ccc|c} x_1 & x_2 & & s_1 & s_2 & s_3 & \\ \hline 0 & 1 & & \frac{1}{7} & \text{-}\frac{3}{14} & 0 & 60 \\ 1 & 0 & & \text{-}\frac{1}{35} & \frac{1}{7} & 0 & 28 \\ 0 & 0 & & \text{-}\frac{22}{35} & \frac{9}{14} & 1 & 36 \\ \hline \text{-}\Delta_1 & 0 & & \frac{2}{7} & \frac{25}{7} & 0 & 1720 \end{array}\right].$$

The pivot for x_1 is in the $r = 2^{\text{nd}}$ row. To make the objective function row nonnegative, we need to add Δ_1 times the second row,

$$\sim \left[\begin{array}{rrc|ccc|c} x_1 & x_2 & & s_1 & s_2 & s_3 & \\ \hline 0 & 1 & & \frac{1}{7} & \text{-}\frac{3}{14} & 0 & 60 \\ 1 & 0 & & \text{-}\frac{1}{35} & \frac{1}{7} & 0 & 28 \\ 0 & 0 & & \text{-}\frac{22}{35} & \frac{9}{14} & 1 & 36 \\ \hline 0 & 0 & & \frac{2}{7}-\frac{1}{35}\Delta_1 & \frac{25}{7}+\frac{1}{7}\Delta_1 & 0 & 1720+28\Delta_1 \end{array}\right].$$

To keep the same basic variables, we need

$$0 \le \frac{2}{7} - \frac{1}{35}\Delta_1 \qquad \text{or} \qquad \Delta_1 \le \frac{2}{7}\cdot\frac{35}{1} = 10,$$
$$0 \le \frac{25}{7} + \frac{1}{7}\Delta_1 \qquad \text{or} \qquad \Delta_1 \ge \text{-}\frac{25}{7}\cdot\frac{7}{1} = \text{-}25.$$

Thus the coefficient satisfies

$$15 = 40 - 25 \leq c_1 \leq 40 + 10 = 50.$$

The value of the objective function for c_1 in this range is $1720 + 28\Delta_1$.

For Δ_2, the pivot for x_2 is in the $r = 1$ row, and a similar calculations shows that

$$-2 = -\frac{2}{7} \cdot \frac{7}{1} \leq \Delta_2 \leq \frac{25}{7} \cdot \frac{14}{3} = \frac{50}{3}.$$

Thus the coefficient satisfies

$$8 = 10 - 2 \leq c_2 \leq 10 + \frac{50}{3} = 26\frac{2}{3},$$

and the value of the objective function is $1720 + 420\Delta_2$. ■

For the general case, consider the changes Δ_k in the coefficient c_k of the variable x_k in the objective function, where x_k is a basic variable in the optimal solution and its pivot is in the r^{th} row, $a'_{rk} = 1$. The changed entry in the original objective row is $-c_k - \Delta_k$, and the optimal tableau changes from 0 to $-\Delta_k$. To keep x_k basic, we need to add $\Delta_k \mathbf{R}'_r$ to the objective row, where the pivot of x_k is in the r^{th} row. The entry in j^{th}-column becomes $c'_j + \Delta_k a'_{rj}$. To keep the same basic variables, the range of change Δ_k is determined so that $c'_j + \Delta_k a'_{rj} \geq 0$ for all j, with artificial variables artificial variable excluded but columns for slack and surplus variables included. For $a'_{rj} > 0$ in r^{th}-pivot row, we need $a'_{rj}\Delta_k \geq -c'_j$ or $\Delta_k \geq -\dfrac{c'_j}{a'_{rj}}$; for $a'_{rj} < 0$ in r^{th}-pivot row, we need $c'_j \geq -a'_{rj}\Delta_k$ or $\dfrac{c'_j}{-a'_{rj}} \geq \Delta_k$. Note that if $c'_j = 0$ for $a'_{rj} > 0$ and $j \neq k$, then we need $\Delta_k \geq 0$, and if $c'_j = 0$ for $a'_{rj} < 0$, then we need $\Delta_k \leq 0$.

Range of change Δ_k with the same set of basic variables:

$$-\min_j \left\{ \frac{c'_j}{a'_{rj}} : a'_{rj} > 0 \ j \neq k \right\} \leq \Delta_k \leq \min_j \left\{ \frac{c'_j}{-a'_{rj}} : a'_{rj} < 0 \right\}.$$

Change in the optimal value for Δ_k in allowable range is $b'_r \cdot \Delta_k$.

Changes in coefficient of objective function for a non-basic variable

If x_k is a non-basic variable, then in the unchanged problem $x_k = 0$. The inequality $c'_k + \Delta_k \geq 0$ insures that x_k remains a non-basic variable, or $\Delta_k \geq -c'_k$. Thus, the entry c'_k in the column for x_k indicates the amount that c_k can decrease while keeping the same set of basic variables. A decrease of more than c'_k would start to make a positive contribution to the optimal value of the objective function.

1.5. Exercises

1.5.1. The optimal tableau for the following linear program

$$\begin{array}{ll} \text{Maximize:} & 2\,x_1 + 3\,x_2 + 5\,x_3 \\ \text{Subject to:} & x_1 + x_2 + x_3 \geq 12 \\ & 5x_1 - 2x_2 + 4x_3 \leq 20 \\ & 0 \leq x_1,\ 0 \leq x_2,\ 0 \leq x_3 \end{array}$$

is the following:

$$\left[\begin{array}{ccc|cc|c} x_1 & x_2 & x_3 & s_1 & s_2 & \\ \hline 1 & 8 & 0 & \frac{5}{3} & \frac{2}{3} & 56 \\ 0 & 2 & 1 & \frac{1}{3} & \frac{1}{3} & 65 \\ \hline 0 & 23 & 0 & 5 & 3 & 177 \end{array}\right]$$

a. How much can each constant b_i increase and decrease, $b_i + \delta_i$, and keep the same set of basic variables in the optimal solution?
What is the change in the optimal value of the objective function with an allowable change δ_i in each b_i?
What is the marginal value per additional unit of a small amount of each resource?

b. Determine the range of values of the objective function coefficient of x_1 such that the optimal basis remains unchanged?

1.5.2. A farmer can grow x_1 acres of corn and x_2 acres of potatoes. The linear program to maximize the profit is the following:

Maximize:	$50\,x_1 + 40\,x_2$	Profit
Subject to:	$x_1 + x_2 \leq 50$	Acres of land
	$3\,x_1 + 2\,x_2 \leq 120$	Days of labor
	$10\,x_1 + 60\,x_2 \leq 1200$	Dollars of capital
	$20\,x_1 + 10\,x_2 \leq 800$	Pounds of fertilizer
	$0 \leq x_1,\ 0 \leq x_2.$	

The optimal tableau is the following:

x_1	x_2	s_1	s_2	s_3	s_4	
0	0	1	$-\frac{5}{16}$	$-\frac{1}{160}$	0	5
0	1	0	$-\frac{1}{16}$	$\frac{3}{160}$	0	15
0	0	0	$-\frac{55}{8}$	$\frac{1}{16}$	1	50
1	0	0	$\frac{3}{8}$	$-\frac{1}{80}$	0	30
0	0	0	$\frac{65}{4}$	$\frac{1}{8}$	0	2100

a. What is the value each additional acre of land, each additional day of labor, each additional dollar of capital, and each additional pound of fertilizer?

b. What range of the number of days of labor is the the value given in part (a) valid?

1.5.3. For the linear program in Exercise 1.5.1, what range of each of the coefficients c_k leaves the set of basic variables the same? In each case, what is the change in maximal value for the changes Δ_k?

1.5.4. For the linear program in Exercise 1.5.2, what range of each of the coefficients c_k leaves the set of basic variables the same? In each case, what is the change in maximal value for the changes Δ_k?

1.6. Theory for Simplex Method

For three vectors $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$, $\dfrac{\mathbf{a}_1 + \mathbf{a}_2 + \mathbf{a}_3}{3}$ is the average of these vectors. The quantity $\dfrac{\mathbf{a}_1 + 2\mathbf{a}_2 + 3\mathbf{a}_3}{6} = \dfrac{\mathbf{a}_1 + \mathbf{a}_2 + \mathbf{a}_2 + \mathbf{a}_3 + \mathbf{a}_3 + \mathbf{a}_3}{6}$ is the weighted average with weights $1/6$, $2/6$, and $3/6$. In general, for a set of vectors $\{\mathbf{a}_i\}_{i=1}^k$ and numbers $\{t_i\}_{i=1}^k$ with $t_i \geq 0$ and $\sum_{i=1}^k t_i = 1$, $\sum_{i=1}^k t_i \mathbf{a}_i$ is a weighted average, and is called a *convex combination* of the points determined by these vectors.

Definition. A set $\mathbf{S} \subset \mathbb{R}^n$ is *convex* provided that if $\mathbf{x}_0$ and $\mathbf{x}_1$ are any two points in $\mathbf{S}$ then the convex combination $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$ is also in $\mathbf{S}$ for all $0 \leq t \leq 1$.

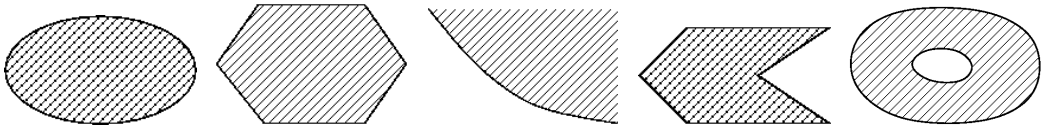


Figure 1.6.1. Examples of convex and non-convex sets

Each resource or requirement constraint, $a_{i1}x_1 + \cdots + a_{in}x_n \leq b_i$ or $\geq b_i$, defines a closed half-space. Each equality constraint, $a_{i1}x_1 + \cdots + a_{in}x_n = b_i$, is a hyperplane. Together, all the constraints define the intersection of such subsets of $\mathbb{R}^n$.

Definition. Any intersection of a finite number of closed half-spaces and possibly some hyperplanes is called a *polyhedron*.

Theorem 1.19. a. *The intersection of convex sets is convex.*

b. *Any polyhedron and so any feasible set for a linear programming problem is convex.*

Proof. (**a**) Assume $\{\mathbf{S}_j\}$ are a collection of convex sets and $\mathbf{x}_0, \mathbf{x}_1 \in \mathbf{S}_j$ for all j. Take $0 \leq t \leq 1$. Then

$$(1-t)\,\mathbf{x}_0 + t\,\mathbf{x}_1 \in \mathbf{S}_j \quad \text{for all } j, \text{ so}$$
$$(1-t)\,\mathbf{x}_0 + t\,\mathbf{x}_1 \in \bigcap_j \mathbf{S}_j.$$

(**b**) Each closed half-space and hyperplane is convex so the intersection is convex. □

Theorem 1.20. *If $\mathbf{S}$ is a convex set, and $\mathbf{p}_i \in \mathbf{S}$ for $1 \leq i \leq k$, then any convex combination $\sum_{i=1}^{k} t_i\,\mathbf{p}_i \in \mathbf{S}$.*

Proof. The proof is by induction on the number of points. For $k = 2$, it follows from the definition of a convex set.

Assume the result is true for $k - 1 \geq 2$ points. If $t_k = 1$, then $\mathbf{p}_k \in \mathbf{S}$, so true. If $t_k < 1$, then $\sum_{i=1}^{k-1} t_i = 1 - t_k > 0$ and $\sum_{i=1}^{k-1} \frac{t_i}{1-t_k} = 1$. The sum $\sum_{i=1}^{k-1} \frac{t_i}{1-t_k}\,\mathbf{p}_i \in \mathbf{S}$ by induction hypothesis. So,

$$\sum_{i=1}^{k} t_i\,\mathbf{p}_i = (1-t_k)\sum_{i=1}^{k-1} \frac{t_i}{1-t_k}\,\mathbf{p}_i + t_k\,\mathbf{p}_k \in \mathbf{S}.$$

□

Definition. A point $\mathbf{p}$ in a nonempty convex set $\mathbf{S}$ is called an *extreme point* provided that whenever $\mathbf{p} = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$ for some $0 < t < 1$ with $\mathbf{x}_0$ and $\mathbf{x}_1$ in $\mathbf{S}$ then $\mathbf{p} = \mathbf{x}_0 = \mathbf{x}_1$. An extreme point for a polyhedral set is also called a *vertex*.

An extreme point of a set must be a boundary point. The disk $\mathbf{D} = \{\,\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\| \leq 1\,\}$ is convex and each point on its boundary circle is an extreme point.

For the rest of the section, we consider a standard linear program in slack-variable form with both slack and surplus variables included in $\mathbf{x}$. The feasible set is given by

$$\mathscr{F} = \{\,\mathbf{x} \in \mathbb{R}_+^{n+m} : \mathbf{A}\mathbf{x} = \mathbf{b}\,\}.$$

Theorem 1.21. *Assume $\mathbf{x} \in \mathscr{F} = \{\,\mathbf{x} \in \mathbb{R}_+^{n+m} : \mathbf{A}\mathbf{x} = \mathbf{b}\,\}$ is a feasible solution to a linear programming problem. Then $\mathbf{x}$ is a point of $\mathscr{F}$ if and only if $\mathbf{x}$ is a basic feasible solution, i.e., if and only if the columns of $\mathbf{A}$ with $x_j > 0$ form a linearly independent set of vectors.*

Proof. By reindexing the columns and variables, we can assume that

$$x_1 > 0, \ldots, x_r > 0 \;\; x_{r+1} = \cdots = x_{n+m} = 0.$$

($\Rightarrow$) Assume that the columns $\{ \mathbf{A}_1, \ldots, \mathbf{A}_r \}$ are linearly dependent, so there are constants $\beta_1, \ldots, \beta_r$ not all zero with

$$\beta_1 \mathbf{A}_1 + \cdots + \beta_r \mathbf{A}_r = \mathbf{0}.$$

If we let $\boldsymbol{\beta}^{\mathsf{T}} = (\beta_1, \ldots, \beta_r, 0, \ldots, 0)$, then $\mathbf{A}\beta = \mathbf{0}$. For all small λ, $\mathbf{w}_1 = \mathbf{x} + \lambda\boldsymbol{\beta} \geq \mathbf{0}$ and $\mathbf{w}_2 = \mathbf{x} - \lambda\boldsymbol{\beta} \geq \mathbf{0}$. Also for $i = 1, 2$, $\mathbf{A}\mathbf{w}_i = \mathbf{A}\mathbf{x} = \mathbf{b}$, so both $\mathbf{w}_1, \mathbf{w}_2 \in \mathscr{F}$ are feasible solutions. Since $\mathbf{x} = \frac{1}{2}\mathbf{w}_1 + \frac{1}{2}\mathbf{w}_2$, $\mathbf{x}$ is not a vertex.

($\Leftarrow$) Conversely, assume that the feasible point $\mathbf{x}$ is not a vertex but a convex combination of the feasible solutions $\mathbf{y}$ and $\mathbf{z}$,

$$\mathbf{x} = t\,\mathbf{y} + (1-t)\,\mathbf{z} \qquad \text{for } 0 < t < 1,$$

with $\mathbf{y} \neq \mathbf{z}$. For $r < j$,

$$0 = x_j = t\,y_j + (1-t)\,z_j.$$

Since both $y_j \geq 0$ and $z_j \geq 0$, both must be zero for $j > r$. Because $\mathbf{y} \neq \mathbf{z}$ are both in $\mathscr{F}$,

$$\begin{aligned} \mathbf{b} = \mathbf{A}\mathbf{y} &= y_1 \mathbf{A}_1 + \cdots + y_r \mathbf{A}_r, \\ \mathbf{b} = \mathbf{A}\mathbf{z} &= z_1 \mathbf{A}_1 + \cdots + z_r \mathbf{A}_r, \qquad \text{and} \\ \mathbf{0} &= (y_1 - z_1) \mathbf{A}_1 + \cdots + (y_r - z_r) \mathbf{A}_r, \end{aligned}$$

and the columns $\{ \mathbf{A}_1, \ldots, \mathbf{A}_r \}$ must be linearly dependent. □

Note that for any linear (or convex) combination the value of the objective function is the corresponding linear combination,

$$f\left(\sum t_j \mathbf{x}_j\right) = \mathbf{c} \cdot \sum t_j \mathbf{x}_j = \sum t_j \mathbf{c} \cdot \mathbf{x}_j = \sum t_j f(\mathbf{x}_j).$$

Theorem 1.22. *Assume that the feasible set is nonempty for a bounded standard maximization linear program in slack-variable form. Then the following hold.*

- **a.** *If $\mathbf{x}^0$ is a feasible solution to a bounded linear program, then there exists a basic feasible solution $\mathbf{x}^b$ such that $f(\mathbf{x}^b) = \mathbf{c} \cdot \mathbf{x}^b \geq \mathbf{c} \cdot \mathbf{x}^0 = f(\mathbf{x}^0)$.*
- **b.** *There is at least one optimal basic solution.*
- **c.** *If two or more basic solutions are optimal, then any convex combination of them is also an optimal solution.*

Proof. (**a**) If $\mathbf{x}^0$ is already a basic feasible solution, then we are done.

Otherwise, the columns $\mathbf{A}$ corresponding to $x_i^0 \neq 0$ are linearly dependent. Let $\mathbf{A}'$ be the matrix with only these columns. Since the columns of $\mathbf{A}'$ are linearly dependent, there is a nonzero vector $\mathbf{y}'$ such that $\mathbf{A}'\mathbf{y}' = \mathbf{0}$. Adding zeroes in the other entries, we obtain a nonzero vector $\mathbf{y}$ such that $\mathbf{A}\mathbf{y} = \mathbf{0}$. Since, $\mathbf{A}(-\mathbf{y}) = \mathbf{0}$, we can assume that $\mathbf{y}$ has $\mathbf{c} \cdot \mathbf{y} \geq 0$ by replacing $\mathbf{y}$ by $-\mathbf{y}$ if necessary. Also,

$$\mathbf{A}\,[\mathbf{x}^0 + t\,\mathbf{y}] = \mathbf{b},$$

so $\mathbf{x}^0 + t\,\mathbf{y}$ is a solution. Remember that by construction, if $x_j^0 = 0$, then $y_j = 0$, so if $y_j \neq 0$, then $x_j^0 > 0$.

Case 1. Assume that $\mathbf{c} \cdot \mathbf{y} > 0$ and some component $y_i < 0$. Then $x_i^0 > 0$ and $x_i^0 + t\,y_i = 0$ for $t_i = {}^{-x_i^0}\!/_{y_i} > 0$. As t increases from 0 to t_i, the objective function increases from $\mathbf{c} \cdot \mathbf{x}^0$ to $\mathbf{c} \cdot [\mathbf{x}^0 + t_i \mathbf{y}]$. If more than one $y_i < 0$, then we select the one with the smallest value of t_i. In this way, we have constructed a new feasible solution $\mathbf{x}^1$ with one more component of the vector zero, fewer components of $y_i < 0$, and a greater value of the objective function.

We can continue in this manner until either the columns are linearly independent or all the components of the $\mathbf{y}$ are nonnegative.

Case 2. If $\mathbf{c} \cdot \mathbf{y} > 0$ $\mathbf{y} \geq \mathbf{0}$, then $\mathbf{x} + t\,\mathbf{y}$ is a feasible solution for all $t > 0$, and

$$\mathbf{c} \cdot [\mathbf{x} + t\,\mathbf{y}] = \mathbf{c} \cdot \mathbf{x} + t\,\mathbf{c} \cdot \mathbf{y}$$

is arbitrarily large. Thus the linear program is unbounded and has no maximum, which is a contradiction.

Case 3. Assume $\mathbf{c} \cdot \mathbf{y} = 0$. Some $y_k \neq 0$. Considering $\mathbf{y}$ and $\texttt{-}\mathbf{y}$, can assume $y_k < 0$. Then there exists first $t_k > 0$ such that $x_k^0 + t_k y_k = 0$. The value of the objective function does not change. There are few positive components of new $\mathbf{x}^0$. Eventually we get the corresponding columns linearly independent, and we are at a basic solution as claimed in part (a).

(**b**) There are only a finite number of basic feasible solutions $\{\mathbf{p}_j\}_{j=1}^N$ since there is a finite number of equations. By part (a),

$$f(\mathbf{x}) \leq \max_{1 \leq j \leq N} f(\mathbf{p}_j) \qquad \text{for } \mathbf{x} \in \mathscr{F}.$$

Thus, a maximum can be found among these finite set of values of the basic solutions.

(**c**) Assume that $f(\mathbf{p}_{j_i}) = M$ is the maximum for some collection of basic feasible solutions $i = 1, \ldots, \ell$. Then any convex combination is also a maximizer:

$$\begin{aligned}
&\mathbf{A}\left(\textstyle\sum_{i=1}^{\ell} t_{j_i}\mathbf{p}_{j_i}\right) = \textstyle\sum_{i=1}^{\ell} t_{j_i}\mathbf{A}\mathbf{p}_{j_i} = \textstyle\sum_{i=1}^{\ell} t_{j_i}\mathbf{b} = \mathbf{b},\\
&\textstyle\sum_{i=1}^{\ell} t_{j_i}\mathbf{p}_{j_i} \geq \mathbf{0} \qquad \text{is feasible,}\\
&f\left(\textstyle\sum_{i=1}^{\ell} t_{j_i}\mathbf{p}_{j_i}\right) = \textstyle\sum_{i=1}^{\ell} t_{j_i} f\left(\mathbf{p}_{j_i}\right) = \textstyle\sum_{i=1}^{\ell} t_{j_i} M = M.
\end{aligned}$$

□

If there are degenerate basic feasible solutions with fewer than m nonzero basic variables, then the simplex method can cycle by row reduction to matrices with the same nonzero basic variables but different zero basic variables so different sets of pivots. Interchanging a basic variable equal to zero for a zero non-basic variable corresponds to the same vertex of the feasible set. See the following example. There are ways to programs computers to avoid repeating a set of basic variables, so avoid cycling. See Jongen et al [**8**]. Humans avoid cycling naturally.

Example 1.23. The following maximization problem has a degenerate basic solution.

$$\begin{aligned}
\text{Maximize:}\quad & 8\,x_1 + 7\,x_2 + 2\,x_3\\
\text{Subject to:}\quad & 2\,x_1 + x_2 + x_3 \leq 15,\\
& 14\,x_1 + 13\,x_2 - 2\,x_3 \leq 105,\\
& 2\,x_1 + 4\,x_2 + 4\,x_3 \leq 30,\\
& x_1 \geq 0, \text{ and } x_2 \geq 0.
\end{aligned}$$

The steps in the simplex method are as follows.

$$\left[\begin{array}{cccccc|c} x_1 & x_2 & x_3 & s_1 & s_2 & s_3 & \\ \hline 2 & 1 & 1 & 1 & 0 & 0 & 15 \\ 14 & 13 & -2 & 0 & 1 & 0 & 105 \\ 2 & 4 & 4 & 0 & 0 & 1 & 30 \\ \hline -8 & -7 & -2 & 0 & 0 & 0 & 0 \end{array}\right] \sim \left[\begin{array}{cccccc|c} x_1 & x_2 & x_3 & s_1 & s_2 & s_3 & \\ \hline 2 & 1 & 1 & 1 & 0 & 0 & 15 \\ 0 & 6 & -9 & -7 & 1 & 0 & 0 \\ 0 & 3 & 3 & -1 & 0 & 1 & 15 \\ \hline 0 & -3 & 2 & 4 & 0 & 0 & 60 \end{array}\right]$$

$$\sim \left[\begin{array}{cccccc|c} x_1 & x_2 & x_3 & s_1 & s_2 & s_3 & \\ \hline 2 & 0 & \frac{5}{2} & \frac{13}{6} & -\frac{1}{6} & 0 & 15 \\ 0 & 2 & -3 & -\frac{7}{3} & \frac{1}{3} & 0 & 0 \\ 0 & 0 & \frac{15}{2} & \frac{5}{2} & -\frac{1}{2} & 1 & 15 \\ \hline 0 & 0 & -\frac{5}{2} & \frac{1}{2} & \frac{1}{2} & 0 & 60 \end{array}\right] \sim \left[\begin{array}{cccccc|c} x_1 & x_2 & x_3 & s_1 & s_2 & s_3 & \\ \hline 2 & 0 & 0 & \frac{4}{3} & 0 & -\frac{1}{3} & 10 \\ 0 & 2 & 0 & -\frac{4}{3} & \frac{2}{15} & \frac{2}{5} & 6 \\ 0 & 0 & 1 & \frac{1}{3} & -\frac{1}{15} & \frac{2}{15} & 2 \\ \hline 0 & 0 & 0 & \frac{4}{3} & \frac{1}{3} & \frac{1}{3} & 65 \end{array}\right].$$

Notice for the first pivoting, the ratio for the first and second row is the same which causes an entry in the augmented column to become zero in the second tableau. Both the second and third tableau have a zero basic variable in addition to the free (non-pivot) variables and have the same degenerate basic solution, $(x_1, x_2, x_3, s_1, s_2, s_3) = (15/2, 0, 0, 0, 0, 15)$, but the basic variables are different: (x_1, s_2, s_3) are basic variables for the second tableau and (x_1, x_2, s_3) are basic variables for the third tableau. When leaving a basic solution, the variable which becomes positive must be a free variable (non-basic variable) and not a zero basic (pivot) variable. The first pivot operation at this degenerate solution interchanges a basic variable equal to 0 and a free variable, so this new free variable made positive with the next pivoting when the value of the objective function is increased; this first pivoting results in all the same values of the variables, so the same point in $\mathscr{F}$. At a degenerate solution, the value will increase after a finite number of pivoting steps unless there is a cycle of pivoting sets (all staying at the same point of the feasible set) that results back with with the original set of pivots. The difficulty of a degenerate vertex and the potential of cycling is a matter of how row reduction relates to a movement on the feasible set. ■

Theorem 1.24. *If a maximal solution exists for a linear programming problem and the simplex algorithm does not cycle among degenerate basic feasible solutions, then the simplex algorithm locates a maximal solution in finitely many steps.*

Proof. Pivoting corresponds to changing from one basic feasible solution to another as in Theorem 1.22(a). Assume that never reach a degenerate basic feasible solution during the steps of the simplex method. Since there are only a finite number of vertices, the process must terminate with a basic feasible solution $\mathbf{p}_0$ for which pivoting to any of the nearby $\mathbf{p}_1,\ldots,\mathbf{p}_k$ has $f(\mathbf{p}_j) \leq f(\mathbf{p}_0)$. (Usually, it will strictly decrease the value.) Complete to the set of all basic feasible solutions (vertices) $\mathbf{p}_0, \mathbf{p}_1,\ldots, \mathbf{p}_\ell$.

The set of all convex combinations is a bounded polyhedron

$$\mathbf{H} = \left\{ \sum_{i=0}^{\ell} t_i \mathbf{p}_i : t_i \geq 0,\ \sum_{i=0}^{\ell} t_i = 1 \right\} \subset \mathscr{F}.$$

An edge of $\mathbf{H}$ from $\mathbf{p}$ to $\mathbf{p}j$ corresponds to pivoting as in the proof of Theorem 1.22(a) where one constraint becomes not equal to b_i and another becomes equal b_j. Positive cone out from $\mathbf{p}_0$ determined by $\{\, \mathbf{p}_i - \mathbf{p}_0 \,\}_{i=1}^{k}$

$$\mathbf{C} = \{\mathbf{p}_0 + \sum_{i=1}^{k} y_i \,(\mathbf{p}_i - \mathbf{p}_0) : y_i \geq 0 \,\} \supset \mathbf{H}.$$

Let $\mathbf{q}$ be any vertex of $\mathbf{H}$ (basic solution), $\mathbf{q} \in \mathbf{H} \subset \mathbf{C}$.

$$\mathbf{q} - \mathbf{p}_0 = \sum_{i=1}^{k} y_i \, (\mathbf{p}_i - \mathbf{p}_0) \quad \text{with all } \; y_i \geq 0.$$

Then,

$$f(\mathbf{q}) - f(\mathbf{p}) = \sum_{i=1}^{k} y_i \, [f(\mathbf{p}_i) - f(\mathbf{p}^*)] \leq 0.$$

This proves that $\mathbf{p}$ is a maximizer for f. □

1.6. Exercises

1.6.1. This exercise corresponds to Case 1 in the proof of Theorem 1.22(a).
Consider the linear program

$$\begin{aligned} \text{Maximize:} \quad & f(x_1, x_2, x_3) = 9\,x_1 + 2\,x_2 + x_3, \\ \text{Subject to:} \quad & 4\,x_1 + 5\,x_2 + 7\,x_3 + s_1 = 20, \\ & x_1 + 3\,x_2 + 2\,x_3 + s_2 = 7, \\ & x_1 \geq 0, \; x_2 \geq 0, \; x_3 \geq 0. \end{aligned}$$

Consider the vectors $\mathbf{x}^0 = (1, 2, 0, 6, 0)$ and $\mathbf{y} = (-3, 1, 0, 7, 0)$ with coordinates $(x_1, x_2, x_3, s_1, s_2)$.

- **a.** Show that $\mathbf{x}^0$ is a nonnegative solution of the linear program. Is it a basic solution? Why or why not?
- **b.** Show that $\mathbf{y}$ is a solution of the corresponding homogeneous equation.
- **c.** Determine a value of t such that $\mathbf{x}^0 + t\,\mathbf{y}$ is a basic solution with $f(\mathbf{a} + t\,\mathbf{b}) \geq f(\mathbf{a})$.

1.6.2. This exercise corresponds to Case 1 in the proof of Theorem 1.22(a).
Consider the linear program

$$\begin{aligned} \text{Maximize:} \quad & f(x_1, x_2, x_3) = 9\,x_1 + 2\,x_2 + x_3, \\ \text{Subject to:} \quad & x_1 + 3\,x_2 + 7\,x_3 + s_1 = 9, \\ & 2\,x_1 + x_2 + 5\,x_3 + s_2 = 12, \\ & x_1 \geq 0, \; x_2 \geq 0, \; x_3 \geq 0. \end{aligned}$$

- **a.** Show that $(x_1, x_2, x_3, s_1, s_2) = (2, 0, 1, 0, 3)$ is a non-basic nonzero solution of the linear program.
- **b.** Let $\mathbf{x}^0 = (2, 0, 1, 0, 3)$ as in part (a). Find a vector $\mathbf{y}$ that is a solution of the homogeneous equations and has only nonzero components in the same coordinates as $\mathbf{x}^0$.
- **c.** Determine a value of t such that $\mathbf{x}^0 + t\,\mathbf{y}$ is a basic solution with $f(\mathbf{a} + t\,\mathbf{b}) \geq f(\mathbf{a})$.

1.6.3. This exercise gives an unbounded problems and corresponds to Case 2 in the proof of Theorem 1.22(a).
Consider the maximization with the nonhomogeneous system of equations

$$\begin{aligned} \text{Maximize:} \quad & f(x_1, x_2, x_3) = 2\,x_1 + 5\,x_2 + 3\,x_3, \\ \text{Subject to:} \quad & 6\,x_1 - x_2 + 5\,x_3 + x_4 = 6, \\ & -4\,x_1 + x_2 + 3\,x_3 - x_5 + x_6 = 2, \\ & \mathbf{x} \geq \mathbf{0}. \end{aligned}$$

a. Set up the tableau (without adding any more artificial variables). Apply the simplex method. You should come to the situation where an entry in the objective function row is negative and all the entries above it are negative. You can use this to give a feasible nonnegative solution.

b. Take the last tableau obtained in part (a) and write out the general solution taking all the free variables to the right hand size. Explain why the variable with the negative entry in the objective function row can be increased arbitrarily large and still give feasible nonnegative solutions. Why does this show that the problem is unbounded?

1.6.4. Consider the maximization with the nonhomogeneous system of equations

$$\begin{aligned} \text{Maximize:} \quad & f(x_1, x_2, x_3, x_4) = 75\,x_1 - 250\,x_2 + 50\,x_3 - 100\,x_4, \\ \text{Subject to:} \quad & x_1 - 4\,x_2 - 4\,x_3 + 6\,x_4 \leq 1, \\ & x_1 - 3\,x_2 - x_3 + x_4 \leq 1, \\ & x_3 \leq 1, \\ & x_1 \geq 0,\, x_2 \geq 0,\, x_3 \geq 0,\, x_4 \geq 0. \end{aligned}$$

a. Row reduce the tableau with slack variables added by choosing the following sequence of new pivots: (i) row 1 and column 1, (ii) row 2 and column 3, then (iii) row three and column 5.

b. For the second and third tableau answer the following: (i) What are the basic variables? (ii) What is the basic feasible solution? (iii) Why are these degenerate basic feasible solutions?

1. Exercises for Chapter 1

1.1. Indicated which of the following statements are *true* and which are *false*. Justify each answer: For a true statement explain why it is true and for a false statement either indicate how to make it true or indicate why the statement is false. The statements relate to a standard maximum linear program with the objective function $f(\mathbf{x}) = \mathbf{c}\cdot\mathbf{x}$ the constraint inequality $\mathbf{Ax} \leq \mathbf{b}$ for an $m \times n$ coefficient matrix $\mathbf{A}$ and constant vector $\mathbf{b} \in \mathbb{R}^m_+$, and $\mathbf{x} \geq \mathbf{0}$.

a. If a standard maximization linear program does not have an optimal solution, then either the objective function is unbounded on the feasible set $\mathscr{F}$ or $\mathscr{F}$ is the empty set.

b. If $\bar{\mathbf{x}}$ is an optimal solution of a standard maximization linear program, then $\bar{\mathbf{x}}$ is an extreme point of the feasible set.

c. A slack variable is used to change an equality into an inequality.

d. A solution is called a basic solution if exactly m of the variables are nonzero.

e. The basic feasible solutions correspond to the extreme points of the feasible region.

f. The bottom entry in the right column of a simplex tableau gives the maximal value of the objective function.

g. For a tableau for a maximization linear program, if there is a column with all negative entries including the one in the row for the objective function, then the linear programming problem has no feasible solution.

h. The value of the objective function for a MLP at any basic feasible solution is always greater than the value at any non-basic feasible solution.

i. If a standard maximization linear program MLP has nonempty feasible set, then it has an optimal basic solution.

j. In the two-phase simplex method, if an artificial variable is positive for the optimal solution for the artificial objective function, then there is no feasible solution to the original linear program.

k. The dual mLP problem is to minimize $\mathbf{y} \in R^m$ subject to $\mathbf{Ay} \geq \mathbf{c}$ and $\mathbf{y} \geq \mathbf{0}$.

l. If $\bar{\mathbf{x}}$ is an optimal solution to the primal MLP and $\hat{\mathbf{y}}$ is a feasible solution to the dual mLP, then $f(\bar{\mathbf{x}}) = g(\hat{\mathbf{y}})$.

m. If a slack variable is $\bar{s}_i > 0$ in an optimal solution, then the addition to the objective function that would be realized by one more unit of the resource corresponding to its inequality is positive.

n. If a maximization linear program MLP and its dual minimization linear problem mLP each have nonempty feasible sets (some feasible point), then each problem has an optimal solution.

o. If the optimal solution of a standard MLP has a slack variable $s_i = 0$, then the i^{th} resource has zero marginal value, i.e., one unit of the i^{th} resource would add nothing to the value of the objective function.

Chapter 2

Unconstrained Extrema

We begin our treatment of nonlinear optimization problems in this chapter with consideration of unconstrained problems where the variables are free to move in any direction in a Euclidean space. The first section presents the mathematical background from calculus and linear algebra and summarizes some terminology and results from real analysis that is used in the rest of the book. The second section gives the standard first and derivative conditions for an extremizer. A standard multi-dimensional calculus course certainly discusses the unconstrained extrema of functions of two variables and three variables and some calculus courses treat most of the material of this chapter. However, we present a unified treatment in any dimension that forms the foundation for the remaining chapters.

2.1. Mathematical Background

In this section, we summarize some terminology and results from an undergraduate course on real analysis so we can use it in the rest of the book. See Wade [**15**] or Rudin [**10**] for more details. This material includes open, closed, and bounded sets and their boundaries. The concept of a continuous functions is introduced, because it is a crucial assumption in many of the results, for example the Extreme Value Theorem on the existence of a maximum or minimum.

Although many of the results can be stated in terms of the gradients of real-valued functions, we consider the derivative of functions between Euclidean spaces as a matrix. This perspective is very useful for the discussion in the following chapter about conditions on constraints and implicit differentiation in the context of several constraints. Colley [**6**] has much of this material on differentiation of multivariable functions. From linear algebra, some material about quadratic forms is reviewed and extended. A basic reference is Lay [**9**]. We present a more practical tests for local extrema than is usually given.

2.1.1. Types of Subsets of $\mathbb{R}^n$

In this subsection, we define the basic types of subsets differently than most books on real analysis. The conditions we give are used because they seem intuitive are usually proved to be equivalent of the usual ones for sets in Euclidean spaces.

Definition. For $\mathbf{p} \in \mathbb{R}^n$ and $r > 0$, the set

$$\mathbf{B}(\mathbf{p}, r) = \{\, \mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{p}\| < r \,\}$$

is called the *open ball about* $\mathbf{p}$ *of radius* r. The set

$$\overline{\mathbf{B}}(\mathbf{p}, r) = \{\, \mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{p}\| \leq r \,\}$$

is called a *closed ball about* $\mathbf{p}$ *of radius* r.

Definition. The *complement* of a set $\mathbf{S}$ in $\mathbb{R}^n$ are the points not in $\mathbf{S}$, $\mathbf{S}^c = \mathbb{R}^n \setminus \mathbf{S}$ $= \{\, \mathbf{x} \in \mathbb{R}^n : \mathbf{x} \notin \mathbf{S} \,\}$.

Definition. The *boundary* of a set $\mathbf{S} \subset \mathbb{R}^n$, denoted by $\partial(\mathbf{S})$, is the set of all points which have points arbitrarily close in both $\mathbf{S}$ and $\mathbf{S}^c$,

$$\partial(\mathbf{S}) = \{\, \mathbf{x} \in \mathbb{R}^n : \mathbf{B}(\mathbf{x}, r) \cap \mathbf{S} \neq \emptyset \text{ and } \mathbf{B}(\mathbf{x}, r) \cap \mathbf{S}^c \neq \emptyset \text{ for all } r > 0 \,\}.$$

Example 2.1. The boundary of an open or a closed ball is the same,

$$\partial(\mathbf{B}(\mathbf{p}, r)) = \partial(\overline{\mathbf{B}}(\mathbf{p}, r)) = \{\, \mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{p}\| = r \,\}.$$ ■

Example 2.2. The boundary of the bounded polyhedral set

$$\begin{aligned} 2 &\geq x_1 + x_2, \\ 16 &\geq 5x_1 + 10x_2 \\ 3 &\geq 2x_1 + x_2, \\ 0 &\leq x_1, \text{ and } 0 \leq x_2, \end{aligned}$$

is the polygonal closed curve made up of five line segments. See Figure 1.2.1. ■

For most of the sets we consider, the boundary will be the curve or surface that encloses the set.

Definition. A set $\mathbf{S} \subset \mathbb{R}^n$ is *closed* provided that its boundary is a contained in $\mathbf{S}$, $\partial(\mathbf{S}) \subset \mathbf{S}$, i.e., if $\mathbf{B}(\mathbf{p}, r) \cap \mathbf{S} \neq \emptyset$ for all $r > 0$, then $\mathbf{p} \in \mathbf{S}$.

Definition. A set $\mathbf{S} \subset \mathbb{R}^n$ is said to be *open* provided that no point of the boundary is an element of $\mathbf{S}$, $\mathbf{S} \cap \partial(\mathbf{S}) = \emptyset$. This is the same as saying that for any point $\mathbf{x}_0 \in \mathbf{S}$ there exists an $r > 0$ such that $\mathbf{B}(\mathbf{x}_0, r) \subset \mathbf{S}$. The intuitive idea of an open set is that for each point in the set, all the nearby points are also in the set.

Since $\partial(\mathbf{S}) = \partial(\mathbf{S}^c)$, it follows that a set $\mathbf{S}$ is closed iff $\mathbf{S}^c$ is open.

Example 2.3. In the real line $\mathbb{R}$, the intervals (a, b), (a, ∞), and $(-\infty, b)$ are open and the intervals $[a, b]$, $[a, \infty)$, and $(-\infty, b]$ are closed. The intervals $[a, b)$ and $(a, b]$ are neither open nor closed. The boundary of the whole line $(-\infty, \infty)$ is empty, and $(-\infty, \infty)$ is both open and closed. ■

Example 2.4. In $\mathbb{R}^n$, the whole space $\mathbb{R}^n$ and the empty set $\emptyset$ are both open and closed. ■

Example 2.5. In $\mathbb{R}^n$, since $\partial(\mathbf{B}(\mathbf{p}, r)) \cap \mathbf{B}(\mathbf{p}, r) = \emptyset$, the open ball $\mathbf{B}(\mathbf{p}, r)$ is open and its complement $\mathbf{B}(\mathbf{a}, r)^c$ is closed.

Alternatively, let $\mathbf{x}_0 \in \mathbf{B}(\mathbf{p}, r)$. Set $r' = r - \|\mathbf{x}_0 - \mathbf{p}\| > 0$. We claim that $\mathbf{B}(\mathbf{x}_0, r') \subset \mathbf{B}(\mathbf{p}, r)$: Take $\mathbf{x} \in \mathbf{B}(\mathbf{x}_0, r')$. Then

$$\|\mathbf{x} - \mathbf{p}\| \leq \|\mathbf{x} - \mathbf{x}_0\| + \|\mathbf{x}_0 - \mathbf{p}\| < r' + \|\mathbf{x}_0 - \mathbf{p}\| = r,$$

and $\mathbf{x} \in \mathbf{B}(\mathbf{p}, r)$. This shows that $\mathbf{B}(\mathbf{p}, r)$ is open. ■

Example 2.6. In $\mathbb{R}^n$, since $\partial(\overline{\mathbf{B}}(\mathbf{p}, r)) \subset \overline{\mathbf{B}}(\mathbf{p}, r)$, the closed ball $\overline{\mathbf{B}}(\mathbf{p}, r)$ is closed and its complement $\overline{\mathbf{B}}(\mathbf{p}, r)^c$ is open.

Alternatively, take $\mathbf{x}_0 \in \overline{\mathbf{B}}(\mathbf{p}, r)^c$. It follows that $\|\mathbf{x}_0 - \mathbf{p}\| > r$. Let $r' = \|\mathbf{x}_0 - \mathbf{p}\| - r > 0$. If $\mathbf{x} \in \mathbf{B}(\mathbf{x}_0, r')$, then

$$\|\mathbf{x} - \mathbf{p}\| \geq \|\mathbf{p} - \mathbf{x}_0\| - \|\mathbf{x}_0 - \mathbf{x}\| > \|\mathbf{x}_0 - \mathbf{p}\| - r' = r$$

and $\mathbf{x} \in \mathbf{B}(\mathbf{p}, r)^c$. Thus, $\mathbf{B}(\mathbf{x}_0, r') \subset \overline{\mathbf{B}}(\mathbf{p}, r)^c$ and $\overline{\mathbf{B}}(\mathbf{p}, r)^c$ is open and $\overline{\mathbf{B}}(\mathbf{p}, r)$ is closed. ■

Definition. The *interior* of $\mathbf{S} \subset \mathbb{R}^n$, denoted by $\operatorname{int}(\mathbf{S})$, is the set with its boundary removed, $\operatorname{int}(\mathbf{S}) = \mathbf{S} \setminus \partial(\mathbf{S})$. It is the largest open set contained in $\mathbf{S}$. It is also the set of all points $\mathbf{p} \in \mathbf{S}$ for which there exists an $r > 0$ such that $\mathbf{B}(\mathbf{p}, r) \subset \mathbf{S}$.

The *closure* of $\mathbf{S} \subset \mathbb{R}^n$, denoted by $\operatorname{cl}(\mathbf{S})$ or $\overline{\mathbf{S}}$, is the union of $\mathbf{S}$ and its boundary $\partial(\mathbf{S})$, $\operatorname{cl}(\mathbf{S}) = \mathbf{S} \cup \partial(\mathbf{S})$. It is the smallest closed set containing $\mathbf{S}$.

Notice that the boundary of a set equals its closure minus its interior, $\partial(\mathbf{S}) = \operatorname{cl}(\mathbf{S}) \setminus \operatorname{int}(\mathbf{S})$.

Example 2.7. For intervals in $\mathbb{R}$, $\operatorname{int}([0,1]) = (0,1)$, $\operatorname{cl}((0,1)) = [0,1]$, and $\partial([0,1]) = \partial((0,1)) = \{0,1\}$. Also, $\operatorname{cl}(\mathbb{Q} \cap (0,1)) = [0,1]$, $\operatorname{int}(\mathbb{Q} \cap (0,1)) = \emptyset$, and $\partial(\mathbb{Q} \cap (0,1)) = [0,1]$.

The analogous object in $\mathbb{R}^n$ to an interval in $\mathbb{R}$ is an open or closed ball. For these sets, $\operatorname{int} \overline{\mathbf{B}}(\mathbf{a}, r) = \mathbf{B}(\mathbf{a}, r)$ and $\operatorname{cl}(\mathbf{B}(\mathbf{a}, r) = \overline{\mathbf{B}}(\mathbf{a}, r)$. The boundary is the same for both of these balls, $\partial\overline{\mathbf{B}}(\mathbf{a}, r) = \partial\mathbf{B}(\mathbf{a}, r) = \{\, \mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{a}\| = r \,\}$. ■

To guarantee that a maximum of a real valued function exists, the domain or feasible set must be closed and bounded: the domain must contain its boundary and cannot "go off to infinity".

Definition. A set $\mathbf{S} \subset \mathbb{R}^n$ is *bounded* provided that there exists a sufficiently large $r > 0$ such that $\mathbf{S} \subset \overline{\mathbf{B}}(\mathbf{0}, r)$, i.e., $\|\mathbf{x}\| \leq r$ for all $\mathbf{x} \in \mathbf{S}$.

Definition. A set $\mathbf{S} \subset \mathbb{R}^n$ is called *compact* provided that it is closed and bounded.

Remark. In a course in real analysis, a compact set is defined differently: either in terms of sequences or covers of the set by open sets. With one of these definitions, the above definition is a theorem about a compact set in $\mathbb{R}^n$.

2.1.2. Continuous Functions

In a calculus course, continuity is usually only mentioned in passing. Since it plays a key role in a more rigorous discussion of optimization, we give a brief introduction to this concept.

Example 2.8. For a function of a single real variable, the intuitive definition of a continuous function is that its graph can be drawn without lifting the pen. There are various ways in which a function can be discontinuous at a point. Consider the following two functions:

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0, \end{cases} \quad \text{and} \quad g(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \sin\left(1/x\right) & \text{if } x > 0, \end{cases}$$

The function f has a *jump* at $x = 0$ so is discontinuous. The function g *oscillates* as x approaches zero and is discontinuous at $x = 0$. Notice that for the function g, there are some points $x > 0$ where the value is near (equal to) $g(0) = 0$ but there are other points near to 0 where the values is far from $g(0) = 0$. ■

The definition of continuity is given in terms of limits, which we define first.

Definition. Let $f : \mathbf{S} \subset \mathbb{R}^n \to \mathbb{R}^m$.

If $\mathbf{a} \in \text{cl}(\mathbf{S})$, then the *limit of* $f(\mathbf{x})$ *at* $\mathbf{a}$ *is* $\mathbf{L}$, which is written as $\lim_{\mathbf{x}\to\mathbf{a}} f(\mathbf{x}) = \mathbf{L}$, provided that for every $\epsilon > 0$ there exists a $\delta > 0$ such that $\|f(\mathbf{x}) - \mathbf{L}\| < \epsilon$ whenever $\|\mathbf{x} - \mathbf{a}\| < \delta$ and $\mathbf{x} \in \mathbf{S} \setminus \{\mathbf{a}\}$.

This definition can be extended to the limit as a real variable goes to infinity: $\lim_{x\to\infty} f(x) = L$ provided that for every $\epsilon > 0$, there exists a K such that $|f(x)-L| < \epsilon$ whenever $x \geq K$.

Example 2.9. In $\mathbb{R}^2$, consider

$$f(x,y) = \begin{cases} \dfrac{y^2}{x^2+y^2} & \text{if } (x,y) \neq (0,0) \\ 0 & \text{if } (x,y) = (0,0). \end{cases}$$

The function f does not have a limit as (x, y) goes to $(0, 0)$ since it approaches different values as (x, y) approaches the origin from different directions:

$$\lim_{y\to 0} f(0,y) = \lim_{y\to 0} \frac{y^2}{y^2} = 1 \qquad \text{and}$$

$$\lim_{x\to 0} f(x,tx) = \lim_{x\to 0} \frac{t^2x^2}{x^2+t^2x^2} = \frac{t^2}{1+t^2} \neq 1.$$

■

Definition. A function $f : \mathbf{S} \subset \mathbb{R}^n \to \mathbb{R}^m$ is *continuous at* $\mathbf{a} \in \mathbf{S}$ provided that $\lim_{\mathbf{x}\to\mathbf{a}} f(\mathbf{x}) = f(\mathbf{a})$, i.e., for all $\epsilon > 0$ there exists a $\delta > 0$ such that $\|f(\mathbf{x}) - f(\mathbf{a})\| < \epsilon$ whenever $\|\mathbf{x} - \mathbf{a}\| < \delta$ and $\mathbf{x} \in \mathbf{S}$.

We say that f is *continuous on a set* $\mathbf{S}$ provided that it is continuous at all points in $\mathbf{S}$.

Continuity at $\mathbf{a}$ means that given a tolerance $\epsilon > 0$ in the values, there is a tolerance $\delta > 0$ in the input such that *all* points within δ of $\mathbf{a}$ have values within ϵ of $f(\mathbf{a})$.

In terms of sequences, f is *continuous at* $\mathbf{a} \in \mathbf{S}$ provided that for *any* sequence of points $\mathbf{x}_k$ in $\mathbf{S}$ that converge to $\mathbf{a}$, the values $f(\mathbf{x}_k)$ converge to $f(\mathbf{a})$.

Example 2.10. We return to the function $g(x) = \sin(1/x)$ for $x > 0$ and $g(x) = 0$ for $x < 0$. For any small $\delta > 0$, there are *some* points $x_n = 1/(n2\pi) > 0$ for which $g(x_n) = 0 = g(0)$, but there are other points $x'_n = 2/(n4\pi + \pi)$ such that $g(x'_n) = 1$ that all have $|g(x'_n) - g(0)| = 1 > 1/2$. Thus, $g(x)$ is not continuous at $x = 0$.

Example 2.11. Let $f : (0,\infty) \to \mathbb{R}$ be defined by $f(x) = 1/x$. We claim that f is continuous at all points $x > 0$. Fix $a > 0$ and $\epsilon > 0$. If $|x - a| < \delta$, then

$$|f(x) - f(a)| = \left|\frac{1}{x} - \frac{1}{a}\right| = \frac{|a-x|}{|xa|} < \frac{\delta}{|xa|}.$$

If $\delta < a/2$, then $x = a - (a - x) > a/2$, $1/x < 2/a$, and

$$|f(x) - f(a)| < \frac{2\delta}{a^2}.$$

Thus, if δ is also less than $\epsilon a^2/2$, then $|f(x) - f(a)| < \epsilon$. Therefore, if we take $\delta < \min\left\{a/2, \epsilon\, a^2/2\right\}$, then we get that $|f(x) - f(a)| < \epsilon$ as desired. ■

It is not hard to show that a vector valued function $\mathbf{F} : \mathbf{S} \subset \mathbb{R}^n \to \mathbb{R}^m$ is continuous at $\mathbf{a}$ iff each of its coordinate functions F_i is continuous at $\mathbf{a}$.

Continuity can be characterized in terms of the inverse image of open or closed sets. We use this property a few times.

Definition. If $f : \mathscr{D} \subset \mathbb{R}^n \to \mathbb{R}^m$ is a function, then the *inverse image of* $\mathbf{U} \subset \mathbb{R}^m$ is

$$f^{-1}(\mathbf{U}) = \{\, \mathbf{x} \in \mathscr{D} : f(\mathbf{x}) \in \mathbf{U} \,\} \subset \mathscr{D}.$$

In this context, the function need not have an inverse, but $f^{-1}(\mathbf{U})$ merely denotes points that map into the set $\mathbf{U}$.

A *level set* of f is the same as the *inverse image of a point* $\mathbf{b} \in \mathbb{R}^m$,

$$\begin{aligned} f^{-1}(\mathbf{b}) &= \{\, \mathbf{x} \in \mathscr{D} : f(\mathbf{x}) = \mathbf{b} \,\} \\ &= \{\, \mathbf{x} \in \mathscr{D} : f_i(\mathbf{x}) = b_i \text{ for } i = 1, \ldots, m \,\} \subset \mathscr{D}. \end{aligned}$$

Theorem 2.12. *Let $f : \mathscr{D} \subset \mathbb{R}^n \to \mathbb{R}^m$. Then the following are equivalent.*

(i) *f is continuous on $\mathscr{D}$.*

(ii) *For each open set $\mathbf{V} \subset \mathbb{R}^m$, there is an open set $\mathbf{U} \subset \mathbb{R}^n$ such that $f^{-1}(\mathbf{V}) = \mathbf{U} \cap \mathscr{D}$, i.e., the inverse image of an open set $f^{-1}(\mathbf{V})$ is open relative to $\mathscr{D}$.*

(iii) *For each closed set $\mathbf{C} \subset \mathbb{R}^m$, there is an closed set $\mathbf{B} \subset \mathbb{R}^n$ such that $f^{-1}(\mathbf{C}) = \mathbf{B} \cap \mathscr{D}$, i.e., the inverse image of a closed set $f^{-1}(\mathbf{C})$ is closed relative to $\mathscr{D}$.*

For a proof, see Wade [**15**] or Rudin [**10**].

Example 2.13. Let $p_i > 0$ for $1 \le i \le n$ be fixed prices and $w > 0$ be the wealth. The *simplex*

$$\mathbf{S} = \{\, \mathbf{x} \in \mathbb{R}^n : x_i \ge 0 \text{ for } 1 \le i \le n,\ p_1 x_1 + \cdots + p_n x_n \le w \,\}$$

is compact.

For a point $\mathbf{x} \in \mathbf{S}$, each coordinate $0 \le x_j \le {}^{w}\!/_{p_j}$ so $\|\mathbf{x}\|_m \le \max_i\{{}^{w}\!/_{p_i}\}$, so the set is bounded.

Intuitively, the set is closed because the inequalities are non-strict, "less than or equal to" or "greater than or equal to".

More formally, the function $f(\mathbf{x}) = p_1 x_1 + \cdots + p_n x_n$ is easily seen to be continuous. The interval $[0, w]$ is closed in $\mathbb{R}$, so the set

$$\{\, \mathbf{x} \in \mathbb{R}^n : 0 \le f(\mathbf{x}) \le w \,\} = f^{-1}([0, w])$$

is closed. Similarly, for any $1 \le i \le n$, $g_i(\mathbf{x}) = x_i$ is continuous and the interval $[0, \infty)$ is closed, so the sets

$$\{\, \mathbf{x} \in \mathbb{R}^n : x_i \ge 0 \,\}$$

are closed. Combining,

$$\begin{aligned} \mathbf{S} &= \{\, \mathbf{x} \in \mathbb{R}^n : x_i \ge 0 \text{ for } 1 \le i \le n,\ p_1 x_1 + \cdots + p_n x_n \le w \,\} \\ &= f^{-1}([0, w]) \cap \bigcap_{i=1}^{n} g_i^{-1}([0, \infty)) \end{aligned}$$

is closed. Since $\mathbf{S}$ is both closed and bounded, it is compact. ■

2.1.3. Existence of Extrema

We can now state a general theorem on the existence of points that maximize and minimize. It does not give a constructive method for finding a point that maximizes a function but merely gives sufficient conditions for a maximum to exist. Such a result gives incentive to search for a maximizer. In this chapter and the next, we will consider techniques to help find the maximizer.

Theorem 2.14 (Extreme Value Theorem). *Assume that $\mathscr{F} \subset \mathbb{R}^n$ is a nonempty compact set. Assume that $f : \mathscr{F} \to \mathbb{R}$ is a continuous function. Then f attains a maximum and a minimum on $\mathscr{F}$, i.e., there exist points $\mathbf{x}_m, \mathbf{x}_M \in \mathscr{F}$ such that*

$$\begin{aligned} f(\mathbf{x}_m) &\leq f(\mathbf{x}) \leq f(\mathbf{x}_M) \qquad \textit{for all } \mathbf{x} \in \mathscr{F}, \ \textit{so} \\ f(\mathbf{x}_m) &= \min_{\mathbf{x}\in\mathscr{F}} f(\mathbf{x}), \qquad\qquad \textit{and} \\ f(\mathbf{x}_M) &= \max_{\mathbf{x}\in\mathscr{F}} f(\mathbf{x}). \end{aligned}$$

For a proof see a book on real analysis such as the ones by Wade [**15**] or Rudin [**10**].

Example 2.15. We give some examples that illustrate why it is necessary to assume that the domain of a function is compact in order to be certain that the function attains a maximal or minimal value.

On the unbounded set $\mathscr{F} = \mathbb{R}$, the function $f(x) = x^3$ is unbounded above and below and has no maximum nor minimum. The function is continuous, but the domain is not bounded.

The same function $f(x) = x^3$ on $(-1, 1)$ is bounded above and below, but it does not attain a maximum or minimum on $(-1, 1)$; the set is bounded but not closed.

Similarly, $g(x) = \tan(x)$ on the interval $(-\pi/2, \pi/2)$ is unbounded and does not have a minimum or maximal value; again, this interval is bounded but not closed.

The function $h(x) = \arctan(x)$ on $\mathbb{R}$ is bounded but the function does not attain a maximal or minimal value. In this case, the domain is closed but is not bounded and the set of values attained is bounded but not closed. ■

Example 2.16. Consider

$$f(x) = \begin{cases} \frac{1}{x} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

This function is not continuous at $x = 0$ and has no maximum nor minimum on $[-1, 1]$ even though the domain is compact. ■

In Section 4.2.1, we define the least upper bound or supremum for a function that is bounded above but does not attain its maximum.

2.1.4. Differentiation in Multi-Dimensions

In this section, we introduce the derivative of a vector-valued function as the matrix of partial derivatives. This approach generalizes the gradient of a scalar-valued function. We use this treatment of the derivative when considering extremizers later in the chapter and also for implicit differentiation with several constraining equations in the next chapter.

Before discussing differentiation of a function defined on $\mathbb{R}^n$, we derive a consequence of differentiation of a function $f : \mathbb{R} \to \mathbb{R}$. It has derivative $f'(p)$ at p provided that

$$\begin{aligned} \lim_{x\to p} \frac{f(x) - f(p)}{x - p} &= f'(p) \qquad \text{or} \\ \lim_{x\to p} \frac{f(x) - f(p) - f'(p)(x - p)}{x - p} &= 0. \end{aligned}$$

Thus, for small $\epsilon > 0$, there exists $\delta > 0$, such that for $|x - p| < \delta$,

$$\begin{aligned} -\epsilon\,|x - p| \leq f(x) - f(p) - f'(p)(x - p) \leq \epsilon\,|x - p| \quad \text{or} \\ f(p)+f'(p)(x - p) - \epsilon\,|x - p| \leq f(x) \leq f(p) + f'(p)(x - p) + \epsilon\,|(x - p|. \end{aligned}$$

These inequalities imply that for x near p, the value of the function $f(x)$ is in a narrow cone about the line $f(p) + f'(p)(x - p)$, and this line is a good affine approximation of the

nonlinear function $f(x)$ near p. See Figure 2.1.1. (An *affine* function is a constant plus a linear function.)

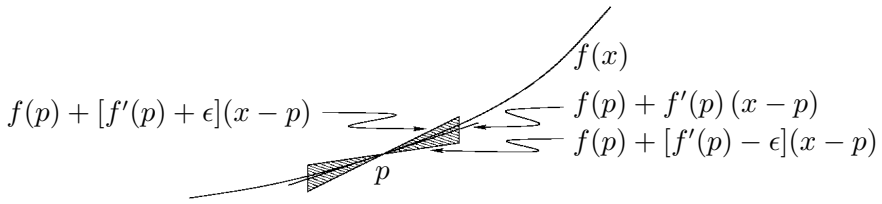


Figure 2.1.1. Cone condition for derivative at p

Definition. A function $f : \mathscr{D} \subset \mathbb{R}^n \to \mathbb{R}^m$ is said to be *differentiable* at $\mathbf{p} \in \text{int}(\mathscr{D})$ provided that all the first order partial derivatives exist at $\mathbf{p}$ and the $m \times n$ matrix

$$Df(\mathbf{p}) = \left(\frac{\partial f_i}{\partial x_j}(\mathbf{p})\right)$$

satisfies

$$\lim_{\mathbf{x}\to\mathbf{p}} \frac{f(\mathbf{x}) - f(\mathbf{p}) - Df(\mathbf{p})(\mathbf{x} - \mathbf{p})}{\|\mathbf{x} - \mathbf{p}\|} = \mathbf{0}, \quad \text{or}$$
$$f(\mathbf{x}) = f(\mathbf{p}) + Df(\mathbf{p})\,(\mathbf{x} - \mathbf{p}) + \widetilde{R}_1(\mathbf{p}, \mathbf{x})\,\|\mathbf{x} - \mathbf{p}\| \quad \text{where}$$
$$\lim_{\mathbf{x}\to\mathbf{p}} \widetilde{R}_1(\mathbf{p}, \mathbf{x}) = \mathbf{0}.$$

When this limit is satisfied, the the matrix $Df(\mathbf{p})$ is called the *derivative* of f at $\mathbf{p}$.

If $m = 1$, then $Df(\mathbf{p})^{\mathsf{T}} = \nabla f(\mathbf{p})$ is the *gradient*, which is a column vector. Then $Df(\mathbf{p})(\mathbf{x} - \mathbf{p}) = \nabla f(\mathbf{p}) \cdot (\mathbf{x} - \mathbf{p}) = \sum_j \dfrac{\partial f}{\partial x_j}(\mathbf{p})\,(x_j - p_j)$.

Remark. The fact that the remainder $R_1(\mathbf{p}, \mathbf{x}) = \widetilde{R}_1(\mathbf{p}, \mathbf{x})\,\|\mathbf{x} - \mathbf{p}\|$ goes to zero faster than $\|\mathbf{x} - \mathbf{p}\|$ means that $f(\mathbf{p}) + Df(\mathbf{p})(\mathbf{x} - \mathbf{p})$ is a good *affine approximation* of the nonlinear function $f(\mathbf{x})$ near $\mathbf{p}$ for all small displacements just as in the one dimensional case.

Since we cannot divide by a vector, the first limit in the definition of differentiability divides by the length of the displacement.

Note that the rows of the derivative matrix $Df(\mathbf{p})$ are determined by the coordinate functions and the columns by the variable of the partial derivative. This choice is important so that the correct terms are multiplied together in the matrix product $Df(\mathbf{p})(\mathbf{x} - \mathbf{p})$.

Definition. A function $f : \mathscr{D} \subset \mathbb{R}^n \to \mathbb{R}^m$ is said to be *continuously differentiable* or C^1 on $\text{int}(\mathscr{D})$ provided that all the first order partial derivatives exist and on continuous on $\text{int}(\mathscr{D})$.

The following theorem shows that a C^1 function is differentiable at all points in the interior of the domain.

Theorem 2.17. *If $f : \mathscr{D} \subset \mathbb{R}^n \to \mathbb{R}^m$ is C^1 on $\text{int}(\mathscr{D})$, then f is differentiable at all points $\mathbf{p} \in \text{int}(\mathscr{D})$.*

Proof. For $\mathbf{x}$ a point near $\mathbf{p}$, let $\mathbf{p}^j = (p_1, \ldots, p_j, x_{j+1}, \ldots, x_n)$ and $\mathbf{r}^j(t) = (1-t)\,\mathbf{p}^{j-1} + t\,\mathbf{p}^{j-1}$. The Mean Value Theorem for a function of one variable can be applied to each coordinate function f_i along the paths $\mathbf{r}^j(t)$:

$$\begin{aligned} f_i(\mathbf{p}^j) - f_i(\mathbf{p}^{j-1}) &= f_i(\mathbf{r}^j(1)) - f_i(\mathbf{r}^j(0)) = (1-0)\ \frac{df_i(\mathbf{r}^j(t))}{dt}\bigg|_{t=t_{ij}} \\ &= \frac{\partial f_i}{\partial x_j}(\mathbf{r}^j(t_{ij}))\,(x_j - p_j). \end{aligned}$$

If we add these up, the sum telescopes,

$$f_i(\mathbf{x}) - f_i(\mathbf{p}) = \sum_{j=1}^{n} f_i(\mathbf{p}^j) - f_i(\mathbf{p}^{j-1}) = \sum_{j=1}^{n} \frac{\partial f_i}{\partial x_j}(\mathbf{r}^j(t_{ij}))\,(x_j - p_j).$$

Then,

$$\begin{aligned} \left| \frac{f_i(\mathbf{x}) - f_i(\mathbf{p}) - Df_i(\mathbf{p})(\mathbf{x}-\mathbf{p})}{\|\mathbf{x}-\mathbf{p}\|} \right| &= \left| \sum_j \frac{\left(\dfrac{\partial f_i}{\partial x_j}(\mathbf{r}^j(t_{ij})) - \dfrac{\partial f_i}{\partial x_j}(\mathbf{r}) \right)(x_j - p_j)}{\|\mathbf{x}-\mathbf{p}\|} \right| \\ &\le \sum_j \left| \frac{\partial f_i}{\partial x_j}(\mathbf{r}^j(t_{ij})) - \frac{\partial f_i}{\partial x_j}(\mathbf{r}) \right|. \end{aligned}$$

This last term goes to zero because the partial derivatives are continuous. Since this is true for each coordinate function, it is true for the vector valued function. □

The following result corresponds to the usual chain rule for functions of one variable and the chain rule for partial derivatives for functions of several variables.

Theorem 2.18 (Chain Rule). *Assume that $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^k$ are C^1, $\mathbf{p} \in \mathbb{R}^n$ and $\mathbf{q} = f(\mathbf{p}) \in \mathbb{R}^m$. Then the composition $g \circ f : \mathbb{R}^n \to \mathbb{R}^k$ is C^1, and*

$$D(g \circ f)(\mathbf{p}) = Dg(\mathbf{q})\,Df(\mathbf{p}).$$

This form of the chain rule agrees with the usual chain rule for functions of several variables written in terms of partial derivatives: Assume that $w = g(\mathbf{x})$ and $\mathbf{x}(t)$, then

$$\frac{dw}{dt} = \left(\frac{\partial w}{\partial x_1}, \ldots, \frac{\partial w}{\partial x_n} \right) \left(\frac{dx_1}{dt}, \ldots, \frac{dx_n}{dt} \right)^{\mathsf{T}} = \sum_i \frac{\partial w}{\partial x_i} \frac{dx_i}{dt}.$$

Proof. We let $\mathbf{v} = \mathbf{x} - \mathbf{p}$ and $\mathbf{w} = \mathbf{y} - \mathbf{q}$. In the limits for the derivative, we write $\widetilde{R}_f(\mathbf{v})$ for $\widetilde{R}_f(\mathbf{p}, \mathbf{p}+\mathbf{v})$ and $\widetilde{R}_g(\mathbf{v})$ for $\widetilde{R}_g(\mathbf{q}, \mathbf{q}+\mathbf{w})$.

$$\begin{aligned} g \circ f(\mathbf{p}+\mathbf{v}) &= g\left(\mathbf{p} + Df(\mathbf{p})\,\mathbf{v} + \widetilde{R}_f(\mathbf{v})\,\|\mathbf{v}\| \right) \\ &= g(\mathbf{q}) + Dg(\mathbf{q})\ \left[Df(\mathbf{p})\,\mathbf{v} + \widetilde{R}_f(\mathbf{v})\|\mathbf{v}\| \right] \\ &\quad + \widetilde{R}_g\left(Df(\mathbf{p})\,\mathbf{v} + \widetilde{R}_f(\mathbf{v})\|\mathbf{v}\| \right) \left\| Df(\mathbf{p})\,\mathbf{v} + \widetilde{R}_f(\mathbf{v})\|\mathbf{v}\| \right\| \\ &= g(\mathbf{q}) + Dg(\mathbf{q})\,Df(\mathbf{p})\,\mathbf{v} + \left[Dg(\mathbf{q})\,\widetilde{R}_f(\mathbf{v}) \right]\ \|\mathbf{v}\| \\ &\quad + \widetilde{R}_g\left(Df(\mathbf{p})\,\mathbf{v} + \widetilde{R}_f(\mathbf{v})\|\mathbf{v}\| \right) . \left\| Df(\mathbf{p})\,\mathbf{v} + \widetilde{R}_f(\mathbf{v})\|\mathbf{v}\| \right\| \end{aligned}$$

The first term $Dg(\mathbf{q})\,\widetilde{R}_f(\mathbf{v})$ goes to zero as needed and is multiplied by $\|\mathbf{v}\|$. For the second term,

$$\frac{\left\| Df(\mathbf{p})\,\mathbf{v} + \widetilde{R}_f(\mathbf{v})\|\mathbf{v}\| \right\|}{\|\mathbf{v}\|} = \left\| Df(\mathbf{p})\,\frac{\mathbf{v}}{\|\mathbf{v}\|} + \widetilde{R}_f(\mathbf{v}) \right\|$$

is bounded as $\|\mathbf{v}\|$ goes to zero, and $\widetilde{R}_g\left(Df(\mathbf{p})\,\mathbf{v} + \widetilde{R}_f(\mathbf{v})\|\mathbf{v}\|\right)$ goes to zero. Thus,

$$g \circ f(\mathbf{p}+\mathbf{v}) = g(\mathbf{q}) + Dg(\mathbf{q})\,Df(\mathbf{p})\,\mathbf{v} + \widetilde{R}_{g\circ f}(\mathbf{v})\,\|\mathbf{v}\|$$

where $\widetilde{R}_{g\circ f}(\mathbf{v})$ goes to zero as $\|\mathbf{v}\|$ goes to zero. It follows that $Dg(\mathbf{q})\,Df(\mathbf{p})$ must be the derivative. □

2.1.5. Second Derivative and Taylor's Theorem

The only higher derivative that we consider is the second derivative of a real valued function of multiple variables. Other higher derivatives are more complicated to express. We use this second derivative to state Taylor's theorem for a multi-variable real valued function, giving the expansion with linear and quadratic terms. We indicate why this multi-variable version follows from Taylor's Theorem for a function of a single variable.

Definition. Let $\mathscr{D} \subset \mathbb{R}^n$ be an open set and $f : \mathscr{D} \to \mathbb{R}$. If all the second partial derivatives $\dfrac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{p})$ exists and are continuous for all $\mathbf{p} \in \mathscr{D}$ then f is said to be *twice continuously differentiable* or C^2.

The matrix of second partial derivatives $\left(\dfrac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{p})\right)$ is called the *second derivative at* $\mathbf{p}$ and is denoted by $D^2 f(\mathbf{p})$. Some authors call it the *Hessian matrix of* f.

In more formal treatments of calculus, $D^2 f(\mathbf{p})$ can be understood as follows. The matrix $Df(\mathbf{x})$ can be considered as a point in $\mathbb{R}^n$, and the map $Df : \mathbf{x} \mapsto Df(\mathbf{x})$ is a function from $\mathbb{R}^n$ to $\mathbb{R}^n$. The derivative of the map Df at $\mathbf{x}_0$ would be an $n \times n$ matrix, which is the second derivative $D^2 f(\mathbf{x}_0)$. This matrix can be applied to two vectors $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^n$; the result is a real number, $\mathbf{v}_1^{\mathsf{T}} D^2 f(\mathbf{p})\,\mathbf{v}_2$. The second derivative is a bilinear map from $\mathbb{R}^n$ to $\mathbb{R}$, i.e., it takes two vectors and gives a number and is linear in each of the vectors separately.

Theorem 2.19. *Let $\mathscr{D} \subset \mathbb{R}^n$ be open and $f : \mathscr{D} \to \mathbb{R}$ be C^2. Then for all pairs $1 \le i, j \le n$,*

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{p}) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{p}),$$

i.e., $D^2 f(\mathbf{p})$ is a symmetric matrix.

The symmetric matrix $D^2 f(\mathbf{p})$ defines a quadratic form

$$(\mathbf{x}-\mathbf{p})^{\mathsf{T}}\, D^2 f(\mathbf{p})\, (\mathbf{x}-\mathbf{p}) = \sum_{i,j} \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{p})(x_j - p_j)(x_i - p_i) \qquad \text{for } \mathbf{x} \in \mathbb{R}^n,$$

which is used in Taylor's Theorem for a function of several variables.

Theorem 2.20 (Taylor's Theorem for a Single Variable). *Assume that $g : \mathscr{D} \subset \mathbb{R} \to \mathbb{R}$ is C^r, i.e., has r continuous derivatives. We denote the k^{th} derivative at x by $g^{(k)}(x)$. Assume that p is in the interior of $\mathscr{D}$. Then*

$$g(x) = g(p) + \sum_{k=1}^{r} \frac{1}{k!} g^{(k)}(p)(x-p)^k + R_r(p,x)$$

where the remainder

$$R_r(p,x) = \frac{1}{(r-1)!} \int_p^x (x-t)^{r-1} \left[g^{(r)}(t) - g^{(r)}(p) \right] \, dt \qquad \textit{and satisfies}$$

$$\lim_{x \to p} \frac{|R_r(p,x)|}{|x-p|^r} = 0.$$

If g is C^{r+1}, then the remainder can also be given by either of the following expressions:

$$\begin{aligned} R_r(p,x) &= \frac{1}{r!} \int_p^x (x-t)^r g^{(r+1)}(t) \, dt \\ &= \frac{1}{(r+1)!} g^{(r+1)}(c)(x-p)^{r+1}, \end{aligned}$$

where c is between p and x.

Proof. The second form of the remainder can be proved by induction using integration by parts. The other two forms of the remainder can be proved from the second. We refer the reader to a book on calculus.

To estimate the remainder we use the first form of the remainder. Let

$$C_r(x) = \sup \left\{ \left| g^{(r)}(t) - g^{(r)}(p) \right| : t \text{ is between } p \text{ and } x \right\}.$$

We treat the case with $x > p$ and leave the details for $x < p$ to the reader. Then,

$$\begin{aligned} |R_r(p,x)| &= \left| \frac{1}{(r-1)!} \int_p^x (x-t)^{r-1} \left[g^{(r)}(t) - g^{(r)}(p) \right] \, dt \right| \\ &\leq \frac{1}{(r-1)!} \int_p^x (x-t)^{r-1} \left| g^{(r)}(t) - g^{(r)}(p) \right| \, dt \\ &\leq \frac{1}{(r-1)!} \int_p^x (x-t)^{r-1} C_r(x) \, dt \\ &= \frac{1}{r!} (x-p)^r C_r(x). \end{aligned}$$

Since $C_r(x)$ goes to zero as x goes to p, we get the desired result. □

Theorem 2.21 (Taylor's Theorem for a Multi-variable Function). *Assume that $F : \mathscr{D} \subset \mathbb{R}^n \to \mathbb{R}$ is C^2 on $\operatorname{int}(\mathscr{D})$ and $\mathbf{p} \in \operatorname{int}(\mathscr{D})$. Then,*

$$F(\mathbf{x}) = F(\mathbf{p}) + DF(\mathbf{p})\,(\mathbf{x}-\mathbf{p}) + \tfrac{1}{2}\,(\mathbf{x}-\mathbf{p})^{\mathsf{T}}\, D^2F(\mathbf{p})\,(\mathbf{x}-\mathbf{p}) + R_2(\mathbf{p},\mathbf{x})$$

where

$$\lim_{\mathbf{x} \to \mathbf{p}} \frac{R_2(\mathbf{p},\mathbf{x})}{\|\mathbf{x}-\mathbf{p}\|^2} = 0,$$

i.e., if $R_2(\mathbf{p},\mathbf{x}) = \widetilde{R}_2(\mathbf{p},\mathbf{x})\|\mathbf{x}-\mathbf{p}\|^2$, then $\lim_{\mathbf{x}\to\mathbf{p}} \widetilde{R}_2(\mathbf{p},\mathbf{x}) = 0$.

Proof. Define

$$\begin{aligned} \mathbf{x}_t &= \mathbf{p} + t(\mathbf{x}-\mathbf{p}) \qquad \text{and} \\ g(t) &= F(\mathbf{x}_t), \end{aligned}$$

so $g(0) = F(\mathbf{p})$ and $g(1) = F(\mathbf{x})$. For $\mathbf{x}$ near enough to $\mathbf{p}$, $\mathbf{x}_t \in \mathcal{D}$ for $0 \le t \le 1$. The derivatives of g in terms of F are

$$\begin{aligned}
g'(t) &= \sum_{i=1}^{n} \frac{\partial F}{\partial x_i}(\mathbf{x}_t)(x_i - p_i) \\
g'(0) &= DF(\mathbf{p})(\mathbf{x} - \mathbf{p}), \\
g''(t) &= \sum_{\substack{i=1,\dots,n \\ j=1,\dots,n}} \frac{\partial^2 F}{\partial x_j \partial x_i}(\mathbf{x}_t)(x_i - p_i)(x_j - p_j) \\
g''(0) &= (\mathbf{x} - \mathbf{p})^{\mathsf{T}} D^2 F(\mathbf{p})(\mathbf{x} - \mathbf{p}).
\end{aligned}$$

Applying usual Taylor's Theorem for a function of one variable to g gives the result including the estimate on the remainder. □

Remark. For a 3×3 symmetric matrix $\mathbf{A} = (a_{ij})$, $a_{21} = a_{12}$, $a_{31} = a_{13}$, and $a_{32} = a_{23}$, so

$$\mathbf{v}^{\mathsf{T}} \mathbf{A} \mathbf{v} = a_{11}\, v_1^2 + a_{22}\, v_2^2 + a_{33}\, v_3^2 + 2\, a_{12}\, v_1\, v_2 + 2\, a_{13}\, v_1\, v_3 + 2\, a_{23}\, v_2\, v_3.$$

If we apply that last formula to the Taylor's expansion of a function $f : \mathbb{R}^3 \to \mathbb{R}$, we get

$$\begin{aligned}
F(\mathbf{x}) = F(&\mathbf{p} + Df(\mathbf{p})(\mathbf{x} - \mathbf{p}) \\
&+ \tfrac{1}{2} \frac{\partial^2 F}{\partial x_1^2}(\mathbf{p})(x_1 - p_1)^2 + \tfrac{1}{2} \frac{\partial^2 F}{\partial x_2^2}(\mathbf{p})(x_2 - p_2)^2 + \tfrac{1}{2} \frac{\partial^2 F}{\partial x_3^2}(\mathbf{p})(x_3 - p_3)^2 \\
&+ \frac{\partial^2 F}{\partial x_1 \partial x_2}(\mathbf{p})(x_1 - p_1)(x_2 - p_2) + \frac{\partial^2 F}{\partial x_1 \partial x_3}(\mathbf{p})(x_1 - p_1)(x_3 - p_3) \\
&+ \frac{\partial^2 F}{\partial x_2 \partial x_3}(\mathbf{p})(x_2 - p_2)(x_3 - p_3) + R_2(\mathbf{p}, \mathbf{x}).
\end{aligned}$$

Example 2.22. Find the second order Taylor expansion of $F(x, y, z) = 3x^2 y + y^3 - 3x^2 - 3y^2 + z^3 - 3z$, about the point $\mathbf{p} = (1, -2, 3)$.

$F(\mathbf{p}) = -8$. The first order partial derivatives are

$$\begin{aligned}
\frac{\partial F}{\partial x} &= 6xy - 6x, & \frac{\partial F}{\partial x}(\mathbf{p}) &= -18, \\
\frac{\partial F}{\partial y} &= 3x^2 + 3y^2 - 6y, & \frac{\partial F}{\partial y}(\mathbf{p}) &= 27, \\
\frac{\partial F}{\partial z} &= 3z^2 - 3, & \frac{\partial F}{\partial z}(\mathbf{p}) &= 24.
\end{aligned}$$

The second order partial derivatives of F are

$$\begin{aligned}
\frac{\partial^2 F}{\partial x^2} &= 6y - 6, & \frac{\partial^2 F}{\partial x^2}(\mathbf{p}) &= -18,\\
\frac{\partial^2 F}{\partial y^2} &= 6y - 6, & \frac{\partial^2 F}{\partial y^2}(\mathbf{p}) &= -18,\\
\frac{\partial^2 F}{\partial z^2} &= 6z, & \frac{\partial^2 F}{\partial z^2}(\mathbf{p}) &= 18,\\
\frac{\partial^2 F}{\partial x \partial y} &= 6x, & \frac{\partial^2 F}{\partial x \partial y}(\mathbf{p}) &= 6,\\
\frac{\partial^2 F}{\partial x \partial z} &= 0, & \frac{\partial^2 F}{\partial x \partial z}(\mathbf{p}) &= 0,\\
\frac{\partial^2 F}{\partial y \partial z} &= 0, & \frac{\partial^2 F}{\partial y \partial z}(\mathbf{p}) &= 0.
\end{aligned}$$

The second order Taylor expansion about $(1, -2, 3)$ is

$$\begin{aligned}
F(x, y, z) = -8 &- 18(x-1) + 27(y+2) + 24(z-3)\\
&- 9(x-1)^2 - 9(y+2)^2 + 9(z-3)^2\\
&+ 6(x-1)(y+2) + R_2(\mathbf{p}, (x, y, z)).
\end{aligned}$$

■

2.1.6. Quadratic Forms

Later in the chapter, we give conditions that insure that a critical point $\mathbf{x}^*$ is a local extremizer based on the quadratic form determined by $D^2 f(\mathbf{x}^*)$. To prepare for that material, we review and extend the material about quadratic forms from linear algebra.

Definition. Let $\mathbf{A} = (a_{ij})$ be an $n \times n$ symmetric matrix. Then

$$\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} = \sum_{i,j=1}^{n} a_{ij} x_i x_j \qquad \text{for } \mathbf{x} \in \mathbb{R}^n$$

is called a *quadratic form*.

Definition. The quadratic form $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}$ is called

- **i.** *positive definite* provided that $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$,
- **ii.** *positive semidefinite* provided that $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} \geq 0$ for all $\mathbf{x}$,
- **iii.** *negative definite* provided that $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} < 0$ for all $\mathbf{x} \neq \mathbf{0}$,
- **iv.** *negative semidefinite* provided that $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} \leq 0$ for all $\mathbf{x}$, and
- **v.** *indefinite* provided that $\mathbf{x}_1^\mathsf{T}\mathbf{A}\mathbf{x}_1 > 0$ for some $\mathbf{x}_1$ and $\mathbf{x}_2^\mathsf{T}\mathbf{A}\mathbf{x}_2 > 0$ for some other $\mathbf{x}_2$.

Since $(s\mathbf{x})^\mathsf{T}\mathbf{A}(s\mathbf{x}) = s^2\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}$, the sign of $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}$ is determined by its sign on unit vectors.

Definition. For an $n \times n$ symmetric matrix $\mathbf{A}$, the *principal submatrices* of $\mathbf{A}$ are the $k \times k$ submatrices in the upper left hand corner,

$$\mathbf{A}_k = (a_{ij})_{1 \leq i,j \leq k} \qquad \text{for } 1 \leq k \leq n.$$

We let $\Delta_k = \det(\mathbf{A}_k)$ be the determinants of the principal submatrices.

Theorem 2.23 (**Test for Definite Matrices**). *Let* $\mathbf{A}$ *be an* $n \times n$ *symmetric matrix.*

a. *The following three conditions are equivalent:*
- **i.** *The matrix* $\mathbf{A}$ *is positive definite.*
- **ii.** *All the eigenvalues of* $\mathbf{A}$ *are positive.*
- **iii.** *The determinant of every principal submatrices is positive,* $\det(\mathbf{A}_k) > 0$ *for* $1 \leq k \leq n$.
- **iv.** *The matrix* $\mathbf{A}$ *can be row reduced without row exchanges or scalar multiplications of rows to an upper triangular matrix that has* n *positive pivots, i.e., all the pivots are positive.*

b. *The following three conditions are equivalent:*
- **i.** *The matrix* $\mathbf{A}$ *is negative definite.*
- **ii.** *All the eigenvalues of* $\mathbf{A}$ *are negative.*
- **iii.** *The determinants of the principal submatrices alternate sign,* $(-1)^k \det(\mathbf{A}_k) > 0$ *for* $1 \leq k \leq n,\ \Delta_1 < 0,\ \Delta_2 > 0,\ \Delta_3 < 0, \ldots$
- **iv.** *The matrix* $\mathbf{A}$ *can be row reduced without row exchanges or scalar multiplications of rows to an upper triangular matrix that has* n *negative pivots, i.e., all the pivots are negative.*

c. *The following two conditions are equivalent:*
- **i.** *The matrix* $\mathbf{A}$ *is indefinite.*
- **ii.** *The matrix* $\mathbf{A}$ *has at least one positive and one negative eigenvalue.*

Any one of the following conditions implies conditions (c.i) and (c.ii):
- **iii.** $\det(\mathbf{A}) = \det(\mathbf{A}_n) \neq 0$ *and the pattern of signs of* $\Delta_k = \det(\mathbf{A}_k)$ *is different than those of both part (a) and (b), allowing one of the other* $\det(\mathbf{A}_k) = 0$.
- **iv.** *The matrix* $\mathbf{A}$ *can be row reduced to an upper triangular matrix without row exchanges or a scalar multiplication of a row and all* n *of the pivots are nonzero and some pivot* $p_j > 0$ *and another* $p_k < 0$.
- **v** *The matrix* $\mathbf{A}$ *cannot be row reduced to an upper triangular matrix without row exchanges.*

We discuss the proof in an appendix at the end of the chapter. For a 3×3 matrix, the calculation of the determinants of the principal submatrices is the most direct method of determining whether a symmetric matrix is positive or negative definite. Row reduction is probably the easiest for matrices larger than 3×3.

Remark. *Negative Definite*: The idea behind the condition that $(-1)^k \det(\mathbf{A}_k) > 0$ for all k for negative definite matrix is that the product of k negative numbers has the same sign as $(-1)^k$.

Nonzero Determinant: If $\det(\mathbf{A}) \neq 0$, then all the eigenvalues are nonzero, and Cases a(iii), b(iii), and c(iii) tell whether it is positive definite, negative definite, or indefinite. (To remember the signs of the determinants, think of the diagonal case.)

Zero Determinant: If $\det(\mathbf{A}) = 0$, then some eigenvalue is zero and $\mathbf{A}$ can be either indefinite or positive semi-definite or negative semi-definite. There is no simple general rule.

The following theorem collects together using row reduction to test for the type of symmetric matrices. It includes the results about positive and negative definite matrices given earlier.

Theorem 2.24 (**Row Reduction Test**). *Let* $\mathbf{A}$ *be an* $n \times n$ *symmetric matrix,* $\mathbf{Q}(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}$ *the related quadratic form.*

a. *Assume that* $\mathbf{A}$ *can be row reduced to an upper triangular matrix without row exchanges or scalar multiplications of rows.*
- **i.** *If all the pivots satisfy* $p_j > 0$ *for* $1 \leq j \leq n$, *then* $\mathbf{A}$ *is positive definite.*
- **ii.** *If all the pivots satisfy* $p_j < 0$ *for* $1 \leq j \leq n$, *then* $\mathbf{A}$ *is negative definite.*

iii. *If all the pivots satisfy $p_j \geq 0$ for $1 \leq j \leq n$, then $\mathbf{A}$ is positive semi-definite.*
iv. *If all the pivots satisfy $p_j \leq 0$ for $1 \leq j \leq n$, then $\mathbf{A}$ is negative semi-definite.*
v. *If some pivot $p_j > 0$ and another $p_k < 0$, then $\mathbf{A}$ is indefinite.*

b. *If $\mathbf{A}$ cannot be row reduced to an upper triangular matrix without row exchanges, then $\mathbf{A}$ is indefinite.*

Remark (Row Reduction Conditions). If the matrix can be row reduced to upper triangular form without row exchanges or scalar multiplication of a row, then $\det(\mathbf{A}_k) = p_1 \cdots p_k$. If all the pivots are positive, then all the determinants of the $\mathbf{A}_k$ are positive. If all the pivots are negative, then the determinants of the $\mathbf{A}_k$ alternate signs are required. If all the pivots are nonzero and both signs appear, then the signs of the determinants do not fit the pattern for either positive definite or negative definite, so $\mathbf{A}$ must be indefinite. Finally, if the matrix cannot be row reduced to an upper triangular matrix without row exchanges, then some submatrix must of the form $\begin{bmatrix} 0 & a \\ a & b \end{bmatrix}$, with 0 on the diagonal. This insures the matrix is indefinite. See Theorem 3.3.12 in [**2**].

Example 2.25. Let

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}.$$

The principal submatrices and their determinants are

$$\begin{aligned} &\mathbf{A}_1 = (2), && \det(\mathbf{A}_1) = 2 > 0, \\ &\mathbf{A}_2 = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}, && \det(\mathbf{A}_2) = 3 > 0, \\ &\mathbf{A}_3 = \mathbf{A} && \det(\mathbf{A}) = 4 > 0. \end{aligned}$$

Since the signs of these determinants are all positive, the quadratic form induced by $\mathbf{A}$ is positive definite. This method is the easiest to calculate for a 3×3 matrix.

Alternatively, we can row reduce without any row exchanges to to an upper triangular matrix. Row reduction is equivalent to multiplying on the left by a matrix. We given that matrix following the procedure in §2.5 of Lay [**9**]. Since the matrix is symmetric, $\mathbf{A}$ can be written as the product of lower triangular, diagonal, and upper triangular matrices.

$$\begin{aligned} \mathbf{A} = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} &= \begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & -1 & 0 \\ 0 & \frac{3}{2} & -1 \\ 0 & -1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{-1}{2} & 1 & 0 \\ 0 & -\frac{2}{3} & 1 \end{bmatrix} \begin{bmatrix} 2 & -1 & 0 \\ 0 & \frac{3}{2} & -1 \\ 0 & 0 & \frac{4}{3} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 \\ 0 & -\frac{2}{3} & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & \frac{3}{2} & 0 \\ 0 & 0 & \frac{4}{3} \end{bmatrix} \begin{bmatrix} 1 & -\frac{1}{2} & 0 \\ 0 & 1 & -\frac{2}{3} \\ 0 & 0 & 1 \end{bmatrix}. \end{aligned}$$

The pivots are all positive so $\mathbf{A}$ is positive definite: $p_1 = 2 > 0$, $p_2 = \frac{3}{2} > 0$, and $p_3 = \frac{4}{3} > 0$. This method would be the easiest to calculate for a matrix larger than 3×3.

The quadratic form can be written as a sum of squares with all positive coefficients:

$$\mathbf{Q}(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{U}^\mathsf{T}\mathbf{D}\mathbf{U}\mathbf{x} = 2\left(x_1 - \frac{1}{2}x_2\right)^2 + \frac{3}{2}\left(x_2 - \frac{2}{3}x_3\right)^2 + \frac{4}{3}x_3^2.$$

The eigenvalues are not especially easy to calculate, but they are 2 and $2 \pm \sqrt{2}$ which are all positive. The signs of these eigenvalues are correct for $\mathbf{A}$ to be positive definite. ■

Proof of Theorem 2.23(a). Lay [**9**] discusses some aspects of the proof of this theorem. A more complete reference is [**13**] by Gilbert Strang.

In the proof, we need to add the following intermediate steps.

(v) For each $1 \leq k \leq n$, the quadratic form associated to $\mathbf{A}_k$ is positive definite.

(vi) All the eigenvalues of $\mathbf{A}_k$ are positive for $1 \leq k \leq n$.

(vii) By completion of the squares, $\mathbf{Q}(\mathbf{x})$ can be represented as a sum of squares, with all positive coefficients,

$$\begin{aligned}\mathbf{Q}(x_1,\ldots,x_n) &= (x_1,\ldots,x_n)\mathbf{U}^\mathsf{T}\mathbf{D}\mathbf{U}(x_1,\ldots,x_n)^\mathsf{T}\\ &= p_1\left(x_1 + u_{1,2}x_2 + \cdots + u_{1,n}x_n\right)^2\\ &\quad + p_2\left(x_2 + u_{2,3}x_3 + \cdots + u_{2,n}x_n\right)^2\\ &\quad + \cdots + p_n x_n^2.\end{aligned}$$

(**i** $\Leftrightarrow$ **ii**) The fact the positive definite is equivalent to all positive eigenvalues is proved in Lay [**9**].

(**i** $\Rightarrow$ **v**) Assume $\mathbf{Q}$ is positive definite. Then, for any $1 \leq k \leq n$ and any $(x_1,\ldots,x_k) \neq \mathbf{0}$,

$$\begin{aligned}0 &< \mathbf{Q}(x_1,\ldots,x_k,0,\ldots,0)\\ &= (x_1,\ldots,x_k,0,\ldots,0)\mathbf{A}(x_1,\ldots,x_k,0,\ldots,0)^\mathsf{T}\\ &= (x_1,\ldots,x_k)\mathbf{A}_k(x_1,\ldots,x_k)^\mathsf{T}.\end{aligned}$$

This shows that (i) implies (v).

(**v** $\Leftrightarrow$ **vi**) This is the same as the result (**i** $\Leftrightarrow$ **ii**) applied to $\mathbf{A}_k$. Notice that the eigenvalues of $\mathbf{A}_k$ are not necessarily eigenvalues of $\mathbf{A}$.

(**vi** $\Rightarrow$ **iii**) For any k, $\det(\mathbf{A}_k)$ is positive since it is the product of its eigenvalues of $\mathbf{A}_k$,

(**iii** $\Rightarrow$ **iv**) Assume (iii). If $\mathbf{A}$ cannot be row reduced without row exchanges, then there is a subblock down the diagonal of the form $\begin{bmatrix}0 & a\\ a & b\end{bmatrix}$ and some $\det(\mathbf{A}_k) = 0$ and $\det(\mathbf{A}_{k+1}) =$ - $\det(\mathbf{A}_{k-1})$. If we can row reduce $\mathbf{A}$ without row exchanges, then it also row reduces all the $\mathbf{A}_k$. Therefore the pivots of the $\mathbf{A}_k$ are pivots of $\mathbf{A}$. Also, the determinant of $\mathbf{A}_k$ is the product of the first k pivots, $\det(\mathbf{A}_k) = p_1 \ldots p_k$. Therefore

$$p_k = (p_1 \ldots p_k)/(p_1 \ldots p_{k-1}) = \det(\mathbf{A}_k)/\det(\mathbf{A}_{k-1} > 0,$$

for each k. This proves (iv).

(**iv** $\Rightarrow$ **vii**) Assume (iv). Row reduction can be realized by matrix multiplication on the left by a lower triangular matrix with ones on the diagonal. Therefore, $\mathbf{L}^{-1}\mathbf{A} = \mathbf{U}_1 = \mathbf{D}\mathbf{U}$, where $\mathbf{U}_1$ and $\mathbf{U}$ are upper triangular, $\mathbf{D}$ is a diagonal matrix with the pivots on the diagonal, and $\mathbf{U}$ has ones on the diagonal. (See §2.5 in [**9**].) Since $\mathbf{A}$ is symmetric, $\mathbf{L}\mathbf{D}\mathbf{U} = \mathbf{A} = \mathbf{A}^\mathsf{T} = \mathbf{U}^\mathsf{T}\mathbf{D}\mathbf{L}^\mathsf{T}$. It can then be shown that the factorization is unique and $\mathbf{U}^\mathsf{T} = \mathbf{L}$, so

$$\begin{aligned}\mathbf{Q}(x_1,\ldots,x_n) &= (x_1,\ldots,x_n)\mathbf{U}^\mathsf{T}\mathbf{D}\mathbf{U}(x_1,\ldots,x_n)^\mathsf{T}\\ &= p_1\left(x_1 + u_{1,2}x_2 + \cdots + u_{1,n}x_n\right)^2\\ &\quad + p_2\left(x_2 + u_{2,3}x_3 + \cdots + u_{2,n}x_n\right)^2\\ &\quad + \cdots + p_n x_n^2.\end{aligned}$$

Thus, we can "complete the squares", expressing $\mathbf{Q}$ as the sum of squares with the pivots as the coefficients. If the pivots are all positive, then all the coefficients p_i are positive. Thus (iv) implies (vii). Note that $\mathbf{z} = \mathbf{U}\mathbf{x}$ is a non-orthonormal change of basis that makes the quadratic form diagonal.

(**vii** $\Rightarrow$ **i**) If $\mathbf{Q}(\mathbf{x})$ can be written as the sum of squares of the above form with positive coefficients, then the quadratic form must be positive. Thus, (vii) implies (i). □

2.1. Exercises

2.1.1. Consider the sets

$$\mathbf{S}_1 = \{\, (x, y) \in \mathbb{R}^2 : \text{-}1 < x < 1 \,\}$$
$$\mathbf{S}_2 = \{\, (x, y) \in \mathbb{R}^2 : x \geq 1,\ y \geq 0 \,\}.$$

a. For the sets $\mathbf{S}_1$ and $\mathbf{S}_2$, discuss which points are in the boundary and which points are not using the definition of the boundary.

b. Discuss why $\mathbf{S}_1$ is open in two ways: (i) $\mathbf{S}_1 \cap \partial(\mathbf{S}_1) = \emptyset$ and (ii) for every point $\mathbf{p} \in \mathbf{S}_1$, there is an $r > 0$ such that $\mathbf{B}(\mathbf{p}, r) \subset \mathbf{S}_1$.

c. Discuss why $\mathbf{S}_2$ is closed in two ways: (i) $\partial(\mathbf{S}_2) \subset \mathbf{S}_2$ and (ii) for every point $\mathbf{p} \in \mathbf{S}_2^c$, there is an $r > 0$ such that $\mathbf{B}(\mathbf{p}, r) \subset \mathbf{S}_2^c$.

2.1.2. Consider the sets

$$\mathbf{S}_1 = \{(x, y) : x > 0,\ y > 0,\ xy < 1 \},$$
$$\mathbf{S}_2 = \{(x, y) : x \geq 0,\ y \geq 0,\ xy \leq 1 \},$$
$$\mathbf{S}_3 = \{(x, y) : x \geq 0,\ y \geq 0,\ 2x + 3y \leq 7 \}.$$

a. For each of the sets $\mathbf{S}_1$, $\mathbf{S}_2$, and $\mathbf{S}_3$, discuss which points are in the boundary and which points are not using the definition of the boundary.

b. Discuss why $\mathbf{S}_1$ is open in two ways: (i) $\mathbf{S}_1 \cap \partial(\mathbf{S}_1) = \emptyset$ and (ii) for every point $\mathbf{p} \in \mathbf{S}_1$, there is an $r > 0$ such that $\mathbf{B}(\mathbf{p}, r) \subset \mathbf{S}_1$.

c. Discuss why $\mathbf{S}_2$ and $\mathbf{S}_3$ are closed in two ways: (i) $\partial(\mathbf{S}) \subset \mathbf{S}$ and (ii) for every point $\mathbf{p} \in \mathbf{S}^c$, there is an $r > 0$ such that $\mathbf{B}(\mathbf{p}, r) \subset \mathbf{S}^c$.

2.1.3. Which of the following sets are open, closed, and or compact? Explain why your answer is correct.

a. Let $g_1(x, y, z) = x^2 + y^2$, $g_2(x, y, z) = x^2 + z^2$, and

$$\begin{aligned}\mathbf{S}_1 &= g_1^{-1}([0, 9]) \cap g_2^{-1}([0, 4]) \\ &= \{\, (x, y, z) \in \mathbb{R}^3 : x^2 + y^2 \leq 9,\ x^2 + z^2 \leq 4 \,\}.\end{aligned}$$

b. Let $g(x, y) = |x - y|$ and

$$\mathbf{S}_2 = g^{-1}([0, 1]) = \{(x, y) \in \mathbb{R}^2 : |x - y| \leq 1 \}.$$

c. Assume that $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ are continuous and

$$\begin{aligned}\mathbf{S}_3 &= f^{-1}(0, \infty) \cap g^{-1}(-\infty, 2) \\ &= \{\, \mathbf{x} \in \mathbb{R}^n : 0 < f(\mathbf{x}),\ g(\mathbf{x}) < 2 \,\}.\end{aligned}$$

2.1.4. Assume that $f : \mathscr{D} \subset \mathbb{R}^n \to \mathbb{R}$ is continuous. The Extreme Value Theorem says that if $\mathscr{F} \subset \mathscr{D}$ is compact, then f attains its maximum and minimum on $\mathscr{F}$.

a Give an example to show that the conclusion is false if $\mathscr{F}$ is not bounded.

b When if $\mathscr{F}$ is not closed, give examples where (i) f is not bounded above and (ii) f is bounded above but does not attain a maximum

2.1.5. Let $f : \mathbb{R}_+ \to \mathbb{R}$ be continuous, $f(0) > 1$, and $\lim_{x \to \infty} f(x) = 0$.

a. Show that there is a $p > 0$ such that the maximal value of $f(x)$ on $[0, p]$ is larger than any value of $f(x)$ for $x > p$. *Hint:* Take p such that $f(x) < \frac{1}{2} f(0)$ for $x \geq p$.

b. Show that $f(x)$ has a maximum on $\mathbb{R}_+$.

c. Does $f(x)$ have to have a minimum on $\mathbb{R}_+$? Explain why or why not.

2.1.6. Show that the function

$$f(x,y) = \frac{2x+y}{4x^2+y^2+8}$$

attains a maximum on $\mathbb{R}^2_+$
Hint: $f(x,y) \le {}^{3R}\!/_{R^2} = {}^{3}\!/_{R}$ for $\|(x,y)\| = R > 0$.

2.1.7. Let $\mathscr{F} = \{\,(x,y) \in \mathbb{R}^2_+ : xy \ge 1\,\}$ and $\mathscr{B} = \{\,(x,y) \in \mathbb{R}^2_+ : x+y \le 10\,\}$. (Note that $\mathscr{F}$ is not compact.) Assume that $f : \mathbb{R}^2_+ \to \mathbb{R}$ is a continuous function with $f(x,y) > f(2,3)$ for $x+y > 10$, i.e., for $(x,y) \in \mathbb{R}^2_+ \setminus \mathscr{B}$.

a. Why must f attain a minimum on $\mathscr{F} \cap \mathscr{B}$?

b. Using reasoning like for Exercise 2.1.4, explain why f attains a minimum on all of $\mathscr{F}$.

2.1.8. Compute the second order Taylor polynomial (without explicit remainder) for $f(x,y) = e^x \cos(y)$ around $(x_0, y_0) = (0,0)$. You do not need to find an expression for the remainder.

2.1.9. Compute the second order Taylor polynomial of $f(x,y) = xy^2$ about $(x^*, y^*) = (2,1)$. You do not need to find an expression for the remainder.

2.1.10. Decide whether the following matrices are positive definite, negative definite, or neither:

(a) $\begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$ (b) $\begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & 1 \\ -1 & 1 & 2 \end{pmatrix}$

(c) $\begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 4 \\ 3 & 4 & 9 \end{pmatrix}$ (d) $\begin{pmatrix} 1 & 2 & 0 & 0 \\ 2 & 6 & -2 & 0 \\ 0 & -2 & 5 & -2 \\ 0 & 0 & -2 & 3 \end{pmatrix}$

2.2. Derivative Conditions

In this section, we consider derivative conditions that are necessary for an extremizer and ones that are sufficient for a local extremizer. We treat the general multi-dimensional case using the derivative and second derivative as matrices, rather than expressing the conditions in two and three dimensions in terms of partial derivatives as is done in many multi-dimensional calculus courses. In contrast with linear functions, nonlinear functions can have local maximizers that are not global maximizers. Therefore, we start by giving definitions that indicate the difference between these two concepts.

Definition. We say that a function $f : \mathscr{F} \subset \mathbb{R}^n \to \mathbb{R}$ has a *maximum* or *global maximum* at a point $\mathbf{x}_M \in \mathscr{F}$ provided that $f(\mathbf{x}) \le f(\mathbf{x}_M)$ for all $\mathbf{x} \in \mathscr{F}$. We also say that the point $\mathbf{x}_M$ is a *maximizer* of f on $\mathscr{F}$. It has a *strict maximum* at $\mathbf{x}_M$ provided that $f(\mathbf{x}) < f(\mathbf{x}_M)$ for all $\mathbf{x} \in \mathscr{F} \setminus \{\mathbf{x}_M\}$.

We say that a function $f : \mathscr{F} \subset \mathbb{R}^n \to \mathbb{R}$ has a *local maximum* at a point $\mathbf{x}_M \in \mathscr{F}$ provided that there exists an $r > 0$ such that

$$f(\mathbf{x}) \le f(\mathbf{x}_M) \qquad \text{for all } \mathbf{x} \in \mathscr{F} \cap \mathbf{B}(\mathbf{x}_M, r).$$

It has a *strict local maximum* at $\mathbf{x}_M$ provided that there exists an $r > 0$ such that

$$f(\mathbf{x}) < f(\mathbf{x}_M) \qquad \text{for all } \mathbf{x} \in \mathscr{F} \cap \mathbf{B}(\mathbf{x}_M, r) \setminus \{\mathbf{x}_M\}.$$

If $\mathbf{x}_M \in \operatorname{int}(\mathscr{F})$ is a (local) maximizer, then f is said to have an *unconstrained (local) maximum* at $\mathbf{x}_M$.

The *minimum*, *global minimum*, *minimizer*, *local minimum*, *strict local minimum*, and *unconstrained local minimum* can be defined in a similar manner.

We say that a function f has an *extremum* at a point $\mathbf{x}^*$ provided that f has either a maximum or a minimum at $\mathbf{x}^*$. In the same way, we say that a function f has a *local extremum* at a point $\mathbf{x}^*$ provided that f has either a local maximum or local minimum at $\mathbf{x}^*$.

2.2.1. First Derivative Conditions

In this section, we concentrate on extremizers that are in the interior of the domain and not on the boundary. We show below that such an extremizer must be a point where either the derivative is equal to zero or the function is not differentiable.

Definition. For a continuous function $f : \mathscr{F} \to \mathbb{R}$, $\mathbf{x}_c$ is a *critical point* of f provided that either (i) $Df(\mathbf{x}_c) = \mathbf{0}$ or (ii) f is not differentiable at $\mathbf{x}_c$. (We will treat points on the boundary or end points separately and do not call them critical points.)

Theorem 2.26. *If $f : \mathscr{F} \subset \mathbb{R}^n \to \mathbb{R}$ is continuous on $\mathscr{F}$ and f has a unconstrained local extremum at $\mathbf{x}^* \in \operatorname{int}(\mathscr{F})$, then $\mathbf{x}^*$ is a critical point of f, i.e., either (i) f is not differentiable at $\mathbf{x}^*$ or (ii) f is differentiable at $\mathbf{x}^*$ and $Df(\mathbf{x}^*) = \mathbf{0}$.*

Proof. We prove the contrapositive: We assume that the point $\mathbf{x}^*$ is not a critical point and prove that the function does not have a maximum nor a minimum at $\mathbf{x}^*$. We are assuming that the column vector (the gradient) exists and $\mathbf{v} = Df(\mathbf{x}^*)^{\mathsf{T}} \neq \mathbf{0}$. We consider the values of f along the line $\mathbf{x}_t = \mathbf{x}^* + t\mathbf{v}$ (in the direction of the gradient) using the remainder of the affine approximation:

$$\begin{aligned} f(\mathbf{x}_t) &= f(\mathbf{x}^*) + Df(\mathbf{x}^*)\,(t\mathbf{v}) + \widetilde{R}_1(\mathbf{x}^*, \mathbf{x}_t)\,\|t\mathbf{v}\| \\ &= f(\mathbf{x}^*) + \mathbf{v}^{\mathsf{T}}(t\mathbf{v}) + \widetilde{R}_1(\mathbf{x}^*, \mathbf{x}_t)\,\|t\mathbf{v}\| \\ &= f(\mathbf{x}^*) + t\left[\|\mathbf{v}\|^2 + \widetilde{R}_1(\mathbf{x}^*, \mathbf{x}_t)\|\mathbf{v}\|\operatorname{sign}(t)\right] \\ &\begin{cases} < f(\mathbf{x}^*) & \text{if } t < 0 \text{ and } t \text{ small enough so that } |\widetilde{R}_1| < \frac{1}{2}\|\mathbf{v}\| \\ > f(\mathbf{x}^*) & \text{if } t > 0 \text{ and } t \text{ small enough so that } |\widetilde{R}_1| < \frac{1}{2}\|\mathbf{v}\|. \end{cases} \end{aligned}$$

This proves that $\mathbf{x}^*$ is neither a maximum nor a minimum, i.e., it is not an extreme point. What we have shown is that if the gradient is nonzero, then the function is decreasing in the direction of the negative gradient and increasing in the direction of the gradient. □

2.2.2. Second Derivative Conditions

In this section, we give conditions on the second derivative that insure that a critical point is a local extremizer. At a critical point $\mathbf{x}^*$, $Df(\mathbf{x}^*) = \mathbf{0}$, so

$$f(\mathbf{x}) - f(\mathbf{x}^*) = \tfrac{1}{2}\,(\mathbf{x} - \mathbf{x}^*)^{\mathsf{T}}\, D^2 f(\mathbf{x}^*)\,(\mathbf{x} - \mathbf{x}^*) + \widetilde{R}_2(\mathbf{x}^*, \mathbf{x})\,\|\mathbf{x} - \mathbf{x}^*\|^2.$$

For $\mathbf{x}$ near $\mathbf{x}^*$, $\widetilde{R}_2$ is small, and the term involving the second derivative dominates and determines whether the right hand side of the last equation is positive or negative.

Theorem 2.27. *Suppose that $f : \mathscr{F} \subset \mathbb{R}^n \to \mathbb{R}$ is C^2 on $\operatorname{int}(\mathscr{F})$ and $\mathbf{x}^* \in \operatorname{int}(\mathscr{F})$.*

a. *If f has a local minimum (resp. local maximum) at $\mathbf{x}^*$, then $D^2f(\mathbf{x}^*)$ is positive semidefinite (resp. negative semidefinite).*
b. *If $Df(\mathbf{x}^*) = \mathbf{0}$ and $D^2f(\mathbf{x}^*)$ is positive definite (resp. negative definite), then f has a strict local minimum (resp. strict local maximum) at $\mathbf{x}^*$.*
c. *If $Df(\mathbf{x}^*) = \mathbf{0}$ and $D^2f(\mathbf{x}^*)$ is indefinite, then $\mathbf{x}^*$ is not an extreme point for f.*

Proof. (**a**) If f has a local minimum at $\mathbf{x}^*$, then it is a critical point. Use the second order Taylor's expansion:

$$f(\mathbf{x}) = f(\mathbf{x}^*) + \tfrac{1}{2}\,(\mathbf{x}-\mathbf{x}^*)^{\mathsf{T}}\, D^2f(\mathbf{x}^*)\;(\mathbf{x}-\mathbf{x}^*) + \widetilde{R}_2(\mathbf{x}^*,\mathbf{x})\,\|\mathbf{x}-\mathbf{x}^*\|^2.$$

Assume the conclusion is false and there is a direction $\mathbf{v}$ such that $\mathbf{v}^{\mathsf{T}}\, D^2f(\mathbf{x}^*)\,\mathbf{v} < 0$. Let $\mathbf{x}_t = \mathbf{x}^* + t\,\mathbf{v}$, so $\mathbf{x}_t - \mathbf{x}^* = t\,\mathbf{v}$. Then

$$\begin{aligned} f(\mathbf{x}) &= f(\mathbf{x}^*) + \frac{t^2}{2}\,\mathbf{v}^{\mathsf{T}}\, D^2f(\mathbf{x}^*)\,\mathbf{v} + \widetilde{R}_2(\mathbf{x}^*,\mathbf{x}_t)\,t^2\|\mathbf{v}\|^2 \\ &= f(\mathbf{x}^*) + t^2\,\left[\tfrac{1}{2}\mathbf{v}^{\mathsf{T}}\, D^2f(\mathbf{x}^*)\,\mathbf{v} + \widetilde{R}_2(\mathbf{x}^*,\mathbf{x}_t)\|\mathbf{v}\|^2\right] \\ &< f(\mathbf{x}^*) \qquad \text{for } t \neq 0 \text{ small enough so that } \widetilde{R}_2(\mathbf{x}^*,\mathbf{x}_t)\|\mathbf{v}\|^2 < -\tfrac{1}{2}\mathbf{v}^{\mathsf{T}}\, D^2f(\mathbf{x}^*)\,\mathbf{v}. \end{aligned}$$

This implies that f would not have a local minimum at $\mathbf{x}^*$. Thus, $D^2f(\mathbf{x}^*)$ must be positive semidefinite.

(**b**) Assume that $D^2f(\mathbf{x}^*)$ is positive definite. The set $\{\,\mathbf{u} : \|\mathbf{u}\| = 1\,\}$ is compact, so

$$m = \min_{\|\mathbf{u}\|=1} \mathbf{u}^{\mathsf{T}} D^2f(\mathbf{x}^*)\,\mathbf{u} > 0.$$

For a $\mathbf{x}$ near $\mathbf{x}^*$, letting $\mathbf{v} = \mathbf{x} - \mathbf{x}^*$ and $\mathbf{u} = \dfrac{1}{\|\mathbf{v}\|}\,\mathbf{v}$,

$$\begin{aligned} (\mathbf{x}-\mathbf{x}^*)^{\mathsf{T}}\, D^2f(\mathbf{x}^*)\;(\mathbf{x}-\mathbf{x}^*) &= (\|\mathbf{v}\|\mathbf{u})^{\mathsf{T}} D^2f(\mathbf{x}^*)\,(\|\mathbf{v}\|\mathbf{u}) \\ &= \|\mathbf{v}\|^2\mathbf{u}^{\mathsf{T}} D^2f(\mathbf{x}^*)\,|\mathbf{u} \geq m\,\|\mathbf{x}-\mathbf{x}^*\|^2. \end{aligned}$$

Since $Df(\mathbf{x}^*) = \mathbf{0}$, the Taylor's expansion with two terms is as follows:

$$f(\mathbf{x}) = f(\mathbf{x}^*) + \tfrac{1}{2}\,(\mathbf{x}-\mathbf{x}^*)^{\mathsf{T}}\, D^2f(\mathbf{x}^*)\;(\mathbf{x}-\mathbf{x}^*) + \widetilde{R}_2(\mathbf{x}^*,\mathbf{x})\,\|\mathbf{x}-\mathbf{x}^*\|^2.$$

There exists a $\delta > 0$ such that

$$|\widetilde{R}_2(\mathbf{x}^*,\mathbf{x})| < \tfrac{1}{4}m \qquad \text{for } \|\mathbf{x}-\mathbf{x}^*\| < \delta.$$

Then, for $\delta > \|\mathbf{x}-\mathbf{x}^*\| > 0$,

$$\begin{aligned} f(\mathbf{x}) &> f(\mathbf{x}^*) + \tfrac{1}{2}m\,\|\mathbf{x}-\mathbf{x}^*\|^2 - \tfrac{1}{4}m\,\|\mathbf{x}-\mathbf{x}^*\|^2 \\ &= f(\mathbf{x}^*) + \tfrac{1}{4}m\,\|\mathbf{x}-\mathbf{x}^*\|^2 \\ &> f(\mathbf{x}^*). \end{aligned}$$

This shows that $\mathbf{x}^*$ is a strict local minimum.

The proof of (**c**) is similar. □

Example 2.28. For the function $F(x,y,z) = 3x^2y + y^3 - 3x^2 - 3y^2 + z^3 - 3z$, find the critical points and classify them as local maximum, local minimum, or neither.

The equations for a critical point are

$$0 = \frac{\partial F}{\partial x} = 6xy - 6x = 6x(y-1)$$
$$0 = \frac{\partial F}{\partial y} = 3x^2 + 3y^2 - 6y$$
$$0 = \frac{\partial F}{\partial z} = 3z^2 - 3.$$

From the third equation $z = \pm 1$. From the first equation, $x = 0$ or $y = 1$. If $x = 0$, then the second equation gives $0 = 3y(y-2)$, $y = 0$ or $y = 2$. Thus we have the points $(0, 0, \pm 1)$ and $(0, 2, \pm 1)$. If $y = 1$ from the first equation, then the second equation becomes $0 = 3x^2 - 3$ and $x = \pm 1$. Thus we have the points $(\pm 1, 1, \pm 1)$. Thus, all the critical points are $(0, 0, \pm 1)$, $(0, 2, \pm 1)$, and $(\pm 1, 1, \pm 1)$.

The second derivative of F is

$$D^2F(x, y, z) = \begin{bmatrix} 6y-6 & 6x & 0 \\ 6x & 6y-6 & 0 \\ 0 & 0 & 6z \end{bmatrix}.$$

At the critical points

$$D^2F(0, 0, \pm 1) = \begin{bmatrix} -6 & 0 & 0 \\ 0 & -6 & 0 \\ 0 & 0 & \pm 6 \end{bmatrix}, \qquad D^2F(0, 2, \pm 1) = \begin{bmatrix} 6 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & \pm 6 \end{bmatrix},$$
$$D^2F(\pm 1, 1, \pm 1) = \begin{bmatrix} 0 & \pm 6 & 0 \\ \pm 6 & 0 & 0 \\ 0 & 0 & \pm 6 \end{bmatrix}.$$

Let $\Delta_k = \det(\mathbf{A}_k)$ be the determinants of the principal submatrices. For this example,

$$\Delta_1 = F_{xx} = 6y - 6,$$
$$\Delta_2 = F_{xx}F_{yy} - F_{xy}^2 = (6y-6)^2 - 36x^2, \qquad \text{and}$$
$$\Delta_3 = F_{zz}\,\Delta_2 = 6z\,\Delta_2.$$

(x, y, z)	$\Delta_1 = F_{xx}$	F_{yy}	F_{xy}	Δ_2	F_{zz}	Δ_3	Type
$(0, 0, 1)$	-6	-6	0	36	6	216	saddle
$(0, 0, -1)$	-6	-6	0	36	-6	-216	local max
$(0, 2, 1)$	6	6	0	36	6	216	local min
$(0, 2, -1)$	6	6	0	36	-6	-216	saddle
$(\pm 1, 1, \pm 1)$	0	0	± 6	-36	± 6	∓ 216	saddle

Therefore, $(0, 0, -1)$ is a local maximum, $(0, 2, 1)$ is a local minimum, and the other points are neither. ■

2.2. Exercises

2.2.1. Find the points at which each of the following functions attains a maximum and minimum on the interval $0 \le x \le 3$. For parts (a) and (b), also find the maximal and minimal values. *Remember* to consider the end points of the interval $[0, 3]$.

a. $f(x) = x^2 - 2x + 2$.

b. $g(x) = -x^2 + 2x + 4$.

c. The function $h(x)$ satisfies $h'(x) > 0$ for all $0 \le x \le 3$.

d. The function $k(x)$ satisfies $k'(x) < 0$ for all $0 \leq x \leq 3$.
e. The function $u(x)$ satisfies $u'(x) = 0$ for all $0 \leq x \leq 3$.

2.2.2. Consider the function $f(x) = \dfrac{x}{1+x^2}$.
a. What are the critical points of $f(x)$?
b. Are the critical points global maximum or minimum?

2.2.3. Find the critical points of the following functions.
a. $f(x,y) = 2xy - 2x^2 - 5y^2 + 4y - 3$.
b. $f(x,y) = x^2 - y^3 - x^2y + y$.
c. $f(x,y) = xy + \dfrac{8}{x} + \dfrac{1}{y}$.

2.2.4. Suppose $f : \mathscr{F} \subset \mathbb{R}^n \to \mathbb{R}$ is C^1 and has a maximum at a point $\mathbf{x}^*$ on the boundary of $\mathscr{F}$. Does $Df(\mathbf{x}^*)$ have to equal $\mathbf{0}$ at $\mathbf{x}^*$? Give an explanation or a counter-example.

2.2.5. Find all the critical points and classify them as local maximum, local minimum, or saddle points for the following functions.
a. $f(x,y,z) = x^4 + x^2 - 6xy + 3y^2 + z^2$
b. $f(x,y,z) = 3x - x^3 - 2y^2 + y^4 + z^3 - 3z$
c. $f(x,y,z) = 3x^2y + y^3 - 3x^2 - 3y^2 + z^3 - 3z$.
d. $f(x,y,z) = 3x^2y - 12xy + y^3 + z^4 - 2z^2$.
e. $f(x,y,z) = 3x^2 + x^3 + y^2 + xy^2 + z^3 - 3z$.
f. $f(x,y,z) = -\frac{1}{2}\,x^2 - 5y^2 - 2z^2 + 3xy - yz + 6x + 5y + 20z$

2.2.6. For the feasible set $\mathscr{F} = \mathbb{R}^2_+ = \{\,(x,y) : x,y \geq 0\,\}$, consider the function

$$f(x,y) = \frac{2x+y}{4x^2+y^2+8}.$$

Its first order partial derivatives are as follows:

$$\frac{\partial f}{\partial x}(x,y) = \frac{16 - 8x^2 + 2y^2 - 8xy}{(4x^2+y^2+8)^2},$$
$$\frac{\partial f}{\partial y}(x,y) = \frac{8 + 4x^2 - y^2 - 4xy}{(4x^2+y^2+8)^2}.$$

a. Find the one critical point $(\bar{x},\bar{y})$ in the interior of $\mathbb{R}^2_+$.
Hint: Show that $\bar{y} = 2\bar{x}$.
b. Classify the critical point $(\bar{x},\bar{y})$ as a strict local maximum, strict local minimum, or neither. You may use the fact that at the critical point the second order partial derivatives are as follows:

$$\frac{\partial^2 f}{\partial x^2}(\bar{x},\bar{y}) = \frac{-16\bar{x} - 8\bar{y}}{(1+\bar{x}^2+\bar{y}^2)^2}, \qquad \frac{\partial^2 f}{\partial x \partial y}(\bar{x},\bar{y}) = \frac{4\bar{y} - 8\bar{x}}{(1+\bar{x}^2+\bar{y}^2)^2},$$
$$\frac{\partial^2 f}{\partial y^2}(\bar{x},\bar{y}) = \frac{-4\bar{x} - 2\bar{y}}{(1+\bar{x}^2+\bar{y}^2)^2}.$$

c. Find the critical point of $g(x) = f(x,0) = {}^{x}\!/_{(2x^2+4)}$ for $x \geq 0$, i.e., along the boundary $y = 0$. Classify this critical point of g.
d. Find and classify the critical point of $h(y) = f(0,y) = {}^{y}\!/_{(y^2+8)}$ for $y \geq 0$.
e. Find the maximal value and the point which gives the maximum. Hint: It can be shown that f attains a maximum on $\mathbb{R}^2_+$. (See Exercise 2.1.6.) The

maximizer must be among the points found in parts (a), (c), (d), and the origin $(0, 0)$.

2.2.7. A firm produces a single output Q determined by the Cobb-Douglas production function of two inputs q_1 and q_2, $Q = q_1^{\frac{1}{3}} q_2^{\frac{1}{2}}$. Let p_1 be the price of q_1, and p_2 be the price of q_2; let the price of the output Q be one. (Either the price of Q is taken as the unit of money or the p_j are the ratios of the prices of q_j to the price of Q.) The profit is given by $\pi = Q - p_1 q_1 - p_2 q_2$.

- **a.** Considering the inputs as variables, $(q_1, q_2) \in \mathbb{R}^2_{++}$, show that there is a unique critical point $q_1^*, q_2^*)$.
- **b.** Show that the critical point is a local maximum.

2. Exercises for Chapter 2

2.1. Indicate which of the following statements are *true* and which are *false*. Justify each answer: For a true statement explain why it is true and for a false statement either indicate how to make it true or indicate why the statement is false.

- **a.** If $\lim_{\mathbf{x}\to\mathbf{a}} f(\mathbf{x})$ exists and is finite, then f is continuous at $\mathbf{x} = \mathbf{a}$.
- **b.** If $\mathbf{f} : \mathbb{R}^3 \to \mathbb{R}^2$ is differentiable at $\mathbf{x} = \mathbf{a}$, then $\mathbf{f}$ is continuous at $\mathbf{x} = \mathbf{a}$.
- **c.** $\mathbf{f} : \mathbb{R}^3 \to \mathbb{R}^2$ is continuous at $\mathbf{x} = \mathbf{a}$, then $\mathbf{f}$ is differentiable at $\mathbf{x} = \mathbf{a}$.
- **d.** If $f(x, y, z)$ is a C^1 function on $\mathbb{R}^3$, then the extrema of f on the set $x^2 + 4y^2 + 9z^2 \leq 100$ must either be a critical point in $x^2 + 4y^2 + 9z^2 < 100$ or an extrema of f on the boundary $x^2 + 4y^2 + 9z^2 = 100$.
- **e.** If $\nabla f(a_1, \ldots, a_n) = \mathbf{0}$, then f has a local extremum at $\mathbf{a} = (a_1, \ldots, a_n)$.
- **f.** A continuous function $f(x, y)$ must attain a maximum on the disk $\{\, (x, y) : x^2 + y^2 < 1. \,\}$.
- **g.** If $\det D^2 f(\mathbf{a}) = 0$, then f has a saddle point at $\mathbf{a}$.
- **h.** For a C^2 function $f : \mathbb{R}^n \to \mathbb{R}$ with a critical point $\mathbf{x}^*$ at which the second derivative (or Hessian matrix) $D^2 f(\mathbf{x}^*)$ is negative definite, then $\mathbf{x}^*$ is a global maximizer of f on $\mathbb{R}^n$.

Chapter 3

Constrained Extrema

This chapter concerns the problem of find the extrema of a function on a feasible set defined by a number of constraints. When there are equality constraints, the Implicit Function Theorem gives conditions on the derivatives to ensure that some of the variables locally can be considered as functions of the other variables. This theorem makes use of the derivative as a matrix in an explicit manner and is the mathematical basis for implicit differentiation, which we illustrate with examples from comparative statics within economics. The Implicit Function Theorem is also used in the proofs of subsequent topics, including in the derivation of the Lagrange Multiplier Theorem that concerns optimization problems for which the constraint functions are set equal to constants. The last two sections concern nonlinear optimization problems where the feasible set is defined by inequality constraints.

3.1. Implicit Function Theorem

For a real valued function $g : \mathbb{R}^n \to \mathbb{R}$, if $g(\mathbf{x}^*) = b$ and $\nabla g(\mathbf{x}^*) \neq \mathbf{0}$, then the level set $g^{-1}(b) = \{\, \mathbf{x} \in \mathbb{R}^n : g(\mathbf{x}) = b \,\}$ is locally a graph near $\mathbf{x}^*$. This fact can be understood in terms of the tangent plane to the level surface that is the set of vectors perpendicular to the gradient, $0 = \nabla g(\mathbf{x}^*) \cdot (\mathbf{x} - \mathbf{x}^*)$. If $\dfrac{\partial g}{\partial x_n}(\mathbf{x}^*) \neq 0$, then the tangent plane is the graph of the variable x_n in terms of the other variables:

$$\begin{aligned} 0 &= \nabla g(\mathbf{x}^*) \cdot (\mathbf{x} - \mathbf{x}^*) \\ &= \frac{\partial g}{\partial x_1}(\mathbf{x}^*)(x_1 - x_1^*) + \cdots + \frac{\partial g}{\partial x_{n-1}}(\mathbf{x}^*)(x_{n-1} - x_{n-1}^*) + \frac{\partial g}{\partial x_n}(\mathbf{x}^*)(x_n - x_n^*), \text{ so} \\ x_n &= x_n^* - \left(\frac{\dfrac{\partial g}{\partial x_1}(\mathbf{x}^*)}{\dfrac{\partial g}{\partial x_n}(\mathbf{x}^*)} \right)(x_1 - x_1^*) - \cdots - \left(\frac{\dfrac{\partial g}{\partial x_{n-1}}(\mathbf{x}^*)}{\dfrac{\partial g}{\partial x_n}(\mathbf{x}^*)} \right)(x_{n-1} - x_{n-1}^*). \end{aligned}$$

The implicit function says that the nonlinear level set is also locally a graph, with x_n determined by the other variables. If a different variable x_m has $\dfrac{\partial g}{\partial x_m}(\mathbf{x}^*) \neq 0$, then the same type of argument shows that the nonlinear level set locally determines x_m as a function of the other variables.

Now we turn the case of a vector constraint or several scalar constraints. Consider a C^1 $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^k$ for $k > 1$ with coordinate functions g_i. For a constant $\mathbf{b} \in \mathbb{R}^k$, the level set where is the function takes on the values $\mathbf{b}$ is denoted by

$$\mathbf{g}^{-1}(\mathbf{b}) = \{\, \mathbf{x} \in \mathbb{R}^n : g_i(\mathbf{x}) = b_i \text{ for } i = 1, \ldots, k \,\}.$$

For $\mathbf{x}^* \in \mathbf{g}^{-1}(\mathbf{b})$, the assumption of nonzero gradient for a scalar function is replaced by the assumption is that the rank of $D\mathbf{g}(\mathbf{x}^*)$ is equal to k, i.e., the gradients $\{\nabla g_i(\mathbf{x}^*)\}_{i=1}^k$ are linearly independent. Then, it is possible to select k variables $x_{m_1}, \ldots, x_{m_k}$ such that

$$\det \begin{pmatrix} \dfrac{\partial g_1}{\partial x_{m_1}}(\mathbf{x}^*) & \cdots & \dfrac{\partial g_1}{\partial x_{m_k}}(\mathbf{x}^*) \\ \vdots & \ddots & \vdots \\ \dfrac{\partial g_k}{\partial x_{m_1}}(\mathbf{x}^*) & \cdots & \dfrac{\partial g_k}{\partial x_{m_k}}(\mathbf{x}^*) \end{pmatrix} \neq 0. \tag{3}$$

We show below that if (3) holds, then the null space of $D\mathbf{g}(\mathbf{x}^*)$ is the graph of the variables $\mathbf{z} = (x_{m_1}, \ldots, x_{m_k})$ in terms of the other $n - k$ variables $\mathbf{w} = (x_{\ell_1}, \ldots, x_{\ell_{n-k}})$. The Implicit Function states that the nonlinear level set is also locally a graph with the $\mathbf{z} = (x_{m_1}, \ldots, x_{m_k})$ determined implicitly as functions of the other variables $\mathbf{w} = (x_{\ell_1}, \ldots, x_{\ell_{n-k}})$.

Theorem 3.1 (Implicit Function Theorem). *Assume there are k C^1 constraint functions, $g_i : \mathbb{R}^n \to \mathbb{R}$ for $1 \leq i \leq k$ with $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), \ldots, g_k(\mathbf{x}))^\mathsf{T}$, such that $\mathbf{g}(\mathbf{x}^*) = \mathbf{b}$ and $\operatorname{rank}(D\mathbf{g}(\mathbf{x}^*)) = k$, i.e., the gradients $\{\nabla g_i(\mathbf{x}^*)\}_{i=1}^k$ are linearly independent.*

Then, the nonlinear level set $\mathbf{g}^{-1}(\mathbf{b})$ is locally a graph near $\mathbf{x}^*$.

If $\mathbf{z} = (x_{m_1}, \ldots, x_{m_k})^\mathsf{T}$ are k variables such that inequality (3) holds, then they are locally determined implicitly as functions of the other $n-k$ variables, $\mathbf{w} = (x_{\ell_1}, \ldots, x_{\ell_{n-k}})^\mathsf{T}$. This implicitly defined function $\mathbf{z} = h(\mathbf{w})$ is as differentiable as $\mathbf{g}$. The partial derivatives of the coordinate functions of $\mathbf{z} = \mathbf{h}(\mathbf{w})$, $\dfrac{\partial h_q}{\partial x_{\ell_j}} = \dfrac{\partial x_{m_q}}{\partial x_{\ell_j}}$, can be calculated by the chain rule and satisfy

$$0 = \frac{\partial g_i}{\partial x_{\ell_j}} + \sum_{q=1}^{k} \frac{\partial g_i}{\partial x_{m_q}} \frac{\partial x_{m_q}}{\partial x_{\ell_j}},$$

or in matrix notation

$$\mathbf{0} = \begin{bmatrix} \dfrac{\partial g_1}{\partial x_{\ell_1}} & \cdots & \dfrac{\partial g_1}{\partial x_{\ell_{n-k}}} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial g_k}{\partial x_{\ell_1}} & \cdots & \dfrac{\partial g_k}{\partial x_{\ell_{n-k}}} \end{bmatrix} + \begin{bmatrix} \dfrac{\partial g_1}{\partial x_{m_1}} & \cdots & \dfrac{\partial g_1}{\partial x_{m_k}} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial g_k}{\partial x_{m_1}} & \cdots & \dfrac{\partial g_k}{\partial x_{m_k}} \end{bmatrix} \begin{bmatrix} \dfrac{\partial x_{m_1}}{\partial x_{\ell_1}} & \cdots & \dfrac{\partial x_{m_1}}{\partial x_{\ell_{n-k}}} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial x_{m_k}}{\partial x_{\ell_1}} & \cdots & \dfrac{\partial x_{m_k}}{\partial x_{\ell_{n-k}}} \end{bmatrix} \tag{ImD}$$

Remark. The hard part of the proof is showing that the function h exists and is differentiable, even though there is no explicit formula for it. Once this is known, its derivative can be calculated by the chain rule and solving the matrix equation (ImD). For a proof, see Wade [**15**] or Rudin [**10**].

Take the grouping of variables as given in the theorem: $\mathbf{z} = (x_{m_1}, \ldots, x_{m_k})^\mathsf{T}$ are k variables such that equation (3) holds, and $\mathbf{w} = (x_{\ell_1}, \ldots, x_{\ell_{n-k}})^\mathsf{T}$ are the other $n - k$ variables. We write $\mathbf{x} = (\mathbf{w}, \mathbf{z})$, as if the submatrix of the last k columns has full rank k. (We could

also relabel the variables so this was true.) Use the following notation for the matrix of partial derivatives with respect to only the $\mathbf{w}$ variables or only the $\mathbf{w}$ variables,

$$D_{\mathbf{w}}\mathbf{g}(\mathbf{x}^*) = \left(\frac{\partial g_i}{\partial x_{\ell_j}}(\mathbf{x}^*)\right)_{1\le i\le k,\ 1\le j\le n-k} \qquad \text{and}$$
$$D_{\mathbf{z}}\mathbf{g}(\mathbf{x}^*) = \left(\frac{\partial g_i}{\partial x_{m_j}}(\mathbf{x}^*)\right)_{1\le i\le k,\ 1\le j\le k}$$

Using this notation, $\operatorname{rank}(D_{\mathbf{z}}\mathbf{g}(\mathbf{x}^*)) = k$ or $\det\left(D_{\mathbf{z}}\mathbf{g}(\mathbf{x}^*)\right) \neq 0$ so $D_{\mathbf{z}}\mathbf{g}(\mathbf{x}^*)$ is invertible. Using this notation, the matrix equation (ImD) in the theorem becomes

$$\mathbf{0} = D_{\mathbf{w}}\mathbf{g}(\mathbf{x}^*) + D_{\mathbf{z}}\mathbf{g}(\mathbf{x}^*)\, D\mathbf{h}(\mathbf{w}^*), \qquad \text{so}$$
$$D\mathbf{h}(\mathbf{w}^*) = \text{-}\left(D_{\mathbf{z}}\mathbf{g}(\mathbf{x}^*)\right)^{\text{-}1} D_{\mathbf{w}}\mathbf{g}(\mathbf{x}^*).$$

In a given example, we write down the equation (ImD), and then solve it for the derivative $D\mathbf{h}(\mathbf{z}^*)$, or just the desired partial derivatives. Notice that (i) the matrix $D_{\mathbf{z}}\mathbf{g}(\mathbf{x}^*)$ includes all the partial derivatives with respect to the dependent variable used to calculate the nonzero determinant and (ii) the matrix $D_{\mathbf{w}}\mathbf{g}(\mathbf{x}^*)$ includes all the partial derivatives with respect to the independent (other) variables.

Before giving examples, we want to note that the assumption of the Implicit Function Theorem is exactly what makes the null space of $D\mathbf{g}(\mathbf{x}^*)$ a graph of the $\mathbf{z}$-coordinates in terms of the $\mathbf{w}$-coordinates. The *null space* of $D\mathbf{g}(\mathbf{x}^*)$ is given by

$$\operatorname{null}\left(D\mathbf{g}(\mathbf{x}^*)\right) = \{\mathbf{v}\in\mathbb{R}^n : D\mathbf{g}(\mathbf{x}^*)\mathbf{v} = \mathbf{0}\} = \{\mathbf{v}\in\mathbb{R}^n : \mathbf{v}\cdot\nabla g_i(\mathbf{x}^*) = 0 \text{ for } 1\le i\le k\}.$$

If $\mathbf{v}$ is in this null space, then we can split it up into the components in the $\mathbf{w}$ and $\mathbf{z}$ directions, $\mathbf{v} = \begin{pmatrix}\mathbf{v}_{\mathbf{w}}\\ \mathbf{v}_{\mathbf{z}}\end{pmatrix}$, where we use notation as if the $\mathbf{z}$-variables were in the last k components.

$$\mathbf{0} = D\mathbf{g}(\mathbf{x}^*)\mathbf{v} = [D_{\mathbf{w}}\mathbf{g}(\mathbf{x}^*), D_{\mathbf{z}}\mathbf{g}(\mathbf{x}^*)]\begin{pmatrix}\mathbf{v}_{\mathbf{w}}\\ \mathbf{v}_{\mathbf{z}}\end{pmatrix} = D_{\mathbf{w}}\mathbf{g}(\mathbf{x}^*)\mathbf{v}_{\mathbf{w}} + D_{\mathbf{z}}\mathbf{g}(\mathbf{x}^*)\mathbf{v}_{\mathbf{z}} \qquad \text{and}$$

$$\mathbf{v}_{\mathbf{z}} = \text{-}\left(D_{\mathbf{z}}\mathbf{g}(\mathbf{x}^*)\right)^{\text{-}1} D_{\mathbf{w}}\mathbf{g}(\mathbf{x}^*)\mathbf{v}_{\mathbf{w}}.$$

The Implicit Functions says that if this is possible at the linear level, then the nonlinear level set is also locally a differentiable graph of these same variable in terms of the other variables.

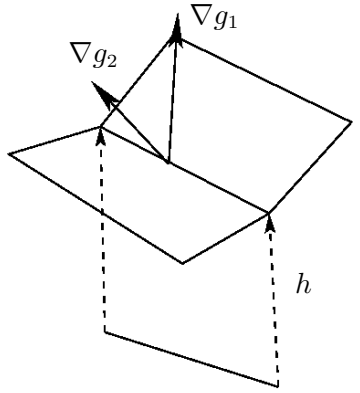


Figure 3.1.1. Null space (and nonlinear level set) as a graph

Example 3.2 (Changing Technology for Production). This example is based on Section 6.6 of [**4**]. A firm uses two inputs to produce a single output. Assume the amounts of the inputs are x and y with p the price of x and q the price of y. The amount produced Q is assumed to be determined by the Cobb-Douglas production function $Q = x^a y^b$, where the technology determines the exponents a and b. By changing the technology, the firm can vary the exponent b while keeping a fixed or vary a while keeping b fixed. It wants to keep the amount produced fixed, Q_0, and the cost of the inputs fixed, $px + qy = 125$. What is the rate of change of the amounts of inputs as functions of a and b at $x = 5$, $y = 50$, $p = 5$, $q = 2$, $a = 1/3$, and $b = 2/3$?

Rather than use the equation $Q_0 = x^a y^b$, we take its logarithm and obtain the two equations

$$\begin{aligned} g_1(x, y, a, b, p, q) &= px + qy = 125 \qquad \text{and} \\ g_2(x, y, a, b, p, q) &= a\ln(x) + b\ln(y) = \ln(Q_0). \end{aligned} \tag{4}$$

These two equations define x and y as functions of a, b, p, and q since equation (3) is

$$\begin{aligned} \det \begin{bmatrix} \dfrac{\partial g_1}{\partial x} & \dfrac{\partial g_1}{\partial y} \\ \dfrac{\partial g_2}{\partial x} & \dfrac{\partial g_2}{\partial y} \end{bmatrix} &= \det \begin{bmatrix} p & q \\ \frac{a}{x} & \frac{b}{y} \end{bmatrix} \\ &= \frac{pb}{y} - \frac{qa}{x} = \frac{pbx - qay}{xy} = \frac{5 \cdot 2 \cdot 5 - 2 \cdot 1 \cdot 50}{3 \cdot 5 \cdot 50} = \frac{-1}{15} \neq 0. \end{aligned}$$

Considering x and y as functions of a, b, p, and q, and differentiating the two equations with respect to the four independent variables gives the following matrix equation (ImD):

$$\begin{aligned} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} &= \begin{bmatrix} \dfrac{\partial g_1}{\partial a} & \dfrac{\partial g_1}{\partial b} & \dfrac{\partial g_1}{\partial p} & \dfrac{\partial g_1}{\partial q} \\ \dfrac{\partial g_2}{\partial a} & \dfrac{\partial g_2}{\partial b} & \dfrac{\partial g_2}{\partial p} & \dfrac{\partial g_2}{\partial q} \end{bmatrix} + \begin{bmatrix} \dfrac{\partial g_1}{\partial x} & \dfrac{\partial g_1}{\partial y} \\ \dfrac{\partial g_2}{\partial x} & \dfrac{\partial g_2}{\partial y} \end{bmatrix} \begin{bmatrix} \dfrac{\partial x}{\partial a} & \dfrac{\partial x}{\partial b} & \dfrac{\partial x}{\partial p} & \dfrac{\partial x}{\partial q} \\ \dfrac{\partial y}{\partial a} & \dfrac{\partial y}{\partial b} & \dfrac{\partial y}{\partial p} & \dfrac{\partial y}{\partial q} \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 & x & y \\ \ln(x) & \ln(y) & 0 & 0 \end{bmatrix} + \begin{bmatrix} p & q \\ \frac{a}{x} & \frac{b}{y} \end{bmatrix} \begin{bmatrix} \dfrac{\partial x}{\partial a} & \dfrac{\partial x}{\partial b} & \dfrac{\partial x}{\partial p} & \dfrac{\partial x}{\partial q} \\ \dfrac{\partial y}{\partial a} & \dfrac{\partial y}{\partial b} & \dfrac{\partial y}{\partial p} & \dfrac{\partial y}{\partial q} \end{bmatrix}, \end{aligned}$$

so

$$\begin{aligned} \begin{bmatrix} \dfrac{\partial x}{\partial a} & \dfrac{\partial x}{\partial b} & \dfrac{\partial x}{\partial p} & \dfrac{\partial x}{\partial q} \\ \dfrac{\partial y}{\partial a} & \dfrac{\partial y}{\partial b} & \dfrac{\partial y}{\partial p} & \dfrac{\partial y}{\partial q} \end{bmatrix} &= -\frac{xy}{pbx - qay} \begin{bmatrix} \frac{b}{y} & -q \\ \frac{-a}{x} & p \end{bmatrix} \begin{bmatrix} 0 & 0 & x & y \\ \ln(x) & \ln(y) & 0 & 0 \end{bmatrix} \\ &= \frac{xy}{qay - pbx} \begin{bmatrix} -q\ln(x) & -q\ln(y) & \frac{bx}{y} & b \\ p\ln(x) & p\ln(y) & -a & \frac{-ay}{x} \end{bmatrix} \\ &= \begin{bmatrix} \frac{-xyq\ln(x)}{qay - pbx} & \frac{-xyq\ln(y)}{qay - pbx} & \frac{bx^2}{qay - pbx} & \frac{bxy}{qay - pbx} \\ \frac{xyp\ln(x)}{qay - pbx} & \frac{xyp\ln(y)}{qay - pbx} & \frac{-axy}{qay - pbx} & \frac{-ay^2\ln(y)}{qay - pbx} \end{bmatrix}. \end{aligned}$$

At the point in question, $qay - pbx = \frac{100-50}{3} = \frac{50}{3}$ and

$$\frac{\partial x}{\partial a} = \frac{-3(5)(50)(2)\ \ln(5)}{50} = -30\ \ln(5), \quad \frac{\partial x}{\partial b} = \frac{-3(5)(50)(2)\ \ln(50)}{50} = -30\ \ln(50),$$
$$\frac{\partial y}{\partial a} = \frac{3(5)(50)(5)\ \ln(5)}{50} = 75\ \ln(5), \quad \frac{\partial y}{\partial b} = \frac{3(5)(50)(5)\ \ln(50)}{50} = 75\ \ln(50).$$

■

Steps for Implicit Differentiation with Several Constraints	
1.	Some equations are given (or derived) relating several variables. The last example has two equations with six variables.
2.	Select the same number of variables as the number of equations that you want to be defined in terms of the other variables. Check that the matrix of partial derivatives of the constraint equations with respect to these variables is invertible at least at the point in question. In the previous example, the matrix of partial derivatives with respect to x and y has nonzero determinant.
3.	Thinking of these equations as defining these variables in terms of the others, take partial derivatives of the equations to give the matrix equation (ImD).
4.	Solve (ImD) for the matrix of partial derivatives of the dependent variables with respect to the independent variables.

Example 3.3 (**Marginal Inputs of Prices).** This example is based on Section 7.3 of [**4**]. A firm produces a single output Q determined by the Cobb-Douglas production function of two inputs q_1 and q_2, $Q = q_1^{\frac{1}{3}} q_2^{\frac{1}{2}}$. Let p_1 be the price of q_1, and p_2 be the price of q_2; let the price of the output Q be one. (Either the price of Q is taken as the unit of money or the p_j are the ratios of the prices of q_j to the price of Q.) The profit is given by $\pi = Q - p_1 q_1 - p_2 q_2$. The inputs that are a critical point satisfy

$$0 = \frac{\partial \pi}{\partial q_1} = \frac{1}{3}\, q_1^{-\frac{2}{3}}\, q_2^{\frac{1}{2}} - p_1 = g_1(p_1, p_2, q_1, q_2) \quad \text{and} \tag{5}$$
$$0 = \frac{\partial \pi}{\partial q_2} = \frac{1}{2}\, q_1^{\frac{1}{3}}\, q_2^{-\frac{1}{2}} - p_2 = g_2(p_1, p_2, q_1, q_2),$$

where the function $\mathbf{g}$ is defined by the last two equations.

We have two equations and four variables q_1, q_2, p_1, and p_2 which clearly define the prices in terms of the inputs. We want to show that it also implicitly determines the two inputs q_1 and q_2 in terms of the prices p_1 and p_2. The derivative of $\mathbf{g}$ with respect to $\mathbf{q}$ is

$$\mathbf{D_q g} = \begin{bmatrix} -\frac{2}{9}\, q_1^{-\frac{5}{3}}\, q_2^{\frac{1}{2}} & \frac{1}{6}\, q_1^{-\frac{2}{3}}\, q_2^{-\frac{1}{2}} \\ \frac{1}{6}\, q_1^{-\frac{2}{3}}\, q_2^{-\frac{1}{2}} & -\frac{1}{4}\, q_1^{\frac{1}{3}}\, q_2^{-\frac{3}{2}} \end{bmatrix} \quad \text{with}$$
$$\Delta_2 = \det(D_{\mathbf{q}}\mathbf{g}) = q_1^{-\frac{4}{3}}\, q_2^{-1} \left[\frac{2}{36} - \frac{1}{36}\right]$$
$$= \frac{1}{36}\, q_1^{-\frac{4}{3}}\, q_2^{-1} > 0.$$

The second derivative $D^2\pi = D_{\mathbf{q}}\mathbf{g}$ is negative definite, so this critical point maximize profits. Because $\det(D\mathbf{g}) \neq 0$, the Implicit Function Theorem implies that these two equations (5) implicitly determine the two inputs q_1 and q_2 in terms of the prices p_1 and p_2. Also, the partial

derivatives of q_1 and q_2 with respect to p_1 and p_2 satisfy the following matrix equation (ImD):

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \dfrac{\partial g_1}{\partial p_1} & \dfrac{\partial g_1}{\partial p_2} \\ \dfrac{\partial g_2}{\partial p_1} & \dfrac{\partial g_2}{\partial p_2} \end{bmatrix} + \begin{bmatrix} \dfrac{\partial g_1}{\partial q_1} & \dfrac{\partial g_1}{\partial q_2} \\ \dfrac{\partial g_2}{\partial q_1} & \dfrac{\partial g_2}{\partial q_2} \end{bmatrix} \begin{bmatrix} \dfrac{\partial q_1}{\partial p_1} & \dfrac{\partial q_1}{\partial p_2} \\ \dfrac{\partial q_2}{\partial p_1} & \dfrac{\partial q_2}{\partial p_2} \end{bmatrix}$$

$$= \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} + \begin{bmatrix} -\frac{2}{9}\, q_1^{-\frac{5}{3}}\, q_2^{\frac{1}{2}} & \frac{1}{6}\, q_1^{-\frac{2}{3}}\, q_2^{-\frac{1}{2}} \\ \frac{1}{6}\, q_1^{-\frac{2}{3}}\, q_2^{-\frac{1}{2}} & -\frac{1}{4}\, q_1^{\frac{1}{3}}\, q_2^{-\frac{3}{2}} \end{bmatrix} \begin{bmatrix} \dfrac{\partial q_1}{\partial p_1} & \dfrac{\partial q_1}{\partial p_2} \\ \dfrac{\partial q_2}{\partial p_1} & \dfrac{\partial q_2}{\partial p_2} \end{bmatrix},$$

$$\begin{bmatrix} \dfrac{\partial q_1}{\partial p_1} & \dfrac{\partial q_1}{\partial p_2} \\ \dfrac{\partial q_2}{\partial p_1} & \dfrac{\partial q_2}{\partial p_2} \end{bmatrix} = \begin{bmatrix} -\dfrac{2}{9}\, q_1^{-\frac{5}{3}}\, q_2^{\frac{1}{2}} & \dfrac{1}{6}\, q_1^{-\frac{2}{3}}\, q_2^{-\frac{1}{2}} \\ \dfrac{1}{6}\, q_1^{-\frac{2}{3}}\, q_2^{-\frac{1}{2}} & -\dfrac{1}{4}\, q_1^{\frac{1}{3}}\, q_2^{-\frac{3}{2}} \end{bmatrix}^{-1}$$

$$\begin{bmatrix} \dfrac{\partial q_1}{\partial p_1} & \dfrac{\partial q_1}{\partial p_2} \\ \dfrac{\partial q_2}{\partial p_1} & \dfrac{\partial q_2}{\partial p_2} \end{bmatrix} = 36\, q_1^{\frac{4}{3}} q_2 \begin{bmatrix} -\dfrac{1}{4}\, q_1^{\frac{1}{3}}\, q_2^{-\frac{3}{2}} & -\dfrac{1}{6}\, q_1^{-\frac{2}{3}}\, q_2^{-\frac{1}{2}} \\ -\dfrac{1}{6}\, q_1^{-\frac{2}{3}}\, q_2^{-\frac{1}{2}} & -\dfrac{2}{9}\, q_1^{-\frac{5}{3}}\, q_2^{\frac{1}{2}} \end{bmatrix}$$

$$= \begin{bmatrix} -9\, q_1^{\frac{5}{3}}\, q_2^{-\frac{1}{2}} & -6\, q_1^{\frac{2}{3}}\, q_2^{\frac{1}{2}} \\ -6\, q_1^{\frac{2}{3}}\, q_2^{\frac{1}{2}} & -8\, q_1^{-\frac{1}{3}}\, q_2^{\frac{3}{2}} \end{bmatrix},$$

and all the partial derivatives $\dfrac{\partial q_i}{\partial p_j}$ are negative, i.e., the inputs decrease with an increase in either price.

Because the sum of the exponents in the production function is less than one, $\frac{1}{3} + \frac{1}{2} < 1$, the production has decreasing return to scale. This property causes $D^2\pi = D\mathbf{g}$ to be negative definite, the critical points to be maximum, and the equations to define the inputs as implicit functions of the prices. ■

Example 3.4 (**Nonlinear Keynesian Model IS-LM Model for National Income).** See §15.3 in [**11**] or §8.6 in [**5**] for more economic discussion. The variables are as follows:

Y	Gross domestic product (GDP)	T	Taxes
C	Consumption	r	Interest rate
I	Investment expenditure	M	Money supply
G	Government spending		

The GDP is assumed to be the sum of the consumption, investment, and government spending, $Y = C + I + G$. The consumption is a function of after taxes income or the difference $Y - T$, $C = C(Y - T)$; the investment is a function of the interest rate, $I = I(r)$; and the money supply is a function (called the liquidity function) of GDP and the interest rate, $M = L(Y, r)$. We can think of the assumptions as yielding the following two functional relationships between the five variables:

$$\begin{aligned} 0 &= -Y + C(Y - T) + I(r) + G \qquad \text{and} \\ 0 &= L(Y, r) - M. \end{aligned} \tag{6}$$

The assumptions on the derivatives of the unspecified functions are

$$0 < C'(x) < 1, \qquad I'(r) < 0, \qquad \frac{\partial L}{\partial Y} > 0, \qquad \text{and} \qquad \frac{\partial L}{\partial r} < 0.$$

For $C(Y-T)$, $\dfrac{\partial C}{\partial Y} = C'$ and $\dfrac{\partial C}{\partial T} = -C'$. Examples of functions that satisfy these derivative conditions are $C(x) = \frac{1}{3}xe^{-x} + \frac{1}{2}x$, $I(r) = \dfrac{2I_0}{e^{10r}+1}$, and $L(Y,r) = \frac{9}{10}Y\left[\dfrac{1+3e^{-10r}}{2+2e^{-10r}}\right]$.

Taking the partial derivatives of the two equations (6) with respect to Y and r is an invertible matrix,

$$\Delta = \det\begin{bmatrix} -1+C' & I' \\ L_Y & L_r \end{bmatrix} = (C'-1)L_r - I'L_Y > 0 :$$

The sign of Δ is positive because $(C'-1) < 0$, $L_r < 0$, $(C'-1)\,L_r > 0$, $I' < 0$, $L_Y > 0$, and $-I'\,L_Y > 0$. Therefore, the Implicit Function Theorem implies the following dependency of the variables: G, M, and T can be considered as the independent or exogenous variables which can be controlled; C and I are intermediate variables (or functions); Y and r can be considered as dependent or endogenous variables determined by G, M, and T through the variables C and I. Be cautioned that the details of solving for partial derivatives in terms of known quantities and derivatives is more complicated than the other examples considered.

Taking the partial derivatives of the two equations (6) with respect to G, T, and M (in that order), and considering Y and r as functions of these variables, we get the following matrix equation, (ImD):

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & -C' & 0 \\ 0 & 0 & -1 \end{bmatrix} + \begin{bmatrix} C'-1 & I' \\ L_Y & L_r \end{bmatrix} \begin{bmatrix} \dfrac{\partial Y}{\partial G} & \dfrac{\partial Y}{\partial T} & \dfrac{\partial Y}{\partial M} \\ \dfrac{\partial r}{\partial G} & \dfrac{\partial r}{\partial T} & \dfrac{\partial r}{\partial M}. \end{bmatrix},$$

$$\begin{bmatrix} \dfrac{\partial Y}{\partial G} & \dfrac{\partial Y}{\partial T} & \dfrac{\partial Y}{\partial M} \\ \dfrac{\partial r}{\partial G} & \dfrac{\partial r}{\partial T} & \dfrac{\partial r}{\partial M}. \end{bmatrix} = \frac{1}{\Delta}\begin{bmatrix} L_r & -I' \\ -L_Y & C'-1 \end{bmatrix}\begin{bmatrix} -1 & C' & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \frac{1}{\Delta}\begin{bmatrix} -L_r & L_rC' & -I' \\ L_Y & -L_YC' & C'-1 \end{bmatrix}.$$

The partial derivatives of Y and r with respect to G are positive:

$$\frac{\partial Y}{\partial G} = \frac{-L_r}{\Delta} > 0 \qquad \text{and} \qquad \frac{\partial r}{\partial G} = \frac{L_Y}{\Delta} > 0.$$

Thus, both the GDP, Y, and the interest rate, r, increase if government spending is increased while holding the money supply and taxes fixed. The partial derivatives of Y and r with respect to T are negative, and both Y and r decrease with increased taxes:

$$\frac{\partial Y}{\partial T} = \frac{L_r\,C'}{\Delta} < 0 \qquad \text{and} \qquad \frac{\partial r}{\partial T} = \frac{-L_Y\,C'}{\Delta} < 0,$$

Finally,

$$\frac{\partial Y}{\partial M} = \frac{-I'}{\Delta} > 0 \qquad \text{and} \qquad \frac{\partial r}{\partial M} = \frac{C'-1}{\Delta} < 0,$$

and Y increases and r decreases with an increased money supply.

This type of analysis can give qualitative information without knowing the particular form of the functional dependence but only the signs of the rates of change. ■

3.1. Exercises

3.1.1. A firm uses two inputs, q_1 and q_2 to produce a single output Q, given by the production function $Q = kq_1^{2/5}q_2^{1/5}$. Let P be the price of the output Q, p_1 be the price of q_1, and p_2 be the price of q_2. The profit is given by $\pi = Pkq_1^{2/5}q_2^{1/5} - p_1q_1 - p_2q_2$. The inputs that maximize profits satisfy

$$0 = \frac{2Pk}{5}\, q_1^{-3/5}\, q_2^{1/5} - p_1 \qquad \text{and}$$
$$0 = \frac{Pk}{5}\, q_1^{2/5}\, q_2^{-4/5} - p_2.$$

- **a.** Show that this two equations can be used to determine the amounts of inputs q_1 and q_2 in terms of the prices p_1, p_2, and P. Show that the relevant matrix has nonzero determinant.
- **b.** Write the matrix equation for the partial derivatives of q_1 and q_2 with respect to p_1, p_2 and P in terms of the variables.
- **c.** Solve for the matrix of partial derivatives of q_1 and q_2 in terms of p_1, p_2 and P.

3.1.2. Assume that the two equations that balance supply and demand for a single product are given by

$$0 = Q - S_0\,(1 - e^{T-P})$$
$$0 = Q - Ye^{-P}.$$

Here Q is the quantity sold, P is the price, T are the taxes, Y is the income of the consumers, and S_0 is a constant.

- **a.** Use the implicit function theorem to show that these two equations define Q and P as functions of T and Y (keeping S_0 fixed).
- **b.** Solve the matrix of partial derivatives of $\dfrac{\partial P}{\partial Y}$, $\dfrac{\partial P}{\partial T}$, $\dfrac{\partial Q}{\partial Y}$, $\dfrac{\partial Q}{\partial T}$. You may leave your answer in matrix form.

3.1.3. A nonlinear Keynesian IS-LM model for national income involves the following quantities:

Y	Gross domestic product (GDP)
G	Government spending
r	Interest rate
M	Money supply

In addition, there are three quantities which are functions of the other variables (intermediate variables). Investment expenditure I is a function of the interest rate given by $I(r) = \dfrac{I_0}{r+1}$. The consumption is a function of Y given by $C(Y) = C_0 + \frac{5}{6}\,Y + \frac{1}{6}\,e^{-Y}$ with C_0 a constant. The gross domestic product is the sum of consumption, investment expenditure, and government spending, $Y = C+I+G = C_0 + \frac{5}{6}\,Y + \frac{1}{6}\,e^{-Y} + \dfrac{I_0}{r+1} + G$. The money supply equals the liquidity function, $M = \dfrac{Y}{r+1}$. With these assumptions, the model yields the following two

equations:

$$0 = C_0 - \frac{1}{6}\,Y + \frac{1}{6}\,e^{-Y} + \frac{I_0}{r+1} + G \qquad \text{and}$$
$$0 = \frac{Y}{r+1} - M.$$

a. Using the Implicit Function Theorem, show that these two equations define Y and r as dependent variables which are determined by the independent variables G and M, i.e., these equations define Y and r as functions of G and M.

b. Write the matrix equation that the partial derivatives of Y and r with respect to G and M must satisfy.

c. Solve for the matrix equation for the partial derivatives of Y and r with respect to G and M.

3.1.4. Consider the three equations

$$\begin{aligned} xyz + u + v^2 &= 6, \\ xy - zy^2 + u^2 + v + w &= 6, \\ xy^3 - zx^2 + u^2 + w &= 4. \end{aligned}$$

a. Show that these equations implicitly define (u, v, w) in terms of (x, y, z).

b. What is the system of equations (ImD) for the partial derivatives near the point $(x, y, z, u, v, w) = (1, 1, 1, 1, 2, 3)$?

3.2. Extrema with Equality Constraints

This section treats optimization with equality constraints. This topic is usually covered in multidimensional calculus course, but we do a more general case than is often covered and we give a proof using the Implicit Function Theorem.

Definition. For C^1 constraints $g_i : \mathbb{R}^n \to \mathbb{R}$ for $i = 1, \ldots, k$, the constraints satisfy the *constraint qualification at a point* $\mathbf{x}^*$ provided that

$$\operatorname{rank}\left(D\mathbf{g}(\mathbf{x}^*)\right) = k,$$

i.e., the gradients $\{\, \nabla g_i(\mathbf{x}^*) \,\}_{i=1}^k$ are linearly independent.

Theorem 3.5 (Method of Lagrange Multipliers). *Assume $f, g_i : \mathbb{R}^n \to \mathbb{R}$ are C^1 functions for $i = 1, \ldots, k$. Suppose that $\mathbf{x}^*$ is a local extreme point of f on the set $\mathbf{g}^{-1}(\mathbf{b}) = \{\, \mathbf{x} \in \mathbb{R}^n : g_i(\mathbf{x}) = b_i,\ i = 1, \ldots, k \,\}$. Then at least one of the following holds:*

1. *There exists $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_k^*) \in \mathbb{R}^k$ such that*

$$Df(\mathbf{x}^*) = \sum_{i=1}^{k} \lambda_i^*\, Dg_i(\mathbf{x}^*). \tag{LM}$$

2. *The constraint qualification fails at $\mathbf{x}^*$, $\operatorname{rank}\left(D\mathbf{g}(\mathbf{x}^*)\right) < k$.*

We give a proof of the theorem later in the section after further discussion and examples.

In a calculus course, this theorem for one or two constraints is considered. The justification given is often as follows. For one constraint $g(\mathbf{x}) = 0$, if $\mathbf{x}^*$ is an extreme point and $\mathbf{v}$ is perpendicular to $\nabla g(\mathbf{x}^*)$ then $\mathbf{v}$ is tangent vector to the level set and the directional derivative of f in the direction of $\mathbf{v}$ must be zero. Thus, $\nabla f(\mathbf{x}^*)$ is perpendicular to the same vectors as $\nabla g(\mathbf{x}^*)$ and so must be parallel to $\nabla g(\mathbf{x}^*)$. To make this precise, we must find a curve in the level set whose tangent vector is equal to $\mathbf{v}$.

A similar justification can be given with two constraints. The tangent vectors to the level set $g_1(\mathbf{x}) = b_1$ and $g_2(\mathbf{x}) = b_2$ are all vectors perpendicular to both $\nabla g_1(\mathbf{x}^*)$ and $\nabla g_2(\mathbf{x}^*)$. This is the same as the null space of $D\mathbf{g}(\mathbf{x}^*)$. Again, the directional derivative of f in the direction of $\mathbf{v}$ must be zero, $\nabla f(\mathbf{x}^*) \cdot \mathbf{v} = 0$. Since $\nabla f(\mathbf{x}^*)$ is perpendicular to all these vectors $\mathbf{v}$, it must be a linear combination of $\nabla g_1(\mathbf{x}^*)$ and $\nabla g_2(\mathbf{x}^*)$.

Example 3.6. It is possible for the constraint qualification to fail at a maximum. Let $g(x, y) = x^3 + y^2 = 0$ and $f(x, y) = y + 2x$.

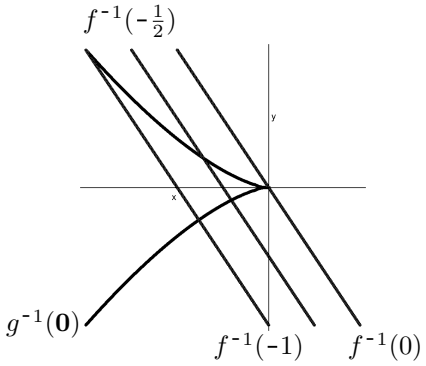


Figure 3.2.1. Maximum at a singular point

The maximum of $f(x, y)$ is at singular point $(0, 0)$, where $\nabla g = \mathbf{0}$.

$$\nabla f(\mathbf{0}) = (2, 1)^\mathsf{T} \neq \mathbf{0} = \lambda \nabla g.$$

■

Example 3.7. This example illustrates the fact that for two (or more) constraints, the maximum of the objective function f can occur where $\nabla g_1(\mathbf{x}^*)$ and $\nabla g_2(\mathbf{x}^*)$ are parallel (a point where the constraint qualification fails). At such a point, it is not always possible to write $\nabla f(\mathbf{x}^*)$ as a a linear combination of the gradients of the constraint equation. (cf. Exercise 3.3.5 with inequality constraints.)

Let $g_1(x, y, z) = x^3 + y^2 + z = 0$, $g_2(x, y, z) = z = 0$, and $f(x, y, z) = y + 2x$. Level set is that of last example in (x, y)-plane,

$$\mathbf{g}^{-1}(\mathbf{0}) = \{\, (x, y, 0) : x^3 + y^2 = 0 \,\}.$$

The maximum of f on $\mathbf{g}^{-1}(\mathbf{0})$ is at $\mathbf{0}$. The gradients of of g_1 and g_2 are parallel at $\mathbf{0}$:

$$\begin{aligned} \nabla g_1(x, y, z) &= (3x^2,\, 2y,\, 1)^\mathsf{T}, & \nabla g_2(x, y, z) &= (0,\, 0,\, 1)^\mathsf{T} \\ \nabla g_1(0, 0, 0) &= (0,\, 0,\, 1)^\mathsf{T}, & \nabla g_2(0, 0, 0) &= (0,\, 0,\, 1)^\mathsf{T}. \end{aligned}$$

Therefore, $\operatorname{rank}(D\mathbf{g}(\mathbf{0})) = \operatorname{rank} \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} = 1 < 2$. Also,

$$\begin{aligned} \nabla f(\mathbf{0}) &= (2, 1, 0)^\mathsf{T} \\ &\neq \lambda_1 (0, 0, 1)^\mathsf{T} + \lambda_2 (0, 0, 1)^\mathsf{T} \\ &= \lambda_1 \nabla g_1(\mathbf{0}) + \lambda_2 \nabla g_2(\mathbf{0}). \end{aligned}$$

■

The proof of the theorem uses the fact that the tangent vectors to the level set is the null space of the derivative of the constraint function. We start with a precise definition of the tangent vectors.

Definition. For a C^1 function $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^k$ with $\mathbf{g}(\mathbf{x}^*) = \mathbf{b}$, denote the set of tangent vectors to the level set $\mathbf{g}^{-1}(\mathbf{b})$ by

$$\mathbf{T}_{\mathbf{g}}(\mathbf{x}^*) = \left\{ \mathbf{v} = \mathbf{r}'(0) : \mathbf{r}(t) \text{ is a } C^1 \text{ curve with } \mathbf{r}(0) = \mathbf{x}^*, \text{ and } \mathbf{g}(\mathbf{r}(t)) = \mathbf{b} \text{ for all small } t \right\}.$$

The linear space $\mathbf{T}_{\mathbf{g}}(\mathbf{x}^*)$ is called the *tangent space to* $\mathbf{g}^{-1}(\mathbf{b})$ *at* $\mathbf{x}^*$.

Proposition 3.8. *Assume that* $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^k$ *is* C^1, $\mathbf{g}(\mathbf{x}^*) = \mathbf{b}$, *and the rank of* $D\mathbf{g}(\mathbf{x}^*)$ *is* k. *Then, the set of tangent vectors at* $\mathbf{x}^*$ *equals to the null space of* $D\mathbf{g}(\mathbf{x}^*)$, $\mathbf{T}_{\mathbf{g}}(\mathbf{x}^*) = \operatorname{null}(D\mathbf{g}(\mathbf{x}^*))$.

Remark. Calculus books often assume implicitly that this proposition is true and that the set of tangent vectors are all vectors perpendicular to the gradients of the constraints.

Proof. For any curve $\mathbf{r}(t)$ with $\mathbf{r}(0) = \mathbf{x}^*$ and $\mathbf{g}(\mathbf{r}(t)) = \mathbf{b}$ for all small t,

$$\mathbf{0} = \frac{d}{dt}\mathbf{g}(\mathbf{r}(t))\bigg|_{t=0} = D\mathbf{g}(\mathbf{r}(0))\mathbf{r}'(0) = D\mathbf{g}(\mathbf{x}^*)\mathbf{r}'(0),$$

so $\mathbf{T}_{\mathbf{g}}(\mathbf{x}^*) \subset \operatorname{null}(D\mathbf{g}(\mathbf{x}^*))$.

The proof that $\operatorname{null}(D\mathbf{g}(\mathbf{x}^*)) \subset \mathbf{T}_{\mathbf{g}}(\mathbf{x}^*)$ uses the Implicit Function Theorem. Assume the variables have been ordered $\mathbf{x} = (\mathbf{w}, \mathbf{z})$ so that $\det(D_{\mathbf{z}}\mathbf{g}(\mathbf{x}^*)) \neq 0$, so the level set is locally a graph $\mathbf{z} = \mathbf{h}(\mathbf{w})$. Take a vector $\mathbf{v} = (\mathbf{v}_{\mathbf{w}}, \mathbf{v}_{\mathbf{z}})^{\mathsf{T}}$ in $\operatorname{null}(D\mathbf{g}(\mathbf{x}^*))$. Consider the curves the line $\mathbf{w}(t) = \mathbf{w}^* + t\mathbf{v}_{\mathbf{w}}$ in $\mathbf{z}$-space and the curve $\mathbf{r}(t) = \begin{pmatrix} \mathbf{w}(t) \\ \mathbf{h}(\mathbf{w}(t)) \end{pmatrix}$ in the level set. Then, $\mathbf{r}(0) = \mathbf{x}^*$, $\mathbf{r}'(0) \in \mathbf{T}_{\mathbf{g}}(\mathbf{x}^*)$,

$$\mathbf{r}'(0) = \begin{pmatrix} \mathbf{v}_{\mathbf{w}} \\ D\mathbf{h}(\mathbf{z}^*)\,\mathbf{v}_{\mathbf{w}} \end{pmatrix}, \qquad \text{and}$$

$$\mathbf{0} = \frac{d}{dt}\, g(\mathbf{r}(t))\bigg|_{t=0} = D\mathbf{g}(\mathbf{x}^*)\,\mathbf{r}'(0).$$

Thus, $\mathbf{r}'(0) \in \operatorname{null}(D\mathbf{g}(\mathbf{x}^*))$ and has the same the same $\mathbf{w}$-components as $\mathbf{v}$. Because $\operatorname{null}(D\mathbf{g}(\mathbf{x}^*))$ is a graph over the $\mathbf{w}$-coordinates, $\mathbf{v} = \mathbf{r}'(0) \in \mathbf{T}_{\mathbf{g}}(\mathbf{x}^*)$. Thus, we have shown that $\operatorname{null}(D\mathbf{g}(\mathbf{x}^*)) \subset \mathbf{T}_{\mathbf{g}}(\mathbf{x}^*)$. Combining, $\operatorname{null}(D\mathbf{g}(\mathbf{x}^*)) = \mathbf{T}_{\mathbf{g}}(\mathbf{x}^*)$. □

Proof of the Lagrange Multiplier Theorem. Assume that $\mathbf{x}^*$ is a local extremizer of f on $g^{-1}(\mathbf{b})$. Then,

$$0 = \frac{d}{dt} f(\mathbf{r}(t))\bigg|_{t=0} = Df(\mathbf{x}^*)\mathbf{v}$$

for all curves $\mathbf{r}(t)$ in $\mathbf{g}^{-1}(\mathbf{b})$ with $\mathbf{r}(0) = \mathbf{x}^*$ and $\mathbf{r}'(0) = \mathbf{v}$, i.e., $Df(\mathbf{x}^*)\mathbf{v} = 0$ for all $\mathbf{v} \in \mathbf{T}_{\mathbf{g}}(\mathbf{x}^*) = \operatorname{null}(D\mathbf{g}(\mathbf{x}^*))$. This says that the null spaces

$$\operatorname{null}(D\mathbf{g}(\mathbf{x}^*)) = \operatorname{null}\begin{pmatrix} D\mathbf{g}(\mathbf{x}^*) \\ Df(\mathbf{x}^*) \end{pmatrix}$$

and of dimension $n - k$. The rank of each matrix is equal to the number of columns minus the dimension of the null space,

$$k = \operatorname{rank}\left(D\mathbf{g}(\mathbf{x}^*)\right) = \operatorname{rank}\begin{pmatrix} D\mathbf{g}(\mathbf{x}^*) \\ Df(\mathbf{x}^*) \end{pmatrix} = n - (n - k) = k.$$

This implies that the last row of the second matrix, $Df(\mathbf{x}^*)$, is a linear combination of the first k rows,

$$Df(\mathbf{x}^*) = \sum_{i=1}^{k} \lambda_i^* Dg_i(\mathbf{x}^*),$$

which is the first order Lagrange multiplier condition (LM). □

Example 3.9. Find the highest point on the set given by $x + y + z = 12$ and $z = x^2 + y^2$.

The function to be maximized is $f(x, y, z) = z$. The two constraint functions are $g(x, y, z) = x + y + z = 12$ and $h(x, y, z) = x^2 + y^2 - z = 0$.

Constraint qualification: If there were a point where constraint qualification fails, then $\nabla g = (1, 1, 1)^\mathsf{T} = s\,\nabla h = s\,(2x, 2y, -1)^\mathsf{T}$. So, $s = -1$ and $x = y = -\frac{1}{2}$. To be on level set, $z = x^2 + y^2 = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$. But then $g\left(-\frac{1}{2}, -\frac{1}{2}, \frac{1}{2}\right) = -\frac{1}{2} \neq 12$. Therefore, there are no points on level set where constraint qualification fails.

First order Lagrange multiplier equations; The first order conditions (LM) are

$$\begin{aligned} f_x &= \lambda g_x + \mu h_x, & 0 &= \lambda + \mu 2x, \\ f_y &= \lambda g_y + \mu h_y, & 0 &= \lambda + \mu 2y, \\ f_z &= \lambda g_z + \mu h_z, & 1 &= \lambda - \mu. \end{aligned}$$

From the third equation, we get $\lambda = 1+\mu$, so we can eliminate this variable from the equations. Substituting in the other equations, they become

$$\begin{aligned} 0 &= 1 + \mu + 2\mu x, \\ 0 &= 1 + \mu + 2\mu y. \end{aligned}$$

Subtracting the second from the first, we get $0 = 2\mu(x - y)$, so $\mu = 0$ or $x = y$.

Consider the first case of $\mu = 0$. But then, $0 = 1+\mu+2\mu x = 1$, which is a contradiction. Therefore, there is no solution with $\mu = 0$.

Next, assume $y = x$. The constraints become $z = 2x^2$ and $12 = 2x + z = 2x + 2x^2$, so $0 = x^2 + x - 6 = (x + 3)(x - 2)$, and $x = 2$ or -3. If $x = 2$, then $y = 2$, $z = 2x^2 = 8$, $0 = 1 + \mu(1 + 2x) = 1 + 5\,\mu$, $\mu = -1/5$, and $\lambda = 4/5$.

If $x = y = -3$, then $z = 2x^2 = 18$, $0 = 1 + \mu(1 + 2x) = 1 - 5\mu$, $\mu = 1/5$, and $\lambda = 6/5$.

We have found two critical points $(\lambda^*, \mu^*, x^*, y^*, z^*) = (4/5, -1/5, 2, 2, 8,)$ and $(6/5, 1/5, -3, -3, 18)$. The values of the objective function at the critical points are $f(2, 2, 8) = 8$ and $f(-3, -3, 18) = 18$. The constraint set is compact so extrema exist and must be one of the critical points. The maximum is at the point $(-3, -3, 18)$ with the maximal value, and the minimum is at the point $(2, 2, 8)$ with the minimum value. ■

Lagrangian. The first derivative conditions (LM) can be seen as the critical point what is called the *Lagrangian*, which is defined by

$$L(\boldsymbol{\lambda}, \mathbf{x}) = f(\mathbf{x}) + \sum_{i=1}^{k} \lambda_i \left(b_i - g_i(\mathbf{x})\right).$$

A point $\mathbf{x}^*$ satisfies the first order Lagrange multiplier conditions (LM) with multipliers $\boldsymbol{\lambda}^*$ iff $(\boldsymbol{\lambda}^*, \mathbf{x}^*)$ is a critical point of L with respect to all its variables,

$$\frac{\partial L}{\partial \lambda_i}(\boldsymbol{\lambda}^*, \mathbf{x}^*) = b_i - g_i(\mathbf{x}^*) = 0 \qquad \text{for } 1 \leq i \leq k \quad \text{and}$$

$$D_{\mathbf{x}} L(\boldsymbol{\lambda}^*, \mathbf{x}^*) = Df(\mathbf{x}^*) - \sum_{i=1}^{k} \lambda_i^*\, Dg_i(\mathbf{x}^*) = \mathbf{0}.$$

To ensure that the constraint qualification does not fail, we need that

$$\operatorname{rank}\left(\left(L_{\lambda_i, x_j}(\boldsymbol{\lambda}^*, \mathbf{x}^*)\right)_{ij}\right) = \operatorname{rank}\left(Dg(\mathbf{x}^*)\right) = k.$$

These conditions on the Lagrangian are not a proof but merely a mnemonic device.

3.2.1. Interpretation of Lagrange Multipliers

For a tight constraint in a maximization linear program, sensitivity analysis showed that the marginal value (in terms of the maximal value of the objective function) of a change in this tight input equals the value of the comparable variable for the optimal solution of the dual linear program. Because this dual variable satisfies complementary slackness, it is the corresponding Lagrange multiplier for the problem. The following theorem gives the comparable result for a nonlinear problem with equality constraints and gives the marginal maximal value with changes in b_i.

Theorem 3.10. *Assume that $f, g_i : \mathbb{R}^n \to \mathbb{R}$ are C^2 for $1 \leq i \leq k < n$. For $\mathbf{b} \in \mathbb{R}^k$, assume that $(\boldsymbol{\lambda}^*(\mathbf{b}), \mathbf{x}^*(\mathbf{b}))$ is a solution of* (LM) *for an extremizer of f on $\mathbf{g}^{-1}(\mathbf{b})$ and satisfies $\operatorname{rank}(D\mathbf{g}(\mathbf{x}^*(\mathbf{b}))) = k$. Let $L(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{b}) = f(\mathbf{x}) + \sum_{j=1}^{k} \lambda_j \left(b_j - g_j(\mathbf{x})\right)$ be the Lagrangian as a function of $\mathbf{b}$ as well as $\mathbf{x}$ and $\boldsymbol{\lambda}$. Also, assume the second derivative of the Lagrangian as a function of $\boldsymbol{\lambda}$ and $\mathbf{x}$ is nondegenerate, i.e., it has a nonzero determinant. Then,*

$$\lambda_i^*(\mathbf{b}) = \frac{\partial}{\partial b_i} f(\mathbf{x}^*(\mathbf{b})).$$

Proof. For fixed $\mathbf{b}$, $(\boldsymbol{\lambda}^*, \mathbf{x}^*) = (\boldsymbol{\lambda}^*(\mathbf{b}), \mathbf{x}^*(\mathbf{b}))$ satisfy

$$\mathbf{0} = \begin{pmatrix} D_{\boldsymbol{\lambda}} L^{\mathsf{T}} \\ D_{\mathbf{x}} L^{\mathsf{T}} \end{pmatrix} = \begin{pmatrix} \mathbf{b} - \mathbf{g}(\mathbf{x}) \\ Df(\mathbf{x})^{\mathsf{T}} - \sum_j \lambda_j \, Dg_j(\mathbf{x})^{\mathsf{T}} \end{pmatrix} = \mathbf{G}(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{b}).$$

The derivative of $\mathbf{G}$ with respect to $\boldsymbol{\lambda}$ and $\mathbf{x}$ is the second derivative of L with respect to $\boldsymbol{\lambda}$ and $\mathbf{x}$ and is the bordered Hessian

$$D_{\boldsymbol{\lambda},\mathbf{x}}\mathbf{G}(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{b})) = (D_{\boldsymbol{\lambda}}\mathbf{G}(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{b}), D_{\mathbf{x}}\mathbf{G}(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{b})) = \begin{bmatrix} \mathbf{0}_k & -D\mathbf{g} \\ -D\mathbf{g}^{\mathsf{T}} & D_{\mathbf{x}}^2 L \end{bmatrix}.$$

This derivative satisfies $\det\left(D_{\boldsymbol{\lambda},\mathbf{x}}\mathbf{G}(\boldsymbol{\lambda}^*(\mathbf{b}), \mathbf{x}^*(\mathbf{b}), \mathbf{b})\right) \neq 0$, since this second derivative of the Lagrangian is assumed to be nondegenerate. See Section 3.5 for a discussion of bordered Hessians and nondegenerate extrema on a level set. Therefore, the variables $(\boldsymbol{\lambda}, \mathbf{x}) = (\boldsymbol{\lambda}^*(\mathbf{b}), \mathbf{x}^*(\mathbf{b})$ are implicitly determined differentiable functions of $\mathbf{b}$ by the equation $\mathbf{G}(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{b}) = \mathbf{0}$.

At these points on the level sets, $f(\mathbf{x}^*(\mathbf{b})) = L(\boldsymbol{\lambda}^*(\mathbf{b}), \mathbf{x}^*(\mathbf{b}), \mathbf{b})$. Taking the partial derivative of this equality with respect to b_i gives

$$\frac{\partial}{\partial b_i} f(\mathbf{x}^*(\mathbf{b})) = D_{\boldsymbol{\lambda}} L \frac{\partial}{\partial b_i} \boldsymbol{\lambda}^*(\mathbf{b}) + D_{\mathbf{x}} L \frac{\partial}{\partial b_i} \mathbf{x}^*(\mathbf{b}) + \frac{\partial L}{\partial b_i}(\mathbf{x}^*(\mathbf{b}), \boldsymbol{\lambda}^*(\mathbf{b}), \mathbf{b}).$$

Because $b_j - g_j(\mathbf{x}^*(\mathbf{b})) = 0$ for all j,

$$D_{\boldsymbol{\lambda}} L \frac{\partial}{\partial b_i} \boldsymbol{\lambda}^*(\mathbf{b}) = \sum_j \left[b_j^* - g_j(\mathbf{x}^*(\mathbf{b})) \frac{\partial}{\partial b_i} \right] \lambda_j(\mathbf{b}) = 0.$$

Because the point satisfies the (LM) conditions,

$$D_{\mathbf{x}} L \frac{\partial}{\partial b_i} \mathbf{x}^*(\mathbf{b}) = \left[Df(\mathbf{x}^*(\mathbf{b})) - \sum_{j=1}^{k} \lambda_j^*(\mathbf{b}) Dg_j(\mathbf{x}^*(\mathbf{b})) \right] \frac{\partial}{\partial b_i} \mathbf{x}^*(\mathbf{b}) = 0.$$

Finally, by taking the partial derivative of the formula for L,

$$\frac{\partial L}{\partial b_i}(\mathbf{x}^*(\mathbf{b}), \boldsymbol{\lambda}^*(\mathbf{b}), \mathbf{b}) = \lambda_i^*(\mathbf{b}).$$

Combining the terms, we get the equality required.

An alternative proof is as follows: Once we know that $(\boldsymbol{\lambda}^*(\mathbf{b}, \mathbf{x}^*(\mathbf{b})$ is a differentiable function of $\mathbf{b}$, by the Chain Rule and since the points solve the Lagrange multiplier problem,

$$\begin{aligned}\frac{\partial}{\partial b_i} f(\mathbf{x}^*(\mathbf{b})) &= Df(\mathbf{x}^*(\mathbf{b})) \frac{\partial}{\partial b_i} \mathbf{x}^*(\mathbf{b}) \\ &= \sum_j \lambda_j^*(\mathbf{b})\, Dg_j(\mathbf{x}^*(\mathbf{b})) \frac{\partial}{\partial b_i} \mathbf{x}^*(\mathbf{x}^*(\mathbf{w})) \\ &= \sum_j \lambda_j^*(\mathbf{b}) \frac{\partial}{\partial b_i} g_j(\mathbf{x}^*(\mathbf{b})) \\ &= \lambda_i^*(\mathbf{b}).\end{aligned}$$

The last equality holds because $g_j(\mathbf{x}^*(\mathbf{b})) = b_j$ for all $\mathbf{b}$, so

$$\frac{\partial}{\partial b_i} g_j(\mathbf{x}^*(\mathbf{b})) = \begin{cases} 0 & \text{when } j \neq i \\ 1 & \text{when } j = i. \end{cases}$$

□

3.2. Exercises

3.2.1. Find the points satisfying the first order conditions for a constrained extrema. Then compare the values and argue which points are global maxima and which are global minima.

a. $f(x, y, z) = xyz$ and $g(x, y, z) = 2x + 3y + z = 6$.

b. $f(x, y, z) = 2x + y^2 - z^2$, $g_1(x, y, z) = x - 2y = 0$, and $g_2(x, y, z) = x + z = 0$.

3.2.2. For each of the following objective and constraint functions, find the maximizer and minimizers.

a. $f(x, y, z) = x^2 + y^2 + z^2$, subject to $g(x, y, z) = x + y + z = 12$ and $h(x, y, z) = x^2 + y^2 - z = 0$.

b. $f(x, y, z) = x+y+z$, subject to $g(x, y, z) = x^2+y^2 = 2$ and $h(x, y, z) = x + z = 1$.

c. Minimize $f(x, y, z) = x^2+y^2+z^2$, subject to $g(x, y, z) = x+2y+3z = 6$ and $h(x, y, z) = y + z = 0$.

3.2.3. Let $u(x, y, z) = x^2y^3z^4$ and the expenditure be $E(x, y, z) = p_1x+p_2y+p_3z$ with $p_1 > 0$, $p_2 > 0$, and $p_3 > 0$. For fixed $w > 0$, let
$\mathbf{X}(w) = \{\, (x, y, z) \in \mathbb{R}^3_+ : u(x, y, z) \geq w \,\}$.

a. Even though $\mathbf{X}(w)$ is not compact, *show* that E attains a minimum on $\mathbf{X}(w)$ using the Extreme Value Theorem. Note that $(x_0, y_0, z_0) = (\sqrt{w}, 1, 1)$ is in $\mathbf{X}(w)$ and $E(\sqrt{w}, 1, 1) = p_1\sqrt{w} + p_2 + p_3$.

b. Since $\nabla E \neq \mathbf{0}$ at all points of $\mathbf{X}(w)$ (or $DE(x, y, z) \neq \mathbf{0}$), it follows that the minimum cannot be in the interior of $\mathbf{X}(w)$ but must be on the boundary $\{(x, y, z) : u(x, y, z) = w\}$. (You do not need to show this.)
Using the first order conditions for Lagrange Multipliers, *find* the point that attains the minimum of E subject to $u(x, y, z) = w$. Explain why this point is the minimizer E on $\mathbf{X}(w)$.

3.3. Extrema with Inequality Constraints: Necessary Conditions

In the rest of this chapter, we use the following notation for a feasible sets defined by resource requirements with $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^m$,

$$\mathscr{F}_{\mathbf{g},\mathbf{b}} = \{\, \mathbf{x} \in \mathbb{R}^n : \mathbf{g}(\mathbf{x}) \leq \mathbf{b} \,\}.$$

Just as for a linear program, a constraint $g_i(\mathbf{x}) \leq b_i$ is said to be *slack* at $\mathbf{x} = \mathbf{p}$ provided that $g_i(\mathbf{p}) < b_i$. A constraint $g_i(\mathbf{x}) \leq b_i$ is said to be *effective or tight at* $\mathbf{x} = \mathbf{p}$ provided that $g_i(\mathbf{p}) = b_i$.

The necessary conditions for an optimal solution allow for the possibility that an extremizer is a point that is the nonlinear analog of a degenerate basic solution for a linear program. The "constraint qualification" is said to fail at such a point as given in the following definition.

Definition. Given a point $\mathbf{p}$, we let $\mathbf{E}(\mathbf{p}) = \{\, i : g_i(\mathbf{p}) = b_i \,\}$ be the set of tight constraints at $\mathbf{p}$, $|\mathbf{E}(\mathbf{p})|$ be the cardinality of $\mathbf{E}(\mathbf{p})$, and $\mathbf{g}_{\mathbf{E}(\mathbf{p})}(\mathbf{x}) = (\, g_i(\mathbf{x}) \,)_{i \in \mathbf{E}(\mathbf{p})}$ be the function with tight constraints only.

The set of constraints satisfies the *constraint qualification at* $\mathbf{p}$ provided that $|\mathbf{E}(\mathbf{p})| = \operatorname{rank}(D\mathbf{g}_{\mathbf{E}(\mathbf{p})}(\mathbf{p}))$, i.e., the gradients of the tight constraints are linearly independent at $\mathbf{p}$.

A set of constraints satisfies the *constraint qualification on the feasible set* $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ provided it is satisfied at all the points on the boundary.

Example 3.11. Consider the constraints

$$\begin{aligned} g_1(x,y) &= x + (y-1)^3 \leq 0, \\ g_2(x,y) &= \text{-}x \leq 0, \\ g_3(x,y) &= \text{-}y \leq 0. \end{aligned}$$

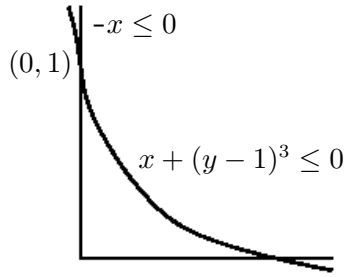


Figure 3.3.1. Constraint qualification fails

At $(0,1)$, $\mathbf{E}(0,1) = \{1,2\}$, $\mathbf{g}_{\mathbf{E}(0,1)}(x,y) = (x + (y-1)^3, \text{-}x)^{\mathsf{T}}$, and

$$\operatorname{rank}(D\mathbf{g}_{\mathbf{E}(0,1)}(0,1)) = \operatorname{rank} \begin{bmatrix} 1 & 3(y-1)^2 \\ \text{-}1 & 0 \end{bmatrix}_{y=1} = \operatorname{rank} \begin{bmatrix} 1 & 0 \\ \text{-}1 & 0 \end{bmatrix} = 1 < 2 = |\mathbf{E}(0,1)|.$$

Therefore, the constraint qualification fails at $(0,1)$. Note that the level sets $g_1^{\text{-}1}(0)$ and $g_2^{\text{-}1}(0)$ are tangent at $(0,1)$ and the gradients $\nabla g_1(0,1)$ and $\nabla g_2(0,1)$ are parallel. ■

Theorem 3.12 (Karush-Kuhn-Tucker). *Suppose that $f, g_i : \mathbb{R}^n \to \mathbb{R}$ are C^1 functions for $1 \leq i \leq m$, and f attains a local extremum at $\mathbf{x}^*$ on $\mathscr{F}_{\mathbf{g},\mathbf{b}}$.*
*Then either (**a**) the constraint qualification fails at $\mathbf{x}^*$ with $\operatorname{rank}(D\mathbf{g}_{\mathbf{E}(\mathbf{x}^*)}(\mathbf{x}^*)) < |\mathbf{E}(\mathbf{x}^*)|$, or*
*(**b**) there exist $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_m^*)$ such that KKT-1,2,3 or KKT-1,2,3′ hold:*

KKT-1. $Df(\mathbf{x}^*) = \sum_{i=1}^m \lambda_i^* \, Dg_i(\mathbf{x}^*)$.
KKT-2. $\lambda_i^* \, (\, b_i - g_i(\mathbf{x}^*) \,) = 0$ *for* $1 \le i \le m$ *(so* $\lambda_i^* = 0$ *for* $i \notin \mathbf{E}(\mathbf{x}^*)$ *).*
KKT-3. *If* $\mathbf{x}^*$ *is a maximizer, then* $\lambda_i^* \ge 0$ *for* $1 \le i \le m$.
KKT-3′. *If* $\mathbf{x}^*$ *is a minimizer, then* $\lambda_i^* \le 0$ *for* $1 \le i \le m$.

Remark. This theorem is the nonlinear version of Proposition 1.13 in Chapter 1 for linear programs. The proof is similar with the Implicit Function Theorem replacing part of the argument done by linear algebra for the linear program.

Remark. Exercise 3.3.5 gives an example for which the maximum occurs at a point where the constraint qualification fails.

Remark. We call KKT-1,2,3 the *first order Karush-Kuhn-Tucker conditions* for a maximum.

Condition KKT-1 implies that $\nabla f(\mathbf{x}^*)$ is perpendicular to the tangent space of the level set of the tight constraints at $\mathbf{x}^*$, $\mathbf{g}^{-1}_{\mathbf{E}(\mathbf{x}^*)}(\mathbf{b}_{\mathbf{E}(\mathbf{x}^*)})$.

Condition KKT-2 is called *complementary slackness*: If a constraint is slack, $g_i(\mathbf{x}^*) < b_i$, then the corresponding multiplier $\lambda_i^* = 0$; if $\lambda_i^* \neq 0$, then the constraint is tight, $g_i(\mathbf{x}^*) = b_i$.

Condition KKT-3, $\lambda_i^* \ge 0$ for a maximum, implies that the gradient of f points out of the feasible set at $\mathbf{x}^*$. The inequalities, $g_i(\mathbf{x}) \le b_i$, are resource type and signs $\lambda_i^* \ge 0$ are compatible with maximization for a linear programming problem.

Condition KKT-3′, $\lambda_i^* \le 0$ for a minimum, implies that the gradient of f points into the feasible set at $\mathbf{x}^*$. The point $\mathbf{x}^*$ is a maximizer of $-f(\mathbf{x})$ and $-\nabla f(\mathbf{x}^*) = \sum_{i=1}^m (-\lambda_i^*) \, \nabla g_i(\mathbf{x}^*)$ with $-\lambda_i^* \ge 0$. So the signs $\lambda_i^* \le 0$ are compatible with minimization for a linear programming problem.

Steps to Find an Optimal Solution using the KKT Theorem 3.12	
1.	Verify that a maximum (resp. minimum) exists by showing either that the feasible set is compact or that $f(\mathbf{x})$ takes on smaller values (resp. larger values) near infinity.
2.	Find all the possible extremizers: (i) Find all the points on the boundary of the feasible set where the constraint qualification fails; (ii) find all the points that satisfy KKT-1,2,3 (resp. KKT-1,2,3′).
3.	Compare the values of $f(\mathbf{x})$ at all the points found in 2(i) and 2(ii).

Proof of Theorem 3.12. Assume the constraint qualification holds at $\mathbf{x}^*$ and $|\mathbf{E}(\mathbf{x}^*)| = k$. We can rearrange the indices of the g_j so that $\mathbf{E}(\mathbf{x}^*) = \{ \, 1, \ldots, k \, \}$, i.e., $g_i(\mathbf{x}^*) = b_i$ for $1 \le i \le k$ and $g_i(\mathbf{x}^*) < b_i$ for $k+1 \le i \le m$. Also, we can rearrange the indices of the x_j so that $\det \left(\dfrac{\partial g_i}{\partial x_j}(\mathbf{x}^*) \right)_{1 \le i,j \le k} \neq 0$.

Set $\lambda_i^* = 0$ for $i \notin \mathbf{E}(\mathbf{x}^*)$, i.e., for $k+1 \le i \le m$. We essentially drop these ineffective constraints from consideration in the proof. The function f also attains a extremum at $\mathbf{x}^*$ on $\{ \, \mathbf{x} : g_i(\mathbf{x}) = b_i \text{ for } i \in \mathbf{E}(\mathbf{x}^*) \, \}$, so by the Lagrange Multiplier Theorem, there exist λ_i^* for $1 \le i \le k$ so that

$$Df(\mathbf{x}^*) = \sum_{1 \le i \le k} \lambda_i^* Dg_i(\mathbf{x}^*).$$

Since $\lambda_i^* = 0$ for $k+1 \le i \le m$, we can change the summation to 1 to m and obtain condition KKT-1. Also, either $\lambda_i^* = 0$ or $b_i - g_i(\mathbf{x}^*) = 0$, so condition KKT-2 holds.

The question remains: For a maximum, why are $\lambda_\ell^* \ge 0$ for $\ell \in \mathbf{E}(\mathbf{x}^*)$? (The case of a minimum is similar.)

We apply the Implicit Function Theorem to show that there is a curve $\mathbf{r}(t)$ in $\mathscr{F}$ such that, for small $t > 0$, (i) $g_\ell(\mathbf{r}(t)) < b_\ell$, (ii) $g_i(\mathbf{r}(t)) = b_i$ for $i \neq \ell$ and $1 \le i \le k$, and (iii) $r_i(t) = x_i^*$ for $k+1 \le i \le n$.

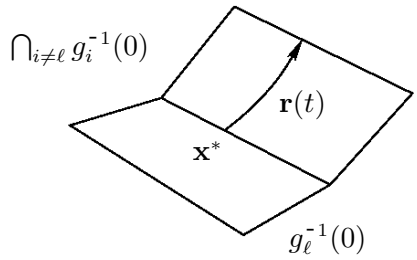


Figure 3.3.2. Curve where only one more constraint becomes slack

Let $\delta_{i\ell} = 0$ if $i \neq \ell$ and $\delta_{\ell\ell} = 1$. We apply the theorem to a function $\mathbf{G}$ defined by

$$G_i(\mathbf{x}, t) = g_i(\mathbf{x}) - b_i + t\,\delta_{i\ell} \qquad \text{for } 1 \leq i \leq k = |\mathbf{E}(\mathbf{x}^*)| \quad \text{and}$$
$$G_i(\mathbf{x}, t) = x_i - x_i^* \qquad \text{for } k+1 \leq i \leq n.$$

The determinant of the derivative of $\mathbf{G}$ with respect to $\mathbf{x}$ is

$$\det(D_{\mathbf{x}}\mathbf{G}(\mathbf{x}^*, 0)) = \det \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_k} & \frac{\partial g_1}{\partial x_{k+1}} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \frac{\partial g_2}{\partial x_1} & \cdots & \frac{\partial g_2}{\partial x_k} & \frac{\partial g_2}{\partial x_{k+1}} & \cdots & \frac{\partial g_2}{\partial x_n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_k}{\partial x_1} & \cdots & \frac{\partial g_k}{\partial x_k} & \frac{\partial g_k}{\partial x_{k+1}} & \cdots & \frac{\partial g_k}{\partial x_n} \\ 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 1 \end{pmatrix}$$
$$= \det \left(\frac{\partial g_i}{\partial x_j}(\mathbf{x}^*) \right)_{1 \leq i,j \leq k} \neq 0.$$

By the Implicit Function Theorem, there exists $\mathbf{x} = \mathbf{r}(t)$ such that $\mathbf{r}(0) = \mathbf{x}^*$ and $\mathbf{G}(\mathbf{r}(t), t) \equiv 0$, i.e., $g_i(\mathbf{r}(t)) = b_i - t\,\delta_{i\ell}$ for $1 \leq i \leq k$ and $x_i = x_i^*$ for $k+1 \leq i \leq n$. This curve $\mathbf{r}(t)$ has the desired properties:

$$Dg_i(\mathbf{x}^*\,\mathbf{r}'(0) = \frac{d}{dt} g_i \circ \mathbf{r}(t)\Big|_{t=0} = -\delta_{i\ell} \qquad \text{for } 1 \leq i \leq k.$$

The function f has a maximum at $\mathbf{x}^*$, so $f(\mathbf{x}^*) \geq f(\mathbf{r}(t))$ for $t \geq 0$, and

$$0 \geq \frac{d}{dt} f \circ \mathbf{r}(t)\Big|_{t=0} = Df(\mathbf{x}^*)\,\mathbf{r}'(0) = \sum_{1 \leq i \leq k} \lambda_i^* Dg_i(\mathbf{x}^*)\,\mathbf{r}'(0) = -\lambda_\ell^*,$$

where the second equality holds by the first order Lagrange Multiplier condition KKT-1 and the third equality holds by the calculation of $Dg_i(\mathbf{x}^*)\mathbf{r}'(0)$. We have shown that that $\lambda_\ell^* \geq 0$ for $1 \leq i \leq k$, so we have KKT-3. □

Example 3.13. Let $f(x, y) = x^2 - y$ and $g(x, y) = x^2 + y^2 \leq 1$. (Note that the feasible set is compact.)

The derivative of the constraint is $Dg(x, y) = (2x, 2y)$, which has rank one at all the points of the boundary of the feasible set. (At least one variable is nonzero at each of the points.) Therefore, the constraint qualification is satisfied at all the points in the feasible set.

The KKT equations KKT-1,2 to be solved are

$$\begin{aligned} 0 &= f_x - \lambda\, g_x = 2x - \lambda\, 2x, \\ 0 &= f_y - \lambda\, g_y = -1 - \lambda\, 2y, \qquad \text{and} \\ 0 &= \lambda\,(1 - x^2 - y^2). \end{aligned}$$

From the first equation, we see that $x = 0$ or $\lambda = 1$.

Case (i): $\lambda = 1 > 0$. We are left with the equations

$$\begin{aligned} 1 &= -2y, \\ 1 &= x^2 + y^2. \end{aligned}$$

From the first equation, $y = \text{-}1/2$, so $x^2 = 1 - 1/4 = 3/4$, or $x = \pm\sqrt{3}/2$.

Case (ii): $x = 0$. We are left with the equations

$$\begin{aligned} 1 &= \text{-}\lambda\, 2y, \\ 0 &= \lambda\,(1 - y^2). \end{aligned}$$

For the first equation, $\lambda \neq 0$. From the second equation, we get that $y = \pm 1$. If $y = 1$, then we get that $0 = -1 - 2\lambda$ and $\lambda = \text{-}1/2 < 0$. This point cannot be a maximum. If $y = \text{-}1$, then $1 = 2\lambda$ and $\lambda = 1/2 > 0$. This is a possible maximizer.

We have found three possible maximizers: $\left(\pm\sqrt{3}/2, \text{-}1/2\right)$ and $(0, \text{-}1)$, with values $f\left(\pm\sqrt{3}/2, \text{-}1/2\right) = 3/4 + 1/2 = 5/4$ and $f(0, \text{-}1) = 1$. Thus the maximum is $5/4$, which is attained at $\left(\pm\sqrt{3}/2, -1/2\right)$.

Notice that although the multiplier is positive at $(0, \text{-}1)$, it is not a maximizer. The function f decreases as it moves into the interior of the region at $(0, \text{-}1)$, but it is a local minimum along the boundary so this point is a type of saddle point on the feasible set. ■

Example 3.14. Maximize $f(x, y, z) = x^2 + 2y^2 + 3z^2$, on the constraint set $1 = x + y + z = g_0(x, y, z)$, $0 \geq \text{-}x = g_1(x, y, z)$, $0 \geq \text{-}y = g_2(x, y, z)$, and $0 \geq \text{-}z = g_3(x, y, z)$. Because the 0^{th}-equation involves an equality, λ_0 can have any sign. For $1 \leq i \leq 3$, we need $\lambda_i \geq 0$.

We want to check that the constraint qualification is satisfied at all points of the feasible set. On the face where $g_0(x, y, z) = 1$ and $g_i(x, y, z) < 0$ for $i = 1, 2, 3$,

$$\operatorname{rank}\left(D\mathbf{g}_{\mathbf{E}}(x, y, z)\right) = \operatorname{rank}\left(Dg_0(x, y, z)\right) = \operatorname{rank}\left(\begin{bmatrix} 1 & 1 & 1 \end{bmatrix}\right) = 1.$$

On the edge where $g_0(0, y, z) = 1$, $g_1(0, y, z) = 0$, and $g_i(0, y, z) < 0$ for $i = 2, 3$,

$$\operatorname{rank}\left(D\mathbf{g}_{\mathbf{E}}(0, y, z)\right) = \operatorname{rank}\left(D(g_0, g_1)^{\mathsf{T}}(0, y, z)\right) = \operatorname{rank}\left(\begin{bmatrix} 1 & 1 & 1 \\ \text{-}1 & 0 & 0 \end{bmatrix}\right) = 2.$$

On the edge where $g_0(x, 0, z) = 1$, $g_2(x, 0, z) = 0$, and $g_i(x, 0, z) < 0$ for $i = 1, 3$,

$$\operatorname{rank}\left(D\mathbf{g}_{\mathbf{E}}(x, 0, z)\right) = \operatorname{rank}\left(D(g_0, g_2)^{\mathsf{T}}(x, 0, z)\right) = \operatorname{rank}\left(\begin{bmatrix} 1 & 1 & 1 \\ 0 & \text{-}1 & 0 \end{bmatrix}\right) = 2.$$

On the edge where $g_0(x, y, 0) = 1$, $g_3(x, y, 0) = 0$, and $g_i(x, y, 0) < 0$ for $i = 1, 2$,

$$\operatorname{rank}\left(D\mathbf{g}_{\mathbf{E}}(x, y, 0)\right) = \operatorname{rank}\left(D(g_0, g_3)^{\mathsf{T}}(x, y, 0)\right) = \operatorname{rank}\left(\begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & \text{-}1 \end{bmatrix}\right) = 2.$$

Finally, at the vertices where three constraints are tight,

$$\operatorname{rank}\Big(D\mathbf{g_E}(1,0,0)\Big) = \operatorname{rank}\Big(D(g_0,g_2,g_3)^{\mathsf{T}}(1,0,0)\Big) = \operatorname{rank}\left(\begin{bmatrix}1 & 1 & 1\\ 0 & -1 & 0\\ 0 & 0 & -1\end{bmatrix}\right) = 3,$$

$$\operatorname{rank}\Big(D\mathbf{g_E}(0,1,0)\Big) = \operatorname{rank}\Big(D(g_0,g_1,g_3)^{\mathsf{T}}(0,1,0)\Big) = \operatorname{rank}\left(\begin{bmatrix}1 & 1 & 1\\ -1 & 0 & 0\\ 0 & 0 & -1\end{bmatrix}\right) = 3,$$

$$\operatorname{rank}\Big(D\mathbf{g_E}(0,0,1)\Big) = \operatorname{rank}\Big(D(g_0,g_1,g_2)^{\mathsf{T}}(0,0,1)\Big) = \operatorname{rank}\left(\begin{bmatrix}1 & 1 & 1\\ -1 & 0 & 0\\ 0 & -1 & 0\end{bmatrix}\right) = 3.$$

This verifies that the constraint qualification is satisfied at all points of the feasible set.

The first order conditions KKT-1,2 are given by

$$\begin{aligned}
0 &= f_x - \lambda_0\, g_{0x} - \lambda_1\, g_{1x} - \lambda_2\, g_{2x} - \lambda_3\, g_{3x} = 2x - \lambda_0 + \lambda_1,\\
0 &= f_y - \lambda_0\, g_{0y} - \lambda_1\, g_{1y} - \lambda_2\, g_{2y} - \lambda_3\, g_{3y} = 4y - \lambda_0 + \lambda_2,\\
0 &= f_z - \lambda_0\, g_{0z} - \lambda_1\, g_{1z} - \lambda_2\, g_{2z} - \lambda_3\, g_{3z} = 6z - \lambda_0 + \lambda_3,\\
1 &= x + y + z, \quad 0 = \lambda_1\, x, \quad 0 = \lambda_2\, y, \quad 0 = \lambda_3\, z.
\end{aligned}$$

Case 1: At an interior point with $x > 0$, $y > 0$, and $z > 0$, $\lambda_i = 0$ for $1 \leq i \leq 3$. Thus

$$\lambda_0 = 2x = 4y = 6z,$$

so $y = {}^{x}\!/_{2}$ and $z = {}^{x}\!/_{3}$. Substituting into g_0, $1 = x + y + z = x\left(1 + {}^{1}\!/_{2} + {}^{1}\!/_{3}\right) = {}^{11\,x}\!/_{6}$, so $x = {}^{6}\!/_{11}$, $y = {}^{3}\!/_{11}$, and $z = {}^{2}\!/_{11}$.

Case 2: $x = 0$, $y > 0$, and $z > 0$. Thus, $\lambda_2 = \lambda_3 = 0$, and we get $4y = \lambda_0 = 6z$, so $z = {}^{2y}\!/_{3}$. Then $1 = y\left(1 + {}^{2}\!/_{3}\right) = {}^{5\,y}\!/_{3}$, and $y = {}^{3}\!/_{5}$. Then $z = \frac{2}{3} \cdot \frac{3}{5} = {}^{2}\!/_{5}$ and $\lambda_0 = 4y = 4\left({}^{3}\!/_{5}\right) = {}^{12}\!/_{5}$. Then, $0 = -\lambda_0 + \lambda_1$, so $\lambda_1 = {}^{12}\!/_{5} > 0$. This is an allowable point.

Case 3: $y = 0$, $x > 0$, and $z > 0$. Thus, $\lambda_1 = \lambda_3 = 0$, and we get $2x = \lambda_0 = 6z$, so $x = 3z$. Then, $1 = z(3 + 1)$, $z = {}^{1}\!/_{4}$, $x = {}^{3}\!/_{4}$, and $\lambda_0 = 2x = {}^{3}\!/_{2} > 0$. Then, $\lambda_2 = \lambda_0 = {}^{3}\!/_{2} > 0$ is allowable.

Case 4: $z = 0$, $x > 0$, and $y > 0$. Thus, $\lambda_1 = \lambda_2 = 0$, and we get $2x = \lambda_0 = 4y$, so $x = 2y$. Then, $1 = y(2 + 1)$, $y = {}^{1}\!/_{3}$, $x = {}^{2}\!/_{3}$, and $\lambda_0 = {}^{4}\!/_{3}$. Then, $\lambda_3 = -\lambda_0 = {}^{4}\!/_{3} > 0$ is allowable.

The vertices $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$ are also possibilities.

The values at these points are as follows:

$$\begin{aligned}
f\left({}^{6}\!/_{11}, {}^{3}\!/_{11}, {}^{2}\!/_{11}\right) &= \frac{36 + 18 + 12}{121} = \frac{66}{121} \approx 0.5454,\\
f\left(0, {}^{3}\!/_{5}, {}^{2}\!/_{5}\right) &= \frac{18 + 12}{25} = \frac{30}{25} = 1.2,\\
f\left({}^{3}\!/_{4}, 0, {}^{1}\!/_{4}\right) &= \frac{9 + 3}{16} = \frac{12}{16} = 0.75,\\
f\left({}^{2}\!/_{3}, {}^{1}\!/_{3}, 0\right) &= \frac{4 + 2}{9} = \frac{2}{3} \approx 0.667,\\
f(1,0,0) &= 1,\\
f(0,1,0) &= 2,\\
f(0,0,1) &= 3.
\end{aligned}$$

The maximal value of 3 is attained at the vertex $(0,0,1)$. ■

There are several aspects that make KKT Theorem 3.12 difficult to apply. First, it is necessary to show that a maximum or minimum exists. Second, it is not easy to show that the constraint qualification holds at all points of $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ or find all the points on the boundary where it fails. Thus, the possibility of the constraint qualification failing makes the theorem difficult to apply in applications. Finally, it is necessary to find all the points that satisfy KKT-1,2,3 or where the constraint qualification fails and compare the values of the objective function at these points.

The deficiencies of Theorem 3.12 are overcome by means of convexity and concavity of constraints and objective function as developed in the next section.

3.3. Exercises

3.3.1. Maximize the revenue

$$\pi = p_1 x_1^{1/2} + p_2 x_1^{1/2} x_2^{1/3}$$

subject to a wealth constraint on the inputs

$$w_1 x_1 + w_2 x_2 \leq C > 0, \qquad x_1 \geq 0 \qquad x_2 \geq 0.$$

a. Write down the constraint functions and the KKT-1,2,3 equations that must be satisfied for the Karush-Kuhn-Tucker Theorem.

b. Take $w_1 = w_2 = 2$, $p_1 = p_2 = 1$, and $C = 8$, and find explicit values of x_1 and x_2 that attains the maximum.

3.3.2. Given $H > 0$, $p_1 > 0$, $p_2 > 0$, and $w > 0$. Let $u(x_1, x_2, y) = x_1^{\alpha_1} x_2^{\alpha_2} - (H - y)^2$, for $\alpha_1 > 0$ and $\alpha_2 > 0$. Consider the feasible set

$$\Phi(p, q, w, H) = \{\, (x_1, x_2, y) \in \mathbb{R}^3_+ : p_1 x_1 + p_2 x_2 + wy \leq wH,\ 2y \leq H \,\}.$$

a. What are the constraint functions g_i which define the inequality constraints in the form of the Kuhn-Tucker Theorem?
Which constraints are effective at $y^* = \frac{1}{2}H$, and $x_i^* = \dfrac{wH\alpha_i}{2p_i(\alpha_1 + \alpha_2)}$ for $i = 1, 2$?
Do the constraints satisfy the rank condition?

b. Why does the feasible set satisfy the Slater condition?

c. What are the KKT-1,2,3 equations?

3.3.3. Consider the following problem:

$$\begin{aligned} \text{Maximize:} \quad & f(x, y) = x^2 + y^2 + 2y \\ \text{Subject to:} \quad & x^2 + y^2 \leq 5 \\ & x + 2y \leq 4 \\ & 0 \leq x,\ 0 \leq y. \end{aligned}$$

a. *Explain why* $f(x, y)$ must have a maximum on the feasible set.

b. Using the fact that the points $(x, y) = (2, 1)$, $(6/5, 7/5)$, and $(0, 2)$ are the only points that satisfy the first order KKT equations for a maximum with the correct signs of the multipliers, *what* is the maximal value of $f(x, y)$ on the feasible set and what is the point that is the maximizer? Explain why this must be the maximizer, including explaining how the theorems apply and what other conditions need to be satisfied.

c. *Verify* that the constraint qualification is satisfied at the three points $(2, 1)$, $(6/5, 7/5)$, and $(0, 2)$.

3.3.4. Assuming the parameters $p > 1$, $w_0 > 0$, $0 < \bar{x}_1 < {}^{w_0}\!/_p$, and $0 < \bar{x}_2 < w_0$, consider the following problem:

$$\begin{aligned} \text{Maximize:} \quad & U(x_1, x_2) = x_1 x_2 \\ \text{Subject to:} \quad & p\,x_1 + x_2 \leq w_0, \\ & 0 \leq x_1 \leq \bar{x}_1, \\ & 0 \leq x_2 \leq \bar{x}_2. \end{aligned}$$

a. Show that the constraint qualification is satisfied on the feasible set.
b. Why must U attain a maximum on the feasible set?
c. What are the KKT-1,2,3 equations?
d. What conditions on the parameters need to be satisfied for U to have a maximum at $X_1 = \bar{x}_1$ and $x_2 = \bar{x}_2$?

3.3.5. Consider the problem

$$\begin{aligned} \text{Maximize:} \quad & f(x, y) = 2 - 2\,y \\ \text{Subject to} \quad & g_1(x, y) = y + (x-1)^3 \leq 0 \\ & g_2(x, y) = \text{-}x \leq 0, \\ & g_3(x, y) = \text{-}y \leq 0. \end{aligned}$$

Carry out the following steps to show that the maximizer is a point at which the constrain qualification fails.

a. By drawing a figure, show that the feasible set is a three sided (nonlinear) region with vertices at $(0, 0)$, $(1, 0)$, and $(0, 1)$.
b. Plot several level curves $f^{\text{-}1}(C)$ of the objective function to your figure from part (a) and conclude geometrically that $(0, 1)$ is a maximizer and $(1, 0)$ is a minimizer of $f(x, y)$ on the feasible set.
c. Show that the constraint qualification fails at $(0, 1)$. Also, show that $Df(0, 1)$ cannot be written as a linear combination of the derivatives $Dg_i(0, 1)$ of the effective constraints.

3.4. Extrema with Inequality Constraints: Sufficient Conditions

We overcome the deficiencies of Theorem 3.12 by means of convexity and concavity. Convex constraints eliminates the need for the constraint qualification. A concave (resp. convex) objective function ensures the objective function has a global maximum (resp. minimum) at a point that satisfies conditions KKT-1,2,3 (resp. KKT-1,2,3′). The convexity and concavity assumptions are like second derivative conditions at all points of the feasible set.

3.4.1. Convex Structures

Before defining convex and concave functions, we repeat the definition of a convex set given in Section 1.6.

Definition. A set $\mathscr{D} \subset \mathbb{R}^n$ is called *convex* provided that if $\mathbf{x}, \mathbf{y}$ are any two points in $\mathscr{D}$, then the convex combination $(1-t)\,\mathbf{x} + t\,\mathbf{y}$ is also in $\mathscr{D}$ for any $0 \leq t \leq 1$. Note that $\mathbf{x}_t = (1-t)\,\mathbf{x} + t\,\mathbf{y}$ for $0 \leq t \leq 1$ is the line segment from $\mathbf{x}$ when $t = 0$ and to $\mathbf{y}$ when $t = 1$.

Figure 3.4.1 shows examples of convex and non-convex sets.

We next define of convex and concave functions in terms of just the values of the function without assuming the function is differentiable. Later in Theorem 3.32, we show that for a C^1

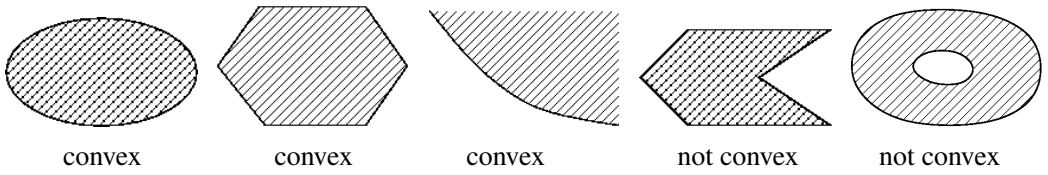


Figure 3.4.1. Convex and non-convex sets

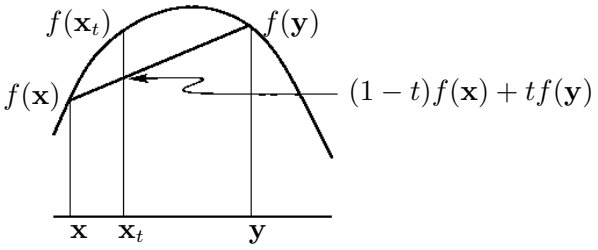


Figure 3.4.2. Concave function

function these conditions can be expressed in terms of the function being above or below the tangent plane at all points.

Definition. A function $f : \mathscr{D} \subset \mathbb{R}^n \to \mathbb{R}$ is *concave* on $\mathscr{D}$ provided that for all $\mathbf{x}, \mathbf{y} \in \mathscr{D}$ and $0 \le t \le 1$,

$$f(\mathbf{x}_t) \ge (1-t)\, f(\mathbf{x}) + t\, f(\mathbf{y}) \qquad \text{for } \mathbf{x}_t = (1-t)\,\mathbf{x} + t\,\mathbf{y}.$$

This is equivalent to assuming that the set of points below the graph,

$$\{(\mathbf{x}, y) \in \mathscr{D} \times \mathbb{R} : y \le f(\mathbf{x})\},$$

is a convex subset of $\mathbb{R}^{n+1}$.

A function $f : \mathscr{D} \to \mathbb{R}$ is *strictly concave* provided that for all $\mathbf{x}, \mathbf{y} \in \mathscr{D}$ with $\mathbf{x} \ne \mathbf{y}$ and $0 < t < 1$,

$$f(\mathbf{x}_t) > (1-t)\, f(\mathbf{x}) + t\, f(\mathbf{y}) \qquad \text{for } \mathbf{x}_t = (1-t)\,\mathbf{x} + t\,\mathbf{y}.$$

A function $f : \mathscr{D} \subset \mathbb{R}^n \to \mathbb{R}$ is *convex* on $\mathscr{D}$ provided that for all $\mathbf{x}, \mathbf{y} \in \mathscr{D}$ and $0 \le t \le 1$,

$$f(\mathbf{x}_t) \le (1-t)\, f(\mathbf{x}) + t\, f(\mathbf{y}) \qquad \text{for } \mathbf{x}_t = (1-t)\,\mathbf{x} + t\,\mathbf{y}.$$

This is equivalent to assuming that the set of points above the graph,

$$\{(\mathbf{x}, y) \in \mathscr{D} \times \mathbb{R} : y \ge f(\mathbf{x})\},$$

is a convex subset of $\mathbb{R}^{n+1}$.

A function $f : \mathscr{D} \to \mathbb{R}$ is *strictly convex* provided that for all $\mathbf{x}, \mathbf{y} \in \mathscr{D}$ with $\mathbf{x} \ne \mathbf{y}$ and $0 < t < 1$,

$$f(\mathbf{x}_t) < (1-t)\, f(\mathbf{x}) + t\, f(\mathbf{y}) \qquad \text{for } \mathbf{x}_t = (1-t)\,\mathbf{x} + t\,\mathbf{y}.$$

Remark. If f is either concave or convex on $\mathscr{D}$ then $\mathscr{D}$ is convex. Also, the condition is on the graph of f and not on the domain of f.

Theorem 3.15. *Let $f : \mathscr{D} \subset \mathbb{R}^n \to \mathbb{R}$ be a concave or convex function on $\mathscr{D}$. Then in the interior of $\mathscr{D}$, f is continuous and possesses all directional derivatives (which can possibly be infinite).*

See [**14**] or [**2**] for a proof. Since concave/convex functions are continuous, they are reasonable functions to maximize or minimize.

Theorem 3.16. *Assume that $\mathscr{D} \subset \mathbb{R}^n$ is an open convex subset, and $g_i : \mathscr{D} \to \mathbb{R}$ are C^1 convex functions for $1 \leq i \leq m$. Then $\mathscr{F}_{\mathbf{g},\mathbf{b}} = \{\, \mathbf{x} \in \mathscr{D} : g_i(\mathbf{x}) \leq b_i \text{ for } 1 \leq i \leq m \,\}$ is a convex set for any $\mathbf{b} \in \mathbb{R}^m$,*

Proof. Take $\mathbf{x}, \mathbf{y} \in \mathscr{F}_{\mathbf{g},\mathbf{b}}$ and let $\mathbf{x}_t = (1-t)\mathbf{x} + t\mathbf{y}$ for $0 \leq t \leq 1$. For any $1 \leq i \leq m$, $g_i(\mathbf{x}_t) \leq (1-t)g_i(\mathbf{x}) + tg_i(\mathbf{y}) \leq (1-t)b_i + tb_i = b_i$, so $\mathbf{x}_t \in \mathscr{F}_{\mathbf{g},\mathbf{b}}$. Thus, $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ is convex. □

The assumption that a function is convex or concave can be verified by a second-derivative condition at all points of the feasible set as shown in the following theorem.

Theorem 3.17 (**Second-Derivative Test**). *Let $\mathscr{D} \subset \mathbb{R}^n$ be open and convex and $f : \mathscr{D} \to \mathbb{R}$ be a C^2 function.*

- **a.** *The function f is convex (respect. convex) on $\mathscr{D}$ iff $D^2 f(\mathbf{x})$ is positive (respect. negative) semidefinite for all $\mathbf{x} \in \mathscr{D}$.*
- **b.** *If $D^2 f(\mathbf{x})$ is positive (respect. negative) definite for all $\mathbf{x} \in \mathscr{D}$, then f is strictly convex (respect. concave) on $\mathscr{D}$.*

We follow the proof in Sundaram [**14**]. The general theorem can be adapted from the special case when $\mathscr{D}$ is all of $\mathbb{R}^n$. The idea is that if $D^2 f(\mathbf{x})$ is positive definite (resp. negative definite), then locally the graph of f lies above (resp. below) the tangent plane. The proof makes this global.

We start by proving a lemma that says it is enough to show f is convex along straight lines in the domain.

Lemma 3.18. *Let $f : \mathbb{R}^n \to \mathbb{R}$. For any $\mathbf{x}, \mathbf{h} \in \mathbb{R}^n$, let $g_{\mathbf{x},\mathbf{h}}(t) = f(\mathbf{x} + t\,\mathbf{h})$ for $t \in \mathbb{R}$. Then the following hold.*

- **a.** *f is convex iff $g_{\mathbf{x},\mathbf{h}}$ is convex for each fixed $\mathbf{x}, \mathbf{h} \in \mathbb{R}^n$.*
- **b.** *If $g_{\mathbf{x},\mathbf{h}}$ is strictly convex for each fixed $\mathbf{x}, \mathbf{h} \in \mathbb{R}^n$ with $\mathbf{h} \neq \mathbf{0}$, then f is strictly convex.*

Proof. First suppose that f is convex. Fix $\mathbf{x}, \mathbf{h} \in \mathbb{R}^n$. For any $t_1, t_2 \in \mathbb{R}$ and any $0 \leq \beta \leq 1$,

$$\begin{aligned} g_{\mathbf{x},\mathbf{h}}(\beta\, t_1 + (1-\beta)\, t_2) &= f\left(\mathbf{x} + \beta\, t_1\mathbf{h} + (1-\beta)\, t_2\mathbf{h}\right) \\ &= f\left(\beta\,(\mathbf{x} + t_1\mathbf{h}) + (1-\beta)\,(\mathbf{x} + t_2\mathbf{h})\right) \\ &\leq \beta\, f\left(\mathbf{x} + t_1\mathbf{h}\right) + (1-\beta)\, f\left(\mathbf{x} + t_2\mathbf{h}\right) \\ &= \beta\, g_{\mathbf{x},\mathbf{h}}(t_1) + (1-\beta)\, g_{\mathbf{x},\mathbf{h}}(t_2). \end{aligned}$$

This shows that $g_{\mathbf{x},\mathbf{h}}$ is convex.

Next, suppose that $g_{\mathbf{x},\mathbf{h}}$ is convex for any $\mathbf{x}, \mathbf{h} \in \mathbb{R}^n$. Pick any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, and let $\mathbf{h} = \mathbf{y} - \mathbf{x}$, and $\mathbf{x}_t = (1-t)\,\mathbf{x} + t\,\mathbf{y} = \mathbf{x} + t\,\mathbf{h}$ for $0 \leq t \leq 1$. Then,

$$\begin{aligned} f\left(\,(1-t)\,\mathbf{x} + t\,\mathbf{y}\,\right) &= g_{\mathbf{x},\mathbf{h}}(t) + g_{\mathbf{x},\mathbf{h}}\left(\,(1-t)0 + t\,1\right) \\ &\leq (1-t)\, g_{\mathbf{x},\mathbf{h}}(0) + t\, g_{\mathbf{x},\mathbf{h}}(1) \\ &= (1-t)\, f(\mathbf{x}) + t\, f(\mathbf{y}). \end{aligned}$$

Since this is true for any $\mathbf{x}, \mathbf{h} \in \mathbb{R}^n$, f is convex.

The proof of (**b**) is similar. □

Proof of Theorem 3.17. We consider part (b) in the case when $D^2 f(\mathbf{x})$ is positive definite and show that f is strictly convex. Pick any $\mathbf{x}, \mathbf{h} \in \mathbb{R}^n$ with $\mathbf{h} \neq \mathbf{0}$, and define $g(t) = g_{\mathbf{x},\mathbf{h}}(t)$ as in the lemma. In the notes for Chapter 6, we showed that

$$g''_{\mathbf{x},\mathbf{h}}(t) = \mathbf{h}^{\mathsf{T}} D^2 f(\mathbf{x}_t)\mathbf{h},$$

which is positive. For any $t < s$ in $\mathbb{R}$ and $0 \leq \beta \leq 1$, let $z = (1-\beta)\,t + \beta\,s$. By the Mean Value Theorem, there exists $t < w_1 < z < w_2 < s$ such that

$$\frac{g(z)-g(t)}{z-t} = g'(w_1) \qquad \text{and} \qquad \frac{g(s)-g(z)}{s-z} = g'(w_2).$$

Since $g''(x) > 0$ for all x, $g'(w_2) > g'(w_1)$ and

$$\begin{aligned} \frac{g(z)-g(t)}{z-t} &< \frac{g(s)-g(z)}{s-z} \\ (s-z)\,(g(z)-g(t)) &< (z-t)\,(g(s)-g(z)) \\ (s-t)g(z) &< (s-z)g(t) + (z-t)g(s) \\ g(z) &< \frac{s-z}{s-t}\,g(t) + \frac{z-t}{s-t}\,g(s). \end{aligned}$$

Since $z = (1-\beta)\,t+\beta\,s$, $\dfrac{s-z}{s-t} = \beta$, and $1-\beta = \dfrac{z-t}{s-t}$, this gives $g\left((1-\beta)\,t+\beta\,s\right) < \beta\,g(t) + (1-\beta)\,g(s)$. This proves that g is strictly convex. By Lemma 3.18, f is strictly convex. □

3.4.2. Karush-Kuhn-Tucker Theorem under Convexity

The main theorem of this section gives necessary and sufficient conditions for an extremizer to exist on a feasible set. To ensure that an extremizer satisfies the conditions KKT-1,2, we need a condition on the feasible set $\mathscr{F}_{\mathbf{g},\mathbf{b}}$.

Definition. Let $g_i : \mathscr{D} \to \mathbb{R}$ for $1 \leq i \leq m$ and $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ as usual. We say that the constraint functions g_i satisfy the *Slater condition* for $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ provided that there exists a point $\overline{\mathbf{x}} \in \mathscr{F}_{\mathbf{g},\mathbf{b}}$ such that $g_i(\overline{\mathbf{x}}) < b_i$ for all $1 \leq i \leq m$. (This assumption says that there is a point with no effective constraint and implies that the constraint set has nonempty interior.)

Theorem 3.19 (Karush-Kuhn-Tucker Theorem under Convexity). *Assume that $\mathbf{b} \in \mathbb{R}^m$, $f, g_i : \mathbb{R}^n \to \mathbb{R}$ are C^1 for $1 \leq i \leq m$, and $\mathbf{x}^* \in \mathscr{F}_{\mathbf{g},\mathbf{b}}$.*

- **a.** *Assume that f is a concave function.*
  - **i.** *If $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ is convex and $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ satisfies KKT-1,2,3 with all the $\lambda_i^* \geq 0$, then f has a maximum on $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ at $\mathbf{x}^*$.*
  - **ii.** *If f has a maximum on $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ at $\mathbf{x}^*$, each of the constraints functions g_i is convex, and $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ satisfies the Slater condition, then there exist $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_m^*)$ such that $(\mathbf{x}^*, \lambda^*)$ satisfies conditions KKT-1,2,3 with all the $\lambda_i^* \geq 0$.*
- **b.** *Assume that f is a convex function.*
  - **i.** *If $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ is convex and $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ satisfies KKT-1,2,3′ with all the $\lambda_i^* \leq 0$, then f has a minimum on $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ at $\mathbf{x}^*$.*
  - **ii.** *If f has a minimum on $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ at $\mathbf{x}^*$, each of the constraints functions g_i is convex, and $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ satisfies the Slater condition, then there exist $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_m^*)$ such that $(\mathbf{x}^*, \lambda^*)$ satisfies conditions KKT-1,2,3′ with all the $\lambda_i^* \leq 0$.*

  (Note that the assumptions on $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ and the g_i stay the same for a minimum as for a maximum.)

Remark. Kuhn and Tucker wrote a paper that popularized this result in 1951. However, Karush had earlier written a thesis at the University of Chicago about a similar result in 1939. Therefore, although this result is sometimes referred to as the Kuhn-Tucker Theorem, people have started using all three names when referring to this theorem. Fritz John had a related result in 1948. See [**2**].

Remark. In most of the examples we use part **a.i** (or **b.i**). Note that for this use it is not necessary to verify Slater's condition or that a maximum must exist (by compactness or similar argument). The feasible set $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ is shown convex by conditions on constraint function. We have shown that if all the $g_i(\mathbf{x})$ are convex then it is true. Later, we allow for rescaling of convex functions. Finally, once one point is found that satisfies KKT-1,2,3, then it is automatically a maximizer; we do not need to verify separately that a maximum exists.

We delay the proof of the Karush-Kuhn-Tucker Theorem under Convexity until Section 3.4.5. In the rest of this section, we give examples of convex/concave functions and applications of the Karush-Kuhn-Tucker Theorem.

Proposition 3.20. *For $\mathbf{a} \in \mathbb{R}^n$ and $b \in \mathbb{R}$, the affine function on $\mathbb{R}^n$ given by $g(\mathbf{x}) = \mathbf{a} \cdot \mathbf{x} + b = a_1x_1 + \cdots + a_nx_n + b$ is both concave and convex.*

Proof. For $\mathbf{p}_0, \mathbf{p}_1 \in \mathbb{R}^n$, set $\mathbf{p}_t = (1-t)\,\mathbf{p}_0 + t\,\mathbf{p}_1$. Then,

$$\begin{aligned} g(\mathbf{p}_t) &= \mathbf{a} \cdot \mathbf{p}_t + b \\ &= \mathbf{a} \cdot [(1-t)\,\mathbf{p}_0 + t\,\mathbf{p}_1] + b \\ &= (1-t)\ [\mathbf{a} \cdot \mathbf{p}_0 + b] + t\ [\mathbf{a} \cdot \mathbf{p}_1 + b] \\ &= (1-t)\,g(\mathbf{p}_0) + t\,g(\mathbf{p}_1). \end{aligned}$$

Thus, we have equality and not just an inequality and g is both concave and convex. □

Example 3.21.

$$\begin{aligned} \text{Minimize :}\quad & f(x,y) = x^4 + y^4 + 12\,x^2 + 6\,y^2 - xy - x + y, \\ \text{Subject to :}\quad & g_1(x,y) = -x - y \leq \text{-}6 \\ & g_2(x,y) = -2x + y \leq \text{-}3 \\ & \text{-}x \leq 0,\ \text{-}y \leq 0. \end{aligned}$$

The constraints are linear and so are convex. The objective function has second derivative

$$D^2 f(x,y) = \begin{bmatrix} 12\,x^2 + 24 & \text{-}1 \\ \text{-}1 & 12\,y^2 + 12 \end{bmatrix}.$$

The determinant is greater than $24(12) - 1 > 0$, so this is positive definite and f is convex.

Let λ_1 and λ_2 be the multipliers for the first two inequalities and μ_1 and μ_2 be the multiplier for $\text{-}x \leq 0$ and $\text{-}y \leq 0$. Condition KKT-1 is

$$\begin{aligned} 0 &= 4x^3 + 24x - y - 1 + \lambda_1 + 2\lambda_2 + \mu_1 \\ 0 &= 4y^3 + 12y - x + 1 + \lambda_1 - \lambda_2 + \mu_2 \end{aligned}$$

If $x = 0$, then $y \leq \text{-}3$, so $\text{-}y > 0$, which is not feasible. Therefore this constraint cannot be tight and $\mu_1 = 0$. We consider the slackness of the other constraints in cases.

If $y = 0$, then $x = \text{-}g_1(x,0) \geq 6$, so $g_2(x,0) = -2x \leq \text{-}12 < \text{-}3$, and $\lambda_2 = 0$. The function $f(x,0) = x^4 + 12\,x^2 - x$, is minimized for $x \geq 6$ at $x = 6$. For that value, the second equation gives

$$\begin{aligned} 0 &= -6 + 1 + \lambda_1 + \mu_2, \qquad \text{or} \\ 5 &= \lambda_1 + \mu_2. \end{aligned}$$

Both these multipliers cannot be negative, so it is not a minimum.

Finally assume that $x, y > 0$, so $\mu_2 = 0$. Points where both g_1 and g_2 are tight satisfies

$$\begin{aligned} \text{-}6 &= g_1(x,y) = -x - y \\ \text{-}3 &= g_2(x,y) = -2x + y \end{aligned}$$

can be solved to yield $x = y = 3$. If this is a solution of the KKT conditions then

$$0 = 4(3^4) + 24(3) - 3 - 1 + \lambda_1 + 2\lambda_2 = \lambda_1 + 2\lambda_2 + 176$$
$$0 = 4(3^3) + 12(3) - 3 + 1 + \lambda_1 - \lambda_2 = \lambda_1 - \lambda_2 + 140.$$

These can be solved to yield $\lambda_1 = -152$ and $\lambda_2 = -12$. Thus, the point $(x^*, y^*) = (3, 3)$ with $\lambda_1 = -152$, $\lambda_2 = -12$, and $\mu_1 = \mu_2 = 0$ does satisfy the KKT conditions with negative multipliers and is the minimizer. ■

Example 3.22 (Lifetime of Equipment). This example is based on Example 7.5.4 in Walker [**16**]. Assume there are two machines with initial costs of \$1,600 and \$5,400 respectively. The operating expense in the j^{th} year of the two machines is $\$50j$ and $\$200j$ respectively. The combined number of years of use of the two machines is desired to be at least 20 years. The problem is to determine the number of years to use each machine to minimize the average total cost per year.

Let x and y be the lifetimes of the respective machines. We require $x+y \geq 20$, $x \geq 0$, and $y \geq 0$. The average amortized capital expense per year for the two machines is $1600/x + 5400/y$. The average operating expenses per year for the two machines are

$$\frac{50 + 2(50) + \cdots + x(50)}{x} = \frac{50}{x} \cdot \frac{x(x+1)}{2} = 25(x+1) \qquad \text{and}$$
$$\frac{200 + 2(200) + \cdots + y(200)}{y} = \frac{200}{y} \cdot \frac{y(y+1)}{2} = 100(y+1).$$

The problem is therefore the following.

$$\begin{aligned} \text{Minimize :} \quad & f(x,y) = 25(x+1) + 100(y+1) + \frac{1600}{x} + \frac{5400}{y} \\ \text{Subject to :} \quad & g_1(x,y) = 20 - x - y \leq 0 \\ & g_2(x,y) = -x \leq 0 \\ & g_3(x,y) = -y \leq 0. \end{aligned}$$

The constraints are linear and so convex. A direct check shows that $D^2 f(x,y)$ is positive definite on $\mathbb{R}^2_{++}$ and so f is convex on $\mathbb{R}^2_{++}$.

The objective function (cost) gets arbitrarily large near $x = 0$ or $y = 0$, so the minimum occurs for $x > 0$ and $y > 0$, and we can ignore those constraints and multipliers.

The KKT-1,2 conditions become

$$0 = 25 - \frac{1600}{x^2} + \lambda$$
$$0 = 100 - \frac{5400}{y^2} + \lambda$$
$$0 = \lambda(20 - x - y).$$

(**i**) First assume the constraint is effective and $y = 20 - x$. The first two equations give

$$\begin{aligned} -\frac{\lambda}{25} = 1 - \frac{64}{x^2} &= 4 - \frac{216}{(20-x)^2}, \\ 0 &= 3x^2(20-x)^2 + 64(20-x)^2 - 216x^2 \\ &= 3x^4 - 120x^3 + 1048x^2 - 2560x + 25600. \end{aligned}$$

This polynomial has positive roots of $x \approx 12.07$ and 28.37. If $x \approx 28.37$, then $y = -8.37 < 0$ so this is not feasible. If $x \approx 12.07$, then $y \approx 7.93$ and $\lambda \approx -25 + \dfrac{1600}{12.07^2} \approx -14.02 < 0$. The multiplier is negative, so this is the minimizer.

(**ii**) Although we do not need to do so, we check that there is no minimizer where the constraint is not effective and $\lambda = 0$.

$$x^2 = \frac{1600}{25} = 64, \qquad x = 8 \qquad \text{and}$$
$$y^2 = 54, \qquad y \approx 7.34.$$

Since $8 + 7.34 < 20$, they do not satisfy the constraint. ■

Proposition 3.23 (Cobb-Douglas in $\mathbb{R}^2$). *Let $f(x, y) = x^a y^b$ with $a, b > 0$. If $a + b \leq 1$, then f is concave on $\mathbb{R}^2_+$ (and $-x^a y^b$ is convex). If $a + b > 1$, then f is neither concave nor convex.*

See Figure 3.4.3.

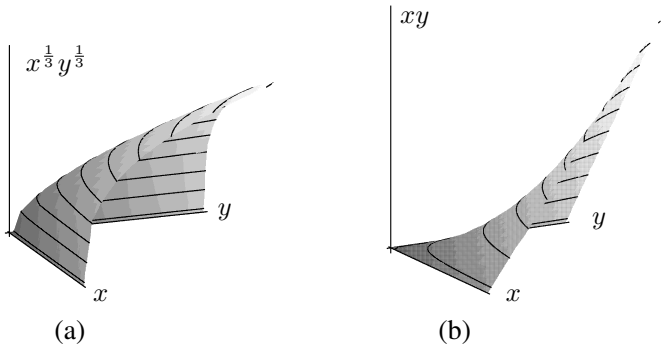


Figure 3.4.3. Cobb-Douglas function for (a) $a = b = 1/3$, (b) $a = b = 1$

Proof.

$$\det\left(D^2 f(x,y)\right) = \det \begin{pmatrix} a(a-1)x^{a-2}y^b & abx^{a-1}y^{b-1} \\ abx^{a-1}y^{1-b} & b(b-1)x^a y^{b-2} \end{pmatrix}$$
$$= ab(1-a-b)x^{2a-2}y^{2b-2} \begin{cases} > 0 & \text{if } a+b<1 \\ = 0 & \text{if } a+b=1 \\ < 0 & \text{if } a+b>1. \end{cases}$$

If $a + b < 1$ (so $a < 1$), $D^2 f(x, y)$ is negative definite and f is strictly concave on $\mathbb{R}^2_{++}$; since f is continuous on $\mathbb{R}^2_+$, it is also concave on $\mathbb{R}^2_+$. If $a + b = 1$, $D^2 f(x, y)$ is negative semidefinite and f is concave on $\mathbb{R}^2_+$; if $a + b > 1$, the $D^2 f(x, y)$ is indefinite and f is neither concave nor convex. For example, xy is neither concave nor convex. □

Proposition 3.24 (Cobb-Douglas in $\mathbb{R}^n$). *Assume that $a_1 + \cdots + a_n < 1$ and $a_i > 0$ for $1 \leq i \leq n$. Then the function $f(\mathbf{x}) = x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}$ is concave on $\mathbb{R}^n_+$ (and $-x_1^{a_1} \cdots x_n^{a_n}$ is convex). If $a_1 + \cdots + a_n > 1$, then f is neither concave nor convex.*

Proof. The reader should try to carry out the calculations directly for the case $n = 3$.

On $\mathbb{R}^n_{++}$, the partial derivatives of f are as follows for $i \neq j$:

$$f_{x_i} = a_i x_1^{a_1} \cdots x_i^{a_i - 1} \cdots x_n^{a_n}$$
$$f_{x_i x_i} = a_i(a_i - 1)x_1^{a_1} \cdots x_i^{a_i - 2} \cdots x_n^{a_n} = a_i(a_i - 1)x_i^{-2} f$$
$$f_{x_i x_j} = a_i a_j x_1^{a_1} \cdots x_i^{a_i - 1} \cdots x_j^{a_j - 1} \cdots x_n^{a_n} = a_i a_j x_i^{-1} x_j^{-1} f.$$

Using linearity on rows and columns, the determinant of the k^{th}-principal submatrix is as follows:

$$\begin{aligned}\Delta_k &= \det \begin{bmatrix} a_1(a_1-1)x_1^{-2}f & \cdots & a_1a_kx_1^{-1}x_k^{-1} \\ \vdots & \ddots & \vdots \\ a_ka_1x_k^{-1}x_1^{-1}f & \cdots & a_k(a_k-1)x_k^{-2}f \end{bmatrix} \\ &= a_1\cdots a_kx_1^{-2}\cdots x_k^{-2}f^k \det \begin{bmatrix} a_1-1 & \cdots & a_k \\ \vdots & \ddots & \vdots \\ a_1 & \cdots & a_k-1 \end{bmatrix} \\ &= a_1\cdots a_kx_1^{-2}\cdots x_k^{-2}f^k\,\bar{\Delta}_k,\end{aligned}$$

where the last equality defines $\bar{\Delta}_k$ as the determinant of the previous matrix. Below, we show by induction that $\bar{\Delta}_k = (-1)^k(1-a_1-\cdots-a_k)$. Once this is established, since signs of the Δ_k alternate as required, D^2f is negative definite on $\mathbb{R}^n_{++}$, and f is strictly concave on $\mathbb{R}^n_{++}$. Since f is continuous, it is concave on the closure of $\mathbb{R}^n_{++}$, i.e., on $\mathbb{R}^n_+$.

We show that $\bar{\Delta}_k = (-1)^k + (-1)^{k-1}(a_1+\cdots+a_k)$ by induction on k, using linearity of the determinant on the last column and column operations on the subsequent matrix:

$$\begin{aligned}\bar{\Delta}_k &= \det \begin{bmatrix} a_1-1 & a_2 & \cdots & a_{k-1} & a_k \\ a_1 & a_2-1 & \cdots & a_{k-1} & a_k \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_1 & a_2 & \cdots & a_{k-1}-1 & a_k \\ a_1 & a_2 & \cdots & a_{k-1} & a_k-1 \end{bmatrix} \\ &= \det \begin{bmatrix} a_1-1 & a_2 & \cdots & a_{k-1} & a_k \\ a_1 & a_2-1 & \cdots & a_{k-1} & a_k \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_1 & a_2 & \cdots & a_{k-1}-1 & a_k \\ a_1 & a_2 & \cdots & a_{k-1} & a_k \end{bmatrix} + \det \begin{bmatrix} a_1-1 & a_2 & \cdots & a_{k-1} & 0 \\ a_1 & a_2-1 & \cdots & a_{k-1} & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_1 & a_2 & \cdots & a_{k-1}-1 & 0 \\ a_1 & a_2 & \cdots & a_{k-1} & -1 \end{bmatrix} \\ \bar{\Delta}_k &= a_k \begin{bmatrix} -1 & 0 & \cdots & 0 & 1 \\ 0 & -1 & \cdots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 1 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix} - \bar{\Delta}_{k-1} \\ &= (-1)^{k-1}a_k - \left[(-1)^{k-1} + (-1)^{k-2}(a_1+\cdots+a_{k-1})\right] \\ &= (-1)^k + (-1)^{k-1}(a_1+\cdots+a_k).\end{aligned}$$

This proves the claim by induction and completes the proof. □

3.4.3. Rescaled Convex Functions

If the constraint functions g_i are convex, then the feasible set $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ is convex. Various books on optimization, including Sundaram [**14**] and Bazaraa et al [**2**], weaken the convexity assumptions on the constraint functions. A function $g : \mathbb{R}^n \to \mathbb{R}$ is *quasi-convex* provided that $\{\,\mathbf{x} \in \mathbb{R}^n : g(\mathbf{x}) \le b\,\}$ is convex for every $b \in \mathbb{R}$. Bazaraa et al [**2**] also weaken the assumption on the objective function to be *pseudo-concave* or *pseudo-convex*. (See the definition

following Theorem 3.32.) Rather than focusing on quasi-convex and pseudo-convex functions, we consider rescalings of convex functions. The next theorem shows that constraints that are rescaled convex functions have a convex feasible set and so are quasi-convex. Then Corollary 3.26 shows that we can use more general exponents in Cobb-Douglas functions than the preceding two proposition allowed.

Definition. A function $g : \mathscr{D} \subset \mathbb{R}^n \to \mathbb{R}$ is a *rescaling* of $\hat{g} : \mathscr{D} \to \mathbb{R}$ provided that there is an increasing function $\phi : \mathbb{R} \to \mathbb{R}$ such that $g(\mathbf{x}) = \phi \circ \hat{g}(\mathbf{x})$. Note that since ϕ has an inverse, $\hat{g}(\mathbf{x}) = \phi^{-1} \circ g(\mathbf{x})$.

We say that ϕ is a C^1 *rescaling* provided that ϕ is C^1 and $\phi'(y) > 0$ for all $y \in \mathbb{R}$.

If $g : \mathscr{D} \subset \mathbb{R}^n \to \mathbb{R}$ is a rescaling of a convex (resp. concave) function $\hat{g} : \mathscr{D} \to \mathbb{R}$, then we say that $g(\mathbf{x})$ is a *rescaled convex function* (resp. *rescaled concave function*). Similarly, if $g : \mathscr{D} \subset \mathbb{R}^n \to \mathbb{R}$ is a C^1 rescaling of a convex (resp. concave) function $\hat{g} : \mathscr{D} \to \mathbb{R}$, then we say that $g(\mathbf{x})$ is a C^1 *rescaled convex function* (resp. C^1 *rescaled concave function*).

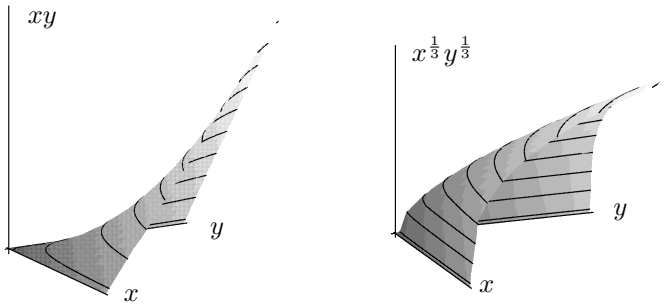
Theorem 3.25 (Rescaling). *Assume that $g : \mathscr{D} \subset \mathbb{R}^n \to \mathbb{R}$ is a rescaling of a convex function $\hat{g} : \mathscr{D} \to \mathbb{R}$, $g(\mathbf{x}) = \phi \circ \hat{g}(\mathbf{x})$. Then $\mathscr{F}_{g,b}$ is convex for any $b \in \mathbb{R}$, and g is quasi-convex.*

The proof follows because $\mathscr{F}_{g,b} = \{\, \mathbf{x} \in \mathscr{D} : \hat{g}(\mathbf{x}) \leq \phi^{-1}(b) \,\}$ is convex.

Proposition 3.26 (Cobb-Douglas). *If $a_1, \ldots, a_n > 0$, then $g(\mathbf{x}) = x_1^{a_1} \cdots x_n^{a_n}$ is a C^1 rescaling of a C^1 concave function on $\mathbb{R}^n_{++}$ and $-g(\mathbf{x})$ is a C^1 rescaling of a C^1 convex function on $\mathbb{R}^n_{++}$.*

Proof. Let $A = a_1 + \cdots + a_n$ and $b_i = {}^{a_i}\!/_{(2A)}$, for $1 \leq i \leq n$, so $b_1 + \cdots + b_n = \frac{1}{2} < 1$. Then $\hat{g}(x, y, z) = x_1^{b_1} \cdots x_n^{b_n}$ is a C^1 concave function on $\mathbb{R}^n_{++}$. The function $\phi(y) = y^{2A}$ is a C^1 rescaling such that $\phi \circ \hat{g}(\mathbf{x}) = g(\mathbf{x})$ on $\mathbb{R}_{++}$. □

Example 3.27. The Cobb-Douglas function $f(x, y) = xy$ is a rescaled concave function, but not concave. See Figure 3.4.4. ■



(a) Non-concave (b) Concave

Figure 3.4.4. Cobb-Douglas function for (a) $a = b = 1$, (b) $a = b = {}^1\!/_3$

Example 3.28. The function $f(x) = x^6 - 2.9\,x^4 + 3\,x^2$ has its graph given in Figure 3.4.5. The derivative is $f'(x) = x[6\,x^4 - 11.6\,x^2 + 6]$, and $f(x)$ has a single critical point at $x = 0$. The second derivative $f''(x) = 30\,x^4 - 34.8\,x^2 + 6$ has zeroes at ± 0.459 and ± 0.974, and $f''(x) < 0$ for $x \in [-0.974, -0.459] \cup [0.459, 0.974]$. Thus, $f(x)$ is not convex. The function $\hat{f}(x) = x^6$ is convex. Using the inverse of $f(x)$ for positive values of x, $\phi(y) = [f^{-1}(y)]^6$ satisfies $\phi \circ f(x) = \hat{f}(x)$ and is a rescaling of $f(x)$ to $\hat{f}(x)$. Thus, $f(x)$ is a rescaled convex function that is not convex. ■

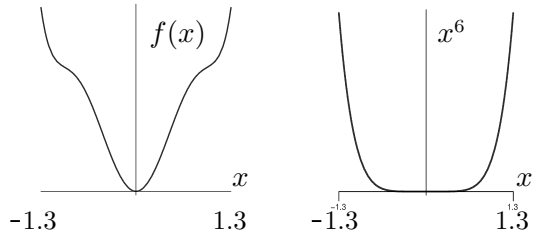


Figure 3.4.5. A rescaled convex function of Example 3.28

The following corollary of the KKT Theorem 3.19 allows us to rescale the objective function as well as the constraints. Thus the general Cobb-Douglas functions considered in the last proposition can be used as objective functions even though they are not concave, and the KKT Theorem 3.19 does not apply.

Corollary 3.29 (KKT for Rescaled Functions). *Assume that* $g_i : \mathscr{D} \subset \mathbb{R}^n \to \mathbb{R}$ *are* C^1 *for* $1 \leq i \leq m$, *each of the constraints* g_i *is a* C^1 *rescaled convex function, and* $\mathbf{x}^* \in \mathscr{F}_{\mathbf{g},\mathbf{b}} = \{\, \mathbf{x} \in \mathscr{D} : g_i(\mathbf{x}) \leq b_i \text{ for } 1 \leq i \leq m \,\}$.

- **a.** *Assume that* $f : \mathbb{R}^n \to \mathbb{R}$ *is a* C^1 *rescaled concave function.*
  - **i.** *If* $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ *satisfies KKT-1,2,3 with all the* $\lambda_i^* \geq 0$, *then* f *has a maximum on* $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ *at* $\mathbf{x}^*$.
  - **ii.** *If* f *has a maximum on* $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ *at* $\mathbf{x}^*$, *and* $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ *satisfies the Slater condition, then there exist* $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_m^*)$ *such that* $(\mathbf{x}^*, \lambda^*)$ *satisfies conditions KKT-1,2,3 with all the* $\lambda_i^* \geq 0$.
- **b.** *Assume that* f *is a* C^1 *rescaled convex function.*
  - **i.** *If* $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ *satisfies KKT-1,2,3′ with all the* $\lambda_i^* \leq 0$, *then* f *has a minimum on* $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ *at* $\mathbf{x}^*$.
  - **ii.** *If* f *has a minimum on* $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ *at* $\mathbf{x}^*$, *and* $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ *satisfies the Slater condition, then there exist* $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_m^*)$ *such that* $(\mathbf{x}^*, \lambda^*)$ *satisfies conditions KKT-1,2,3′ with all the* $\lambda_i^* \leq 0$.

Proof. Assume $\hat{g}_i(\mathbf{x}) = \phi_i \circ g_i(\mathbf{x})$ with $\hat{g}$ a convex C^1 function, $\phi_i : \mathbb{R} \to \mathbb{R}$ C^1, and $\phi_i'(b_i) > 0$ for all $b_i \in \mathbb{R}$. Similarly, assume $\hat{f}$, $\hat{f}(\mathbf{x}) = T \circ f(\mathbf{x})$ with $\hat{f}$ a concave (respect. convex) C^1 function, $T : \mathbb{R} \to \mathbb{R}$ C^1, and $T'(y) > 0$ for all $y = f(\mathbf{x})$ with $\mathbf{x} \in \mathscr{F}_{\mathbf{g},\mathbf{b}}$.

Let $b_i' = \phi_i(b_i)$. If $g_i(\mathbf{x}^*) = b_i$ is tight, then $\hat{g}_i(\mathbf{x}^*) = b_i'$,

$$D\hat{g}_i(\mathbf{x}^*) = \phi_i'(b_i') Dg_i(\mathbf{x}^*) \qquad \text{and} \qquad D\hat{f}(\mathbf{x}^*) = T'(f(\mathbf{x}^*))\, D\hat{f}(\mathbf{x}^*).$$

(a.i) $\mathscr{F}_{\mathbf{g},\mathbf{b}} = \{\, \mathbf{x} \in \mathbf{U} : g_i(\mathbf{x}) \leq b_i \,\} = \{\, \mathbf{x} \in \mathbf{U} : \hat{g}_i(\mathbf{x}) \leq b_i' \,\}$ is convex. If f satisfies KKT-1. then

$$D\hat{f}(\mathbf{x}^*) = T'(f(\mathbf{x}^*))\, Df(\mathbf{x}^*) = T'(f(\mathbf{x}^*)) \sum_i \lambda_i\, Dg_i(\mathbf{x}^*),$$

so $\hat{f}$ satisfies KKT-1,2 with multipliers $T'(f(\mathbf{x}^*))\,\lambda_i > 0$. By Theorem 3.19(a.i), $\hat{f}$ has a maximum at $\mathbf{x}^*$. Since T is increasing, f has a maximum at $\mathbf{x}^*$.

(a.ii) If f has a maximum at $\mathbf{x}^*$ then since T is increasing, $\hat{f}$ has a maximum at $\mathbf{x}^*$. Applying Theorem 3.19(a.ii) to $\hat{f}$ and the $\hat{g}_i$ on $\mathscr{F}_{\mathbf{g},\mathbf{b}}$, we get that

$$T'(f(\mathbf{x}^*))\, Df(\mathbf{x}^*) = D\hat{f}(\mathbf{x}^*) = \sum_i \lambda_i\, D\hat{g}_i(\mathbf{x}^*) = \sum_i \lambda_i\, \phi_i'(b_i) Dg_i(\mathbf{x}^*),$$

where we have use the fact that $\lambda_i = 0$ unless $g_i(\mathbf{x}^*) = b_i$. Since, $T'(f(\mathbf{x}^*)) > 0$ and $\phi_i'(b_i) > 0$ for all effective i, we get that conditions KKT-1.2 hold for f and the g_i with multipliers $\lambda_i\, T_i'(b_i) \big/ T'(f(\mathbf{x}^*)) > 0$. □

Example 3.30. Find the maximum of $f(x,y,z) = xyz$ subject to $g_1(x,y,z) = 2x + y + 2z - 5 \leq 0$, $g_2(x,y,z) = x + 2y + z - 4 \leq 0$, $g_3(x,y,z) = \text{-}x \leq 0$, $g_4(x,y,z) = \text{-}y \leq 0$, and $g_5(x,y,z) = \text{-}z \leq 0$.

The feasible set is $\mathscr{F} = \{\,(x,y,z) \in \mathbb{R}^3 : g_i(x,y,z) \leq 0 \text{ for } 1 \leq i \leq 5\,\}$. The constraints are linear so convex and $\mathscr{F}$ is convex. The objective function f is a C^1 rescaled concave function on $\mathbb{R}^3_+$. We could maximize the function $(xyz)^{\frac{1}{4}}$, but the KKT equations are more complicated for this objective function.

Since the values with at least one variable zero is zero, $0 = f(0,y,z) = f(x,0,z) = f(x,y,0)$, and $f(x,y,z) > 0$ on $\mathbb{R}^3_{++}$, the maximum occurs in $\mathbb{R}^3_{++} \cap \mathscr{F}$. Therefore, the constraints g_i for $3 \leq i \leq 5$ are slack at the maximum and $\lambda_3 = \lambda_4 = \lambda_5 = 0$.

On $\mathbb{R}^n_{++}$, the conditions KKT-1,2 are

$$\begin{aligned} yz &= 2\lambda_1 + \lambda_2, && \text{(KKT-1)}\\ xz &= \lambda_1 + 2\lambda_2,\\ xy &= 2\lambda_1 + \lambda_2,\\ 0 &= \lambda_1(5 - 2x - y - 2z), && \text{(KKT-2)}\\ 0 &= \lambda_2(4 - x - 2y - z). \end{aligned}$$

From the first and third equation, we see that $yz = xy$, so $x = z$. The two remaining complementary slackness equations are

$$\begin{aligned} 0 &= \lambda_1(5 - 4x - y)\\ 0 &= \lambda_2(4 - 2x - 2y). \end{aligned}$$

If both constraints are effective, then we can solve the equations

$$\begin{aligned} 5 &= 4y + y\\ 2 &= x + y, \end{aligned}$$

to get that $x = y = z = 1$. For this point, the first two equations of KKT-1 become

$$\begin{aligned} 1 &= 2\lambda_1 + \lambda_2\\ 1 &= \lambda_1 + 2\lambda_2, \end{aligned}$$

which have a solution $\lambda_1 = \lambda_2 = 1/3 > 0$. (It can be checked that there is no other solution of the first order KKT conditions on $\mathbb{R}^3_{++} \cap \mathscr{F}$, i.e., when one or two of the constraints are effective, but this is not necessary.)

We have shown that f is a rescaling of a concave function, all the g_i are all convex functions on $\mathbb{R}^3_+$, and that $\mathbf{p}^* = (1,1,1)$ multipliers $\lambda_1 = \lambda_2 = 1/3 > 0$ satisfy satisfies conditions KKT-1,2,3. By the Karush-Kuhn-Tucker Theorem under Convexity, f must have a maximum on $\mathbb{R}^3_{++} \cap \mathscr{F}$ at $\mathbf{p}^*$. But as we remarked earlier, this must be the maximum on all of $\mathscr{F}$ since $f(x,y,z) = 0$ when one or more variable is zero. ■

Remark. Although we do not need it, the feasible set $\mathscr{F}$ satisfies the Slater condition since the point $(0.5, 0.5, 0.5)$ in $\mathscr{D} = \{\,(x,y,z) \in \mathbb{R}^3_+ : g_i(x,y,z) \leq 0 \text{ for } 1 \leq i \leq 5\,\}$ has all the g_i positive, the constraint functions satisfy the Slater condition.

We could also check that the constraint qualification is indeed satisfied on all of $\mathscr{F}$. However, if we add another constraint, $x + y + z - 3 \leq 0$, then $\mathbf{p}^*$ can be shown to be a solution of KKT-1,2,3 in $\mathbb{R}^3_{++}$. By the Karush-Kuhn-Tucker Theorem under Convexity, $\mathbf{p}^*$ is a maximizer. For this example, there are three effective constraints at $\mathbf{p}^*$, but the rank is still 2. Therefore, this system does not satisfy the constraint qualification.

3.4.4. Global Extrema for Concave Functions

We return to general facts about concave/convex functions that are the main steps to show that a solution of KKT-1,2,3 is a maximizer. If $M = \max\{ f(\mathbf{x}) : \mathbf{x} \in \mathscr{F} \} < \infty$ exists, we denote the set of maximizers by

$$\mathscr{F}^* = \{ \mathbf{x} \in \mathscr{F} : f(\mathbf{x}) = M \}.$$

If there is no maximum then $\mathscr{F}^*$ is the empty set.

Theorem 3.31. *Assume that $f : \mathscr{F} \subset \mathbb{R}^n \to \mathbb{R}$ is concave. Then the following hold:*

- **a.** *Any local maximizer of f is a global maximizer.*
- **b.** *The set of maximizers $\mathscr{F}^*$ is either empty or convex.*
- **c.** *If f is strictly concave, then $\mathscr{F}^*$ is either empty or a single point.*

Proof. (**a**) If not, then there exists a local maximizer $\mathbf{x}^*$ and $\mathbf{z} \neq \mathbf{x}^*$ such that $f(\mathbf{z}) > f(\mathbf{x}^*)$. For $\mathbf{x}_t = (1-t)\,\mathbf{x}^* + t\,\mathbf{z}$ and $0 < t < 1$,

$$f(\mathbf{x}_t) \geq (1-t)f(\mathbf{x}^*) + t\,f(\mathbf{z}) > (1-t)f(\mathbf{x}^*) + t\,f(\mathbf{x}^*) = f(\mathbf{x}^*).$$

Since $f(\mathbf{x}_t) > f(\mathbf{x}^*)$ for small t, f cannot have a local maximum at $\mathbf{x}^*$.

(**b**) Let $M = \max\{ f(\mathbf{x}) : \mathbf{x} \in \mathscr{F} \}$, $\mathbf{x}_0, \mathbf{x}_1 \in \mathscr{F}^*$, and $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$. Then $M \geq f(\mathbf{x}_t) \geq (1-t)f(\mathbf{x}_0) + t\,f(\mathbf{x}_1) = (1-t)M + tM = M$, so $f(\mathbf{x}_t) = M$ and $\mathbf{x}_t$ is in the set of optimizing points for all $0 \leq t \leq 1$. This proves the desired convexity.

(**c**) The proof is like part (b) except there is a strict inequality if $\mathbf{x}_0 \neq \mathbf{x}_1$. This gives a contradiction $M > M$, so there can be only one point. □

The next theorem shows that for C^1 functions, a function is convex if and only if a condition on the relationship between the function and its tangent plane is satisfied. This first order condition is not used to check convexity or concavity of a function but is used in the succeeding theorem to give first order conditions for a maximizer.

Theorem 3.32. *Let $\mathscr{D} \subset \mathbb{R}^n$ be open and convex. Assume $f : \mathscr{D} \to \mathbb{R}$ is C^1.*

- **a.** *f is convex iff $f(\mathbf{x}) \geq f(\mathbf{p}) + Df(\mathbf{p})(\mathbf{x}-\mathbf{p})$ for all $\mathbf{x}, \mathbf{p} \in \mathscr{D}$ (the graph of $f(\mathbf{x})$ lies above the tangent plane at $\mathbf{p}$.)*
- **b.** *f is concave iff $f(\mathbf{x}) \leq f(\mathbf{p}) + Df(\mathbf{p})(\mathbf{x}-\mathbf{p})$ for all $\mathbf{x}, \mathbf{p} \in \mathscr{D}$ (the graph of $f(\mathbf{x})$ lies below the tangent plane at $\mathbf{p}$.)*

Proof. We consider part (a) only since (b) is similar. ($\Rightarrow$) Assume that f is convex, $\mathbf{x}, \mathbf{p} \in \mathscr{D}$, and $\mathbf{x}_t = \mathbf{p} + t\,(\mathbf{x}-\mathbf{p}) = (1-t)\mathbf{p} + t\mathbf{x}$. Then, $f\,(\mathbf{x}_t) \leq (1-t)f(\mathbf{p}) + tf(\mathbf{x})$, so

$$\begin{aligned} Df(\mathbf{p})(\mathbf{x}-\mathbf{p}) &= \lim_{t\to 0+} \frac{f\,(\mathbf{x}_t) - f(\mathbf{p})}{t} \\ &\leq \lim_{t\to 0+} \frac{(1-t)\,f(\mathbf{p}) + t\,f(\mathbf{x}) - f(\mathbf{p})}{t} \\ &= \lim_{t\to 0+} \frac{t\,[f(\mathbf{x}) - f(\mathbf{p})]}{t} \\ &= f(\mathbf{x}) - f(\mathbf{p}). \end{aligned}$$

($\Leftarrow$) Assume that f satisfies the first derivative condition of the theorem. Let $\mathbf{x}_t = (1-t)\,\mathbf{p} + t\,\mathbf{x}$ and $\mathbf{w}_t = \mathbf{x} - \mathbf{x}_t = (1-t)(\mathbf{x}-\mathbf{p})$, so $\mathbf{p} - \mathbf{x}_t = -\,(t/1-t)\,\mathbf{w}_t$. Then,

$$\begin{aligned} &f(\mathbf{p}) - f(\mathbf{x}_t) \geq Df(\mathbf{x}_t)(\mathbf{p}-\mathbf{x}_t) = -\,(t/1-t)\,Df(\mathbf{x}_t)\mathbf{w}_t \qquad \text{and} \\ &f(\mathbf{x}) - f(\mathbf{x}_t) \geq Df(\mathbf{x}_t)(\mathbf{x}-\mathbf{x}_t) = Df(\mathbf{x}_t)\mathbf{w}_t. \end{aligned}$$

Multiplying the first inequality by $(1-t)$, the second by t, and adding these two inequalities together, we get

$$(1-t)\,f(\mathbf{p}) + t\,f(\mathbf{x}) - f(\mathbf{x}_t) \geq 0, \qquad \text{or}$$
$$(1-t)\,f(\mathbf{p}) + t\,f(\mathbf{x}) \geq f(\mathbf{x}_t).$$

This proves that the function is convex. □

The first derivative condition of the previous theorem inspires the following definition.

Definition. A function f is *pseudo-concave* on a set $\mathscr{D}$ provided that if $f(\mathbf{x}) > f(\mathbf{p})$ for $\mathbf{x}, \mathbf{p} \in \mathscr{D}$, then $Df(\mathbf{p})(\mathbf{x}-\mathbf{p}) > 0$. A direct check shows that a C^1 rescaling of a concave function is pseudo-concave.

The following generalizes the condition of being a critical point for a point that is on the boundary.

Theorem 3.33. *Assume that $f : \mathscr{F} \subset \mathbb{R}^n \to \mathbb{R}$ is concave or pseudo-concave and $\mathbf{x}^* \in \mathscr{F}$. Then, $\mathbf{x}^*$ maximizes f on $\mathscr{F}$ iff $Df(\mathbf{x}^*)\mathbf{v} \leq 0$ for all vectors $\mathbf{v}$ that point into $\mathscr{F}$ at $\mathbf{x}^*$.*

In particular for $\mathbf{x}^* \in \operatorname{int}(\mathscr{F})$, f attains a maximum at $\mathbf{x}^*$ on $\mathscr{F}$ iff $\mathbf{x}^*$ is a critical point of f.

Proof. Since the directional derivatives exist for a concave function, we do not need to assume that the function is C^1.

($\Rightarrow$) Assume $\mathbf{x}^*$ maximizes f over $\mathscr{F}$. Take a vector $\mathbf{v}$ that points into $\mathscr{F}$ at $\mathbf{x}^*$. Then for small $t \geq 0$, $\mathbf{x}^* + t\mathbf{v} \in \mathscr{F}$ and $f(\mathbf{x}^* + t\mathbf{v}) \leq f(\mathbf{x}^*)$. Taking the derivative with respect to t, we get that

$$Df(\mathbf{x}^*)\mathbf{v} = \lim_{t\to 0+} \frac{f(\mathbf{x}^* + t\mathbf{v}) - f(\mathbf{x}^*)}{t} \leq 0.$$

This proves the desired inequality on these directional derivatives.

($\Leftarrow$) Now, assume that $Df(\mathbf{x}^*)\mathbf{v} \leq 0$ for all vectors $\mathbf{v}$ that point into $\mathscr{F}$ at $\mathbf{x}^*$. If f does not have a maximum at $\mathbf{x}^*$, there exists a point $\mathbf{z} \in \mathscr{F}$ such that $f(\mathbf{z}) > f(\mathbf{x}^*)$. Let $\mathbf{v} = \mathbf{z} - \mathbf{x}^*$. If f is pseudo-concave, this leads to the contradiction that $Df(\mathbf{x}^*)\mathbf{v} = Df(\mathbf{x}^*)(\mathbf{z}-\mathbf{x}^*) > 0$. If f is concave, then $\mathbf{v}$ points into $\mathscr{F}$ at $\mathbf{x}^*$. Also, for $\mathbf{x}_t = \mathbf{x}^* + t\,\mathbf{v} = (1-t)\,\mathbf{x}^* + t\,\mathbf{z}$,

$$f(\mathbf{x}_t) \geq (1-t)\,f(\mathbf{x}^*) + t\,f(\mathbf{z}) = f(\mathbf{x}^*) + t\,[f(\mathbf{z}) - f(\mathbf{x}^*)] \qquad \text{so}$$
$$Df(\mathbf{x}^*)\mathbf{v} = \lim_{t\to 0+} \frac{f(\mathbf{x}_t) - f(\mathbf{x}^*)}{t} \geq \lim_{t\to 0+} \frac{t\,[f(\mathbf{z}) - f(\mathbf{x}^*)]}{t} = f(\mathbf{z}) - f(\mathbf{x}^*) > 0.$$

This contradicts the inequality on the directional derivatives and shows that f must have a maximum at $\mathbf{x}^*$. □

3.4.5. Proof of Karush-Kuhn-Tucker Theorem

Proof of Theorem 3.19.a.i. Assume that f is concave and satisfies conditions KKT-1,2 at $(\boldsymbol{\lambda}^*, \mathbf{x}^*)$ with all the $\lambda_i^* \geq 0$ and $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ is convex. Therefore, f restricted to $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ is concave. Let $\mathbf{E}$ be the set of indices of effective constraints at $\mathbf{x}^*$. Suppose $\mathbf{v}$ is a vector that points into $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ at $\mathbf{x}^*$. If $i \notin \mathbf{E}$, then $g_i(\mathbf{x}^*) < b_i$, $\lambda_i^* = 0$, and $\lambda_i^* Dg_i(\mathbf{x}^*)\mathbf{v} = 0$. If $i \in \mathbf{E}$, then $g_i(\mathbf{x}^* + t\mathbf{v}) \leq b_i = g_i(\mathbf{x}^*)$ for $t > 0$,

$$\frac{g_i(\mathbf{x}^* + t\mathbf{v}) - g_i(\mathbf{x}^*)}{t} \leq 0 \qquad \text{for } t > 0, \text{ and}$$
$$Dg_i(\mathbf{x}^*)\mathbf{v} = \lim_{t\to 0+} \frac{g_i(\mathbf{x}^* + t\mathbf{v}) - g_i(\mathbf{x}^*)}{t} \leq 0.$$

Since all the $\lambda_i^* \geq 0$, $\sum_i \lambda_i^* Dg_i(\mathbf{x}^*)\mathbf{v} \leq 0$. By the first order condition KKT-1

$$Df(\mathbf{x}^*)\mathbf{v} = \sum_i \lambda_i^* \, Dg_i(\mathbf{x}^*)\mathbf{v} \leq 0.$$

Since the directional derivative is negative for any vector pointing into $\mathscr{F}_{\mathbf{g},\mathbf{b}}$, f has a maximum at $\mathbf{x}^*$ by Theorem 3.33. □

Proof of 3.19.a.ii. Assume that f is concave and has a maximum on $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ at $\mathbf{x}^*$. Let the Lagrangian be $L(\boldsymbol{\lambda}, \mathbf{x}) = f(\mathbf{x}) + \sum_i \lambda_i(b_i - g_i(\mathbf{x}))$ as usual. We use two disjoint convex sets to show that for correctly chosen fixed $\boldsymbol{\lambda}^* \geq 0$ the Lagrangian has an interior maximum at $\mathbf{x}^*$.

The set $\mathscr{Y} = \{\, (w, \mathbf{z}) \in \mathbb{R} \times \mathbb{R}^m : w > f(\mathbf{x}^*) \;\&\; \mathbf{z} \gg \mathbf{0} \,\}$ is convex for the constant $f(\mathbf{x}^*)$.

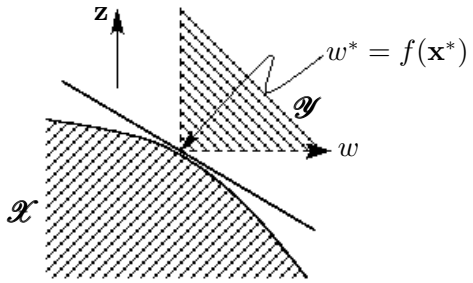


Figure 3.4.6. Separation of $\mathscr{Y}$ and $\mathscr{X}$ by a hyperplane

We claim that

$$\mathscr{X} = \{\, (w, \mathbf{z}) \in \mathbb{R} \times \mathbb{R}^m : w \leq f(\mathbf{x}) \;\;\&\;\; \mathbf{z} \leq \mathbf{b} - \mathbf{g}(\mathbf{x}) \text{ for some } \mathbf{x} \in \mathbb{R}^n \,\}$$

is also convex. Note that $\mathscr{X}$ is defined using $\mathbb{R}^n$ and not $\mathscr{F}_{\mathbf{g},\mathbf{b}}$. Let $(w_0, \mathbf{z}_0), (w_1, \mathbf{z}_1) \in \mathscr{X}$ have corresponding points $\mathbf{x}_0, \mathbf{x}_1 \in \mathbb{R}^n$. Set $w_t = (1-t)w_0 + tw_1$, $\mathbf{z}_t = (1-t)\mathbf{z}_0 + t\mathbf{z}_1$, and $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$. Then,

$$f(\mathbf{x}_t) \geq (1-t)\, f(\mathbf{x}_0) + t\, f(\mathbf{x}_1) \geq (1-t)w_0 + tw_1 = w_t$$
$$g(\mathbf{x}_t) \leq (1-t)\, g(\mathbf{x}_0) + t\, g(\mathbf{x}_1) \leq (1-t)(\mathbf{b} - \mathbf{z}_0) + t(\mathbf{b} - \mathbf{z}_1) = \mathbf{b} - \mathbf{z}_t.$$

This shows that $(w_t, \mathbf{z}_t) \in \mathscr{X}$ and $\mathscr{X}$ is convex.

We next claim that $\mathscr{X} \cap \mathscr{Y} = \emptyset$. Assume that there exists a $(w, \mathbf{z}) \in \mathscr{X} \cap \mathscr{Y}$. Because the point is in $\mathscr{Y}$, $\mathbf{z} \gg \mathbf{0}$ and $w > f(\mathbf{x}^*)$. Because the pair is in $\mathscr{X}$, there exists $\mathbf{x} \in \mathbb{R}^n$ with $w \leq f(\mathbf{x})$ and $\mathbf{z} \leq \mathbf{b} - \mathbf{g}(\mathbf{x})$. The n $\mathbf{b} - \mathbf{g}(\mathbf{x}) \geq \mathbf{z} \gg \mathbf{0}$, so $\mathbf{x} \in \mathscr{F}_{\mathbf{g},\mathbf{b}}$. Combining, $f(\mathbf{x}) \geq w > f(\mathbf{x}^*)$ for $\mathbf{x} \in \mathscr{F}_{\mathbf{g},\mathbf{b}}$, which contradicts the fact that f has a maximum at $\mathbf{x}^*$ on $\mathscr{F}_{\mathbf{g},\mathbf{b}}$. This contradiction shows that they have empty intersection.

By the separation theorem for disjoint convex sets (Theorem 1.68 in [**14**]), there exist a nonzero vector $(p, \mathbf{q}) \in \mathbb{R} \times \mathbb{R}^m$ such that

$$p\, w + \mathbf{q} \cdot \mathbf{z} \leq p\, u + \mathbf{q} \cdot \mathbf{v} \qquad \text{for all } (w, \mathbf{z}) \in \mathscr{X},\ (u, \mathbf{v}) \in \mathscr{Y}. \tag{7}$$

We claim that $(p, \mathbf{q}) \geq 0$. Assume one of the components is negative. Fix $(w, \mathbf{z}) \in \mathscr{X}$ with corresponding point $\mathbf{x}$. By taking the corresponding coordinate of $(u, \mathbf{v}) \in \mathscr{Y}$ large and positive, $p\, u + \mathbf{q} \cdot \mathbf{v}$ can be made arbitrarily negative, contradicting the separation inequality (7). Thus, $(p, \mathbf{q}) \geq 0$.

Taking any $\mathbf{x} \in \mathbb{R}^n$, setting $w = f(\mathbf{x})$ and $\mathbf{z} = \mathbf{b} - \mathbf{g}(\mathbf{x})$, and letting $(u, \mathbf{v}) \in \mathscr{Y}$ converge to $(f(\mathbf{x}^*), \mathbf{0})$, we get that

$$p\, f(\mathbf{x}) + \mathbf{q} \cdot (\mathbf{b} - \mathbf{g}(\mathbf{x})) \leq p\, f(\mathbf{x}^*) \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n. \tag{8}$$

We want to show $p \neq 0$. If $p = 0$, then inequality (8) yields $\mathbf{q} \cdot (\mathbf{b} - \mathbf{g}(\mathbf{x})) \leq 0$ for all $\mathbf{x} \in \mathbb{R}^n$. Taking $\mathbf{x} = \overline{\mathbf{x}}$ given by the Slater condition (with $\mathbf{g}(\overline{\mathbf{x}}) \ll \mathbf{b}$), we get that $\mathbf{q} \leq \mathbf{0}$ and so $\mathbf{q} = \mathbf{0}$. Thus, if $p = 0$ then $\mathbf{q} = \mathbf{0}$, which contradicts the fact that $(p, \mathbf{q})$ is not identically zero. Thus, we have shown that $p \neq 0$. Setting $\boldsymbol{\lambda}^* = (1/p)\,\mathbf{q} = (q_1/p, \ldots, q_m/p) \geq 0$ gives $\lambda_i^* \geq 0$ for $1 \leq i \leq m$, or condition KKT-3.

Inequality (8) and the definition of $\boldsymbol{\lambda}^*$ show that

$$L(\boldsymbol{\lambda}^*, \mathbf{x}) = f(\mathbf{x}) + \sum_i \lambda_i^* \,(b_i - g_i(\mathbf{x})) \leq f(\mathbf{x}^*) \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n. \tag{9}$$

For $\mathbf{x} = \mathbf{x}^*$, this gives that $\sum_i \lambda_i^* \,(b_i - g_i(\mathbf{x}^*)) \leq 0$. But $\lambda_i^* \geq 0$ and $b_i - g_i(\mathbf{x}^*) \geq 0$, so

$$\lambda_i^* \,(b_i - g_i(\mathbf{x}^*)) = 0 \qquad \text{for } 1 \leq i \leq m, \text{ or condition KKT-2.}$$

Also $L(\boldsymbol{\lambda}^*, \mathbf{x}^*) = f(\mathbf{x}^*)$, so substituting into (9),

$$L(\boldsymbol{\lambda}^*, \mathbf{x}) \leq L(\boldsymbol{\lambda}^*, \mathbf{x}^*) \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n.$$

Thus with $\boldsymbol{\lambda}^*$ fixed, $L(\boldsymbol{\lambda}^*, \mathbf{x})$ has an interior maximum at $\mathbf{x}^*$ and

$$0 = D_{\mathbf{x}} L(\boldsymbol{\lambda}^*, \mathbf{x}^*) = Df(\mathbf{x}^*) - \sum_i \lambda_i^* \, Dg_i(\mathbf{x}^*),$$

or condition KKT-1. □

3.4. Exercises

3.4.1. Which of the following functions are convex, concave, or neither? Why?

- **a.** $f(x, y) = 2x^2 - 4xy - 7x + 5y$.
- **b.** $f(x, y) = xe^{-x-5y}$.
- **c.** $f(x, y, z) = -x^2 + 2xy - 3y^2 + 9x - 7y$.
- **d.** $f(x, y, z) = 2x^2 + y^2 + 2z^2 + xy - 3xz$.
- **e.** $f(x, y, z) = -2x^2 - 3y^2 - 2z^2 + 2xy + 3xz + yz$.

3.4.2. Let $f : \mathbb{R}^n_{++} \to \mathbb{R}$ be defined by $f(x_1, \ldots, x_n) = \ln(x_1^{\alpha_1} \cdots x_n^{\alpha_n})$, where all the $\alpha_i > 0$. Is f convex or concave?

3.4.3. Let $\mathscr{D}$ be a convex set and $h : \mathscr{D} \to \mathbb{R}$ a concave function.

- **a.** Show that
$$\mathscr{F} = \{\, (x, y) \in \mathscr{D} : 0 \leq x,\ \ 0 \leq y,\ \ 0 \leq h(x, y) \,\}$$
is convex.
- **b.** Assume $f : \mathscr{D} \to \mathbb{R}$ is convex and $f(\mathbf{y}) < f(\mathbf{x})$ for $\mathbf{x}, \mathbf{y} \in \mathscr{D}$. Let $\mathbf{x}_t = (1 - t)\mathbf{x} + t\mathbf{y}$.
Show that $f(\mathbf{x}_t) < f(\mathbf{x})$ for $0 < t \leq 1$.

3.4.4. Consider the following problem:

$$\begin{aligned} \text{Minimize:} \quad & f(x_1, x_2, x_3) = 2\,x_1^2 + 5\,x_2^2 + 3\,x_3^2 - 2\,x_1 x_2 - 2\,x_2 x_3 \\ \text{Subject to:} \quad & 25 \leq 4\,x_1 + 6\,x_2 + x_3 \\ & x_i \geq 0 \ \text{ for } i = 1, 2, 3. \end{aligned}$$

- **a.** What are the KKT-1,2,3′ equations for this problem to have a minimum? Be sure an list all the equations that must be satisfied. Then, solve these equations for a solution (x_1^*, x_2^*, x_3^*).
- **b.** Explain why the objective function and constraints satisfy the assumptions for a minimum of the Karush-Kuhn-Tucker Theorem under Convexity.
Note that the function f is a positive definite quadratic function.
- **c.** Explain why (x_1^*, x_2^*, x_3^*) must be a minimizer, including explaining how the theorems apply and what other conditions need to be satisfied.

3.4.5. This is a problem about maximization of the social welfare function

$$W(x_1, x_2, x_3) = a_1 \ln(x_1) + a_2 \ln(x_2) + a_3 \ln(x_3).$$

for the production of three outputs x_1, x_2, and x_3, where $a_1, a_2, a_3 > 0$. There are 600 units of labor available and 550 units of land. Because of the requirements for the production of each product, we have the constraints

$$\begin{aligned} 2x_1 + x_2 + 3x_3 &\leq 600, \\ x_1 + 2x_2 + x_3 &\leq 550, \\ 1 \leq x_1,\ \ 1 \leq &x_2,\ \ 1 \leq x_3. \end{aligned}$$

(Notice that for $0 < x_i < 1$, $a_i \ln(x_i) < 0$ would contribute a negative amount to the social welfare.)

- **a.** Write down the KKT-1,2,3 equations that must be solved to find a maximum of W. Find the solution of these equations (x_1^*, x_2^*, x_3^*).
- **b.** Explain why the objective function and constraints satisfy the assumptions for a maximum of the Karush-Kuhn-Tucker Theorem under Convexity.
- **c.** Explain why (x_1^*, x_2^*, x_3^*) must be a maximizer, including explaining how the theorems apply and what other conditions need to be satisfied.

3.4.6. Consider the following problem:

$$\begin{aligned} \text{Maximize :}\quad & f(x, y) = -(x-9)^2 - (y-8)^2, \\ \text{Subject to :}\quad & 4y - x^2 \geq 0 \\ & x + y \leq 24 \\ & x \geq 0 \\ & y \geq 0. \end{aligned}$$

- **a.** Write out the KKT-1,2,3 equations for this maximization problem and find a solution (x^*, y^*).
- **b.** Why is the objective f concave and the constraints convex?
- **c.** Why must the point (x^*, y^*) be a global maximizer on the feasible set $\mathscr{F}$?
- **d.** Draw a rough sketch of the feasible set $\mathscr{F}$ of points satisfying the constraint equations. Why does it satisfy the Slater conditions?

3.4.7. Consider the problem

$$\begin{aligned} \text{Maximization:}\quad & f(x, y, z) = xyz \\ \text{Subject to:}\quad & 2x + y + 2z \leq 5, \\ & x + 2y + z \leq 4, \\ & x + y + z \leq 3, \\ & 0 \leq x,\ 0 \leq y,\ 0 \leq z. \end{aligned}$$

- **a.** Write down the KKT-1,2,3 equations for a maximum on the feasible set. Find a solution $\mathbf{p}^*$ and $\boldsymbol{\lambda}^*$ to these equations.
- **b.** Why are all the constraints convex on $\mathbb{R}^3_+$? Why is f a rescaled concave function on $\mathbb{R}^3_+$?
- **c.** Why must $\mathbf{p}^*$ be a maximizer of f on the feasible set?
- **d.** Show that the feasible set satisfies the Slater condition.

3.4.8. Consider the problem

$$\begin{aligned} \text{Minimize:}\quad & f(x, y) = x^4 + 12\,x^2 + y^4 + 6\,y^2 - xy - x - y \\ \text{Subject to:}\quad & x + y \geq 6, \\ & x - y \geq 3, \\ & 0 \leq x,\ 0 \leq y. \end{aligned}$$

a. Write down the KKT-1,2,3′ equations for a minimum on the feasible set.
b. Find a solution to these equations where both constraints are tight.
c. Why must the solution found in part (b) be a minimizer of f on the feasible set?

3.4.9. A firm produces a single output q with two inputs x and y, with production function $q = xy$. The output must be at least q_0 units, $xy \geq q_0 > 0$. The firm is obligated to use at least one unit of x, $x \geq 1$. The prices of x and y are p and 1 respectively. Assume that the firm wants to minimize the cost of the inputs $f(x, y) = px + y$.

a. Is the feasible set closed? Compact? Convex?
b. Write down the KKT-1,2,3′ equation for a minimum.
c. Find the minimizer by solving the KKT-1,2,3′ equations.
Hints: (i) Note that one of the equations for KKT-1 implies that the multiplier for $0 \geq q_0 - xy$ is nonzero and so this constraint must be effective at a solution.
(ii) If $0 \geq 1 - x$ is tight, then $q \leq p$ because both multiplier must be less than or equal to zero.
(iii) If the multiplier for $0 \geq 1 - x$ is zero, then $q \geq p$ because $x \geq 1$.

3.4.10. Consider the problem

$$\begin{aligned} \text{Minimize:} \quad & \sum_{j=1}^{n} \frac{c_j}{x_j}, \\ \text{Subject to:} \quad & \textstyle\sum_{j=1}^{n} a_j x_j = b, \\ & 0 \leq x_j \qquad \text{for } j = 1, \ldots, n, \end{aligned}$$

where a_j, b, and c_j are all positive constants.

a. Write down the KKT-1,2,3′ equations for a minimum on the feasible set.
Hint: The equality constraint can be written as two inequality constraints, $\sum_{j=1}^{n} a_j x_j \leq b$ and $\sum_{j=1}^{n} -a_j x_j \leq -b$. Also, $\mathbf{x}^* \gg \mathbf{0}$.
b. Find a solution $\mathbf{x}^*$ and $\boldsymbol{\lambda}^*$ to these equations.
c. Why must $\mathbf{p}^*$ be a minimizer of f on the feasible set?

3.4.11. Let $T > 1$ be a fixed integer and $0 < \delta < 1$. Consider the following maximization problem.

$$\begin{aligned} \text{Maximize:} \quad & \sum_{j=1}^{\mathsf{T}} \delta^j x_j^{\frac{1}{2}}, \\ \text{Subject to:} \quad & \textstyle\sum_{j=1}^{n} x_j \leq 1, \\ & x_j \geq 0 \text{ for } j = 1, \ldots T. \end{aligned}$$

a. Write down the KKT-1,2,3 equations.
b. Consider the case when all the $x_j > 0$. Solve the KKT-1,2,3 equations for a solution. *Hint:* Why must the multiplier be nonzero, so the constraint tight? How must x_j be related to x_1? Using the tight constraint, what must x_1 equal?
c. Why must the solution found in part (b) be a maximizer?

3.4.12. Let $I, p_i > 0$ for $1 \leq i \leq n$. Show that

$$\mathscr{B}(\mathbf{p}, I) = \{\mathbf{x} \in \mathbb{R}^n_+ : p_1 x_1 + \cdots + p_n x_n \leq I\}$$

satisfies Slater's condition. *Hint:* Split up $\frac{1}{2}I$ evenly among the amounts spent on the various commodities, i.e., $\frac{1}{2n}I$ on each.

3.4.13. Assume that $g_i : \mathscr{D} \subset \mathbb{R}^n \to \mathbb{R}$ for $1 \leq i \leq k$ are convex and bounded on $\mathscr{D}$.
- **a.** Show that $f(\mathbf{x}) = \max_{1\leq i\leq k} g_i(\mathbf{x})$ is a convex function.
  Hint: $\max_i\{\, a_i + b_i \,\} \leq \max_i\{\, a_i \,\} + \max_i\{\, b_i \,\}$.
- **b.** Is $g(x) = \min_{1\leq i\leq k} g_i(\mathbf{x})$ convex? Why or why not?
  Hint: $\min\{\, a_i + b_i \,\} \geq \min\{\, a_i \,\} + \min\{\, b_i \,\}$.
- **c.** If the g_i are concave, is $\min_{1\leq i\leq k} g_i(\mathbf{x})$ concave?

3.4.14. Assume that $f, g : \mathbb{R}^n \to \mathbb{R}$ are C^1, f is concave, and g is convex. Let $\mathscr{F}_{g,b} = \{\, \mathbf{x} \in \mathbb{R}^n_+ : g(\mathbf{x}) \leq b \,\}$. Assume that $\mathscr{F}_{g,b}$ satisfies Slater condition. Further assume that f attains a maximum on $\mathscr{F}_{g,b}$ at $\mathbf{p}^*$ with $\mathbf{p}^* \in \mathbb{R}^n_{++}$ and $Df(\mathbf{p}^*) \neq \mathbf{0}$.
- **a.** Explain why $g(\mathbf{p}^*)$ must equal b.
- **b.** Explain why $\mathbf{p}^*$ is a minimizer of $g(\mathbf{y})$ on $\{\, \mathbf{y} \in \mathbb{R}^n_+ : f(\mathbf{y}) \geq f(\mathbf{p}^*) \,\}$.

3.4.15. Give an example of a concave function $f : \mathbb{R}_+ \to \mathbb{R}$ that is bounded above but does not attain a maximum. Check whether the constraint qualification is satisfied at $\mathbf{p}^*$.

3.4.16. Let $f(x, y) = (xy)^{1/3}$ and $\mathbf{D} = \{(x, y) \in \mathbb{R}^2_+ : x + y \leq 2\}$. You may use the fact that the maximal value of f on $\mathbf{D}$ is 1. Let

$$\mathscr{X} = \{\, (w, z_1, z_2, z_3) \in \mathbb{R}^4 : w \leq f(x, y),\ z_1 \leq x,\ z_2 \leq y,\ z_3 \leq 2 - x - y \text{ for some } (x, y) \in \mathbb{R}^2_+ \,\},$$

$$\mathscr{Y} = \{\, (w, z_1, z_2, z_3) \in \mathbb{R}^4 : w > 1,\ z_1 > 0,\ z_2 > 0,\ z_3 > 0 \,\}.$$

- **a.** Show that f is concave on $\mathbb{R}^2_{++}$, continuous on $\mathbb{R}^2_+$, and so concave on $\mathbb{R}^2_+$.
- **b.** Show that $\mathscr{X}$ and $\mathscr{Y}$ are convex.
- **c.** Show that $\mathscr{X} \cap \mathscr{Y} = \emptyset$.
- **d.** Let $(p, q_1, q_2, q_3) = (1, 0, 0, 1/3)$. Show that

  $$(1, 0, 0, 1/3) \cdot (w, z_1, z_2, z_3) \leq (1, 0, 0, 1/3) \cdot (u, v_1, v_2, v_3),$$

  for all $(w, z_1, z_2, z_3) \in \mathscr{X}$ and $(u, v_1, v_2, v_3) \in \mathscr{Y}$.
- **e.** Conclude that $f(x, y) + (2 - x - y)/3 \leq 1$ for all $(x, y) \in \mathbb{R}^2_+$.

3.5. Second-Order Conditions for Extrema of Constrained Functions

In this section we derive a second derivative test for local extrema with equality constraints. This material is optional and is mainly provided as a reference.

Lemma 3.34. *Assume that $\mathbf{x}^*$ and $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_k^*)$ satisfy the first-order Lagrange multiplier conditions. If $\mathbf{r}(t)$ is a curve in the level set $\mathbf{g}^{-1}(\mathbf{b})$ with $\mathbf{r}(0) = \mathbf{x}^*$ and $\mathbf{r}'(0) = \mathbf{v}$, then the second derivative of the composition $f \circ \mathbf{r}(t)$ is given by the following formula:*

$$\frac{d^2}{dt^2} f(\mathbf{r}(t))\bigg|_{t=0} = \mathbf{v}^{\mathsf{T}} \left[D^2 f(\mathbf{x}^*) - \sum_{\ell=1}^{k} \lambda_\ell^* D^2 g_\ell(\mathbf{x}^*) \right] \mathbf{v} = \mathbf{v}^{\mathsf{T}} D^2_{\mathbf{x}} L(\boldsymbol{\lambda}^*, \mathbf{x}^*)\, \mathbf{v}.$$

Proof. Using the chain rule and product rule,

$$\frac{d}{dt} f(\mathbf{r}(t)) = Df(\mathbf{r}(t))\, \mathbf{r}'(t) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{r}(t))\, r_i'(t)$$

(by the chain rule)

and

$$\begin{aligned}\left.\frac{d^2}{dt^2} f(\mathbf{r}(t))\right|_{t=0} &= \sum_{i=1}^{n} \left.\frac{d}{dt}\frac{\partial f}{\partial x_i}(\mathbf{r}(t))\right|_{t=0} r_i'(0) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{r}(0))\, r_i''(0) \\ &\qquad \text{(by the product rule)} \\ &= \sum_{\substack{i=1,\dots,n \\ j=1,\dots,n}} \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}^*)\, r_i'(0)\, r_j'(0) + Df(\mathbf{x}^*)\, \mathbf{r}''(0) \\ &\qquad \text{(by the chain rule)} \\ &= (\mathbf{r}'(0))^{\mathsf{T}} D^2 f(\mathbf{x}^*)\, \mathbf{r}'(0) + \sum_{\ell=1}^{k} \lambda_\ell^* D(g_\ell)(\mathbf{x}^*)\, \mathbf{r}''(0).\end{aligned}$$

In the last equality, we used the fact that $Df(\mathbf{x}^*) = \sum_{\ell=1}^{k} \lambda_\ell^* D(g_\ell)(\mathbf{x}^*)$ and the definition of $D^2 f$.

We can perform a similar calculation for the constraint equation $b_\ell = g_\ell(\mathbf{r}(t))$ whose derivatives are zero:

$$\begin{aligned}0 &= \frac{d}{dt} g_\ell(\mathbf{r}(t)) = \sum_{i=1,\dots,n} \left(\frac{\partial g_\ell}{\partial x_i}(\mathbf{r}(t))\right) \mathbf{r}_i'(t), \\ 0 &= \left.\frac{d^2}{dt^2} g_\ell(\mathbf{r}(t))\right|_{t=0} = \sum_{i=1,\dots,n} \left.\frac{d}{dt}\left(\frac{\partial g_\ell}{\partial x_i}(\mathbf{r}(t))\right) \mathbf{r}_i'(t)\right|_{t=0} \\ &= \sum_{\substack{i=1,\dots,n \\ j=1,\dots,n}} \left(\frac{\partial^2 g_\ell}{\partial x_j \partial x_i}(\mathbf{x}^*)\right) r_i'(0) r_j'(0) + D(g_\ell)(\mathbf{x}^*)\mathbf{r}''(0), \qquad \text{so} \\ \lambda_\ell^* D(g_\ell)(\mathbf{x}^*)\mathbf{r}''(0) &= -\lambda_\ell^* \left(\mathbf{r}'(0)\right)^{\mathsf{T}} D^2(g_\ell)(\mathbf{x}^*)\mathbf{r}'(0).\end{aligned}$$

Substituting this equality into the expression for the second derivative of $f(\mathbf{r}(t))$,

$$\left.\frac{d^2}{dt^2} f(\mathbf{r}(t))\right|_{t=0} = \mathbf{v}^{\mathsf{T}} \left[D^2 f(\mathbf{x}^*) - \sum_{\ell=1}^{k} \lambda_\ell^* D^2 g_\ell(\mathbf{x}^*) \right] \mathbf{v},$$

where $\mathbf{v} = \mathbf{r}'(0)$. This is what is claimed. □

The next theorem uses the above lemma to derive conditions for local maxima and minima in terms of the second derivative of the Lagrangian on the null space $\operatorname{null}(Dg(\mathbf{x}^*))$.

Theorem 3.35. *Assume $f, g_i : \mathbb{R}^n \to \mathbb{R}$ are C^2 for $1 \le i \le k$. Assume that $\mathbf{x}^* \in \mathbb{R}^n$ and $\boldsymbol{\lambda}^* = (\lambda_1^*, \dots, \lambda_k^*)$ satisfy the first-order conditions of the Theorem of Lagrange with $\operatorname{rank}(D\mathbf{g}(\mathbf{x}^*)) = k$. Set $D_{\mathbf{x}}^2 L^* = D_{\mathbf{x}}^2 L(\boldsymbol{\lambda}^*, \mathbf{x}^*) = D^2 f(\mathbf{x}^*) - \sum_{\ell=1}^{k} \lambda_\ell^* D^2 g_\ell(\mathbf{x}^*)$.*

- **a.** *If f has a local maximum on $\mathbf{g}^{-1}(\mathbf{b})$ at $\mathbf{x}^*$, then $\mathbf{v}^{\mathsf{T}} D_{\mathbf{x}}^2 L^* \mathbf{v} \le 0$ for all $\mathbf{v} \in \operatorname{null}(D\mathbf{g}(\mathbf{x}^*))$.*
- **b.** *If f has a local minimum on $\mathbf{g}^{-1}(\mathbf{b})$ at $\mathbf{x}^*$, then $\mathbf{v}^{\mathsf{T}} D_{\mathbf{x}}^2 L^* \mathbf{v} \ge 0$ for all $\mathbf{v} \in \operatorname{null}(D\mathbf{g}(\mathbf{x}^*))$.*
- **c.** *If $\mathbf{v}^{\mathsf{T}} D_{\mathbf{x}}^2 L^* \mathbf{v} < 0$ for all $\mathbf{v} \in \operatorname{null}(D\mathbf{g}(\mathbf{x}^*)) \setminus \{\mathbf{0}\}$, then f has a strict local maximum on $\mathbf{g}^{-1}(\mathbf{b})$ at $\mathbf{x}^*$.*
- **d.** *If $\mathbf{v}^{\mathsf{T}} D_{\mathbf{x}}^2 L^* \mathbf{v} > 0$ for all $\mathbf{v} \in \operatorname{null}(D\mathbf{g}(\mathbf{x}^*)) \setminus \{\mathbf{0}\}$, then f has a strict local minimum on $\mathbf{g}^{-1}(\mathbf{b})$ at $\mathbf{x}^*$.*
- **e.** *If $\mathbf{v}^{\mathsf{T}} D_{\mathbf{x}}^2 L^* \mathbf{v}$ is positive for some vector $\mathbf{v} \in \operatorname{null}(D\mathbf{g}(\mathbf{x}^*))$ and negative for another such vector, then f has is neither a local maximum nor a local minimum of f on $\mathbf{g}^{-1}(\mathbf{b})$ at $\mathbf{x}^*$.*

Proof. (**b**) We consider the case of minima. (The case of maximum just reverses the direction of the inequality.) Lemma 3.34 shows that

$$\frac{d^2}{dt^2} f(\mathbf{r}(t))\bigg|_{t=0} = \mathbf{v}^\mathsf{T} D^2_{\mathbf{x}} L^* \mathbf{v},$$

where $\mathbf{v} = \mathbf{r}'(0)$. If $\mathbf{x}^*$ is a local minimum on $g^{-1}(\mathbf{b})$ then

$$\frac{d^2}{dt^2} f(\mathbf{r}(t))\bigg|_{t=0} \geq 0$$

for any curves $\mathbf{r}(t)$ in $\mathbf{g}^{-1}(\mathbf{b})$ with $\mathbf{r}(0) = \mathbf{x}^*$. Thus, $\mathbf{v}^\mathsf{T} D^2_{\mathbf{x}} L^* \mathbf{v} \geq 0$ for any vector $\mathbf{v}$ in $\mathbf{T}_{\mathbf{g}}(\mathbf{x}^*)$. But we had by Proposition 3.8 that $\mathbf{T}_{\mathbf{g}}(\mathbf{x}^*) = \operatorname{null}(D\mathbf{g}(\mathbf{x}^*))$, so part (b) of the theorem is proved.

(**d**) If $\mathbf{v}^\mathsf{T} D^2_{\mathbf{x}} L^* \mathbf{v} > 0$ for all vectors $\mathbf{v} \neq \mathbf{0}$ in $\operatorname{null}(D\mathbf{g}(\mathbf{x}^*))$, then by Proposition 3.8 and Lemma 3.34,

$$\frac{d^2}{dt^2} f(\mathbf{r}(t))\bigg|_{t=0} = \mathbf{r}'(0)^\mathsf{T} D^2_{\mathbf{x}} L^* \, \mathbf{r}'(0) > 0$$

for any curves $\mathbf{r}(t)$ in $\mathbf{g}^{-1}(\mathbf{b})$ with $\mathbf{r}(0) = \mathbf{x}^*$ and $\mathbf{r}'(0) \neq \mathbf{0}$. This latter condition implies that $\mathbf{x}^*$ is a local minimizer on $\mathbf{g}^{-1}(\mathbf{b})$.

For part (**e**), if $\mathbf{v}^\mathsf{T} D^2_{\mathbf{x}} L^* \mathbf{v}$ is both positive and negative, then there are some curves where the value of f is greater than at $\mathbf{x}^*$ and others on which the value is less. □

The preceding theorem shows that we need to consider the quadratic form $\mathbf{x}^\mathsf{T} D^2_{\mathbf{x}} L^* \mathbf{x}$ on the null space $\operatorname{null}(D\mathbf{g}(\mathbf{x}^*))$. The next theorem shows that this restricted quadratic form can be shown to be positive or negative definite by determinants of a submatrices of $DL(\boldsymbol{\lambda}^*, \mathbf{x}^*)$.

Definition. Let $L(\boldsymbol{\lambda}, \mathbf{x}) = f(\mathbf{x}) + \sum_{1 \leq i \leq k} \lambda_i \, (b_i - g_i(\mathbf{x}))$ be the Lagrangian. The derivative of L with respect to all its variables is

$$\mathbf{H}_n = D^2_{\mathbf{x}} L(\boldsymbol{\lambda}^*, \mathbf{x}^*) = \begin{pmatrix} \mathbf{0}_k & -Dg(\mathbf{x}^*) \\ -Dg(\mathbf{x}^*)^\mathsf{T} & D^2_{\mathbf{x}} L^* \end{pmatrix},$$

which is obtained by "bordering" the $n \times n$ matrix $D^2_{\mathbf{x}} L^*$ with the $k \times n$ matrix $-Dg(\mathbf{x}^*)$.

We assume that the rank of $D\mathbf{g}(\mathbf{x}^*)$ is $k < n$, so there is a $k \times k$ submatrix with nonzero determinant. To form the correct submatrices, we need to assume that the the variables have been rearranged so the first k columns suffice and this $k \times k$ submatrix has nonzero determinant. For $1 \leq \ell \leq n$, the *bordered Hessians* $\mathbf{H}_\ell$ are $(k + \ell) \times (k + \ell)$ submatrices of $\mathbf{H}_n = DL(\boldsymbol{\lambda}^*, \mathbf{x}^*)$ given as follows:

$$\mathbf{H}_\ell = \begin{pmatrix} 0 & \dots & 0 & -\frac{\partial g_1}{\partial x_1} & \dots & -\frac{\partial g_1}{\partial x_\ell} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & -\frac{\partial g_k}{\partial x_1} & \dots & -\frac{\partial g_k}{\partial x_\ell} \\ -\frac{\partial g_1}{\partial x_1} & \dots & -\frac{\partial g_k}{\partial x_1} & \frac{\partial^2 L^*}{\partial x_1^2} & \dots & \frac{\partial^2 L^*}{\partial x_1 \partial x_\ell} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ -\frac{\partial g_1}{\partial x_\ell} & \dots & -\frac{\partial g_k}{\partial x_\ell} & \frac{\partial^2 L^*}{\partial x_\ell \partial x_1} & \dots & \frac{\partial^2 L^*}{\partial x_\ell^2} \end{pmatrix}.$$

Theorem 3.36. *Assume that* $f, g_i : \mathbb{R}^n \to \mathbb{R}$ *are* C^2 *for* $1 \leq i \leq k$ *and* $(\boldsymbol{\lambda}^*, \mathbf{x}^*)$ *is a critical point of* L *satisfying* (LM) *with* $\det\left(\dfrac{\partial g_i}{\partial x_j}(\mathbf{x}^*)\right)_{1\leq i,j\leq k} \neq 0$ *and bordered Hessians* $\mathbf{H}_\ell$.

- **a**. *The point* $\mathbf{x}^*$ *is a local minimum of* $f(\mathbf{x})$ *restricted to* $\mathbf{g}^{-1}(\mathbf{b})$ *iff* $(-1)^k \det(\mathbf{H}_\ell) > 0$ *for all* $k+1 \leq \ell \leq n$.
  (Notice that the sign given by $(-1)^k$ *depends on the rank* k *and not* ℓ*.)*
- **b**. *The point* $\mathbf{x}^*$ *is a local maximum of* $f(\mathbf{x})$ *restricted to* $\mathbf{g}^{-1}(\mathbf{b})$ *iff* $(-1)^k \det(\mathbf{H}_{k+1}) < 0$, $(-1)^k \det(\mathbf{H}_{k+2}) > 0$, *and they continue to alternate signs up to* $\det(\mathbf{H}_n)$. *These conditions can be written as* $(-1)^\ell \det(\mathbf{H}_\ell) > 0$ *for all* $k+1 \leq \ell \leq n$.

Remark. For k constraints on n variables, the constraint set is parametrized by $n-k$ variables, and the test for a local extremum requires checking the sign of the determinant of $n-k$ submatrices, $\det(\mathbf{H}_\ell)$ for $k+1 \leq \ell \leq n$. The conditions can also be given by bordering $D^2_{\mathbf{x}} L^*$ with $Dg(\mathbf{x}^*)$, instead of $-Dg(\mathbf{x}^*)$ because the determinants do not change.

Proof. This proof is based on [**7**].

Let

$$\mathbf{B}_1 = \left(-\frac{\partial g_i}{\partial x_j}\right)_{1\leq i,j\leq k} \quad \text{and} \quad \mathbf{B}_2 = \left(-\frac{\partial g_i}{\partial x_j}\right)_{1\leq i\leq k,\, k+1\leq j\leq n}.$$

Let $\mathbf{w} = (x_1, \ldots, x_k)^\mathsf{T}$ be the first k coordinates and $\mathbf{z} = (x_{k+1}, \ldots, x_n)^\mathsf{T}$ be the last $n-k$ coordinates. Then the null space can be expressed in terms of the $\mathbf{z}$ variables:

$$\begin{aligned} \mathbf{0} &= \mathbf{B}_1\mathbf{w} + \mathbf{B}_2\mathbf{z} \\ \mathbf{w} &= -\mathbf{B}_1^{-1}\mathbf{B}_2\mathbf{z} = \mathbf{J}\mathbf{z} \end{aligned}$$

where $\mathbf{J} = -\mathbf{B}_1^{-1}\mathbf{B}_2$. Partitioning $D^2_{\mathbf{x}} L^* = \mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{12}^\mathsf{T} & \mathbf{A}_{22} \end{bmatrix}$ into blocks, where $\mathbf{A}_{11}$ is $k \times k$, $\mathbf{A}_{12}$ is $k \times (n-k)$, and $\mathbf{A}_{22}$ is $(n-k) \times (n-k)$, the quadratic form on the null space has the following symmetric matrix $\mathbf{E}$:

$$\begin{aligned} \mathbf{E} = \begin{bmatrix} \mathbf{J}^\mathsf{T} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{12}^\mathsf{T} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{J} \\ \mathbf{I} \end{bmatrix} &= \begin{bmatrix} \mathbf{J}^\mathsf{T} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11}\mathbf{J} + \mathbf{A}_{12} \\ \mathbf{A}_{12}^\mathsf{T}\mathbf{J} + \mathbf{A}_{22} \end{bmatrix} \\ &= \mathbf{J}^\mathsf{T}\mathbf{A}_{11}\mathbf{J} + \mathbf{J}^\mathsf{T}\mathbf{A}_{12} + \mathbf{A}_{12}^\mathsf{T}\mathbf{J} + \mathbf{A}_{22}. \end{aligned}$$

On the other hand, we can perform a (non-orthogonal) change of basis of the $n+k$-dimensional space on which the quadratic form $\mathbf{H}_n$ is defined:

$$\begin{bmatrix} \mathbf{I}_k & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{J}^\mathsf{T} & \mathbf{I}_{n-k} \end{bmatrix} \begin{bmatrix} \mathbf{0}_k & \mathbf{B}_1 & \mathbf{B}_2 \\ \mathbf{B}_1^\mathsf{T} & \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{B}_2^\mathsf{T} & \mathbf{A}_{12}^\mathsf{T} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{I}_k & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_k & \mathbf{J} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_{n-k} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{B}_1 & \mathbf{0} \\ \mathbf{B}_1^\mathsf{T} & \mathbf{A}_{11} & \mathbf{C}_{12} \\ \mathbf{0} & \mathbf{C}_{12}^\mathsf{T} & \mathbf{E} \end{bmatrix}$$

Here the matrix $\mathbf{E}$ induces the quadratic form on the null space as we showed above. Since the determinant of the change of basis matrix is one, this change of basis preserves the determinant of $\mathbf{H}_n$, and also the determinants of $\mathbf{H}_\ell$ for $2+1 \leq \ell \leq n$.

By using k row interchanges

$$\begin{aligned}\det(\mathbf{H}_n) = \det\begin{bmatrix}\mathbf{0} & \mathbf{B}_1 & \mathbf{0}\\ \mathbf{B}_1^\mathsf{T} & \mathbf{A}_{11} & \mathbf{C}_{12}\\ \mathbf{0} & \mathbf{C}_{12}^\mathsf{T} & \mathbf{E}\end{bmatrix} &= (\text{-}1)^k \det\begin{bmatrix}\mathbf{B}_1^\mathsf{T} & \mathbf{A}_{11} & \mathbf{C}_{12}\\ \mathbf{0} & \mathbf{B}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{C}_{12}^\mathsf{T} & \mathbf{E}\end{bmatrix}\\ &= (\text{-}1)^k \det(\mathbf{B}_1^\mathsf{T}) \det\begin{bmatrix}\mathbf{B}_1 & \mathbf{0}\\ \mathbf{C}_{12}^\mathsf{T} & \mathbf{E}\end{bmatrix}\\ &= (\text{-}1)^k \det(\mathbf{B}_1)^2 \det(\mathbf{E}).\end{aligned}$$

This calculation carries over to all the $\mathbf{H}_\ell$, $(\text{-}1)^k \det(\mathbf{H}_\ell) = \det(\mathbf{B}_1)^2 \det(\mathbf{E}_{\ell-k})$. Therefore, we can use the signs of the determinants of the $\mathbf{H}_\ell$ for $k+1 \le \ell \le n$ to check the signs of the determinants of the principal submatrices $\mathbf{E}_{j-k}$ with size ranging from 1 to $n-k$.

The quadratic form $\mathbf{Q}$ for $\mathbf{A}$ is positive definite on the null space iff the quadratic form for $\mathbf{E}$ is positive definite iff

$$(\text{-}1)^k \det(\mathbf{H}_\ell) = \det(\mathbf{B}_1)^2 \det(\mathbf{E}_{\ell-k}) > 0 \qquad \text{for } k+1 \le \ell \le n.$$

For the negative definite case, the quadratic form $\mathbf{Q}$ for $\mathbf{A}$ is negative definite on the null space iff the quadratic form for $\mathbf{E}$ is negative definite iff

$$(\text{-}1)^{\ell-k} \det(\mathbf{H}_\ell) = (\text{-}1)^{\ell-2k} \det(\mathbf{B}_1)^2 \det(\mathbf{E}_{\ell-k}) > 0 \qquad \text{for } k+1 \le \ell \le n.$$

□

Example 3.37. We check the second order conditions for the critical points of Example 3.9: $f(x,y,z) = z$ on the set given by $x+y+z = 12$ and $z = x^2+y^2$.

We considered this example earlier and found the two critical points $(\lambda^*, \mu^*, x^*, y^*, z^*) = (4/5, \text{-}1/5, 2, 2, 8,)$ and $(6/5, 1/5, \text{-}3, \text{-}3, 18)$.

For this example, $n = 3$ and $k = 2$, so $n-k = 1$ determinant. The Lagrangian is $L = z + \lambda(12 - x - y - z) - \mu(z - x^2 - y^2)$. The bordered Hessian is

$$\mathbf{H}_3 = D^2 L = \begin{bmatrix}0 & 0 & \text{-}1 & \text{-}1 & \text{-}1\\ 0 & 0 & \text{-}2x & \text{-}2y & 1\\ \text{-}1 & \text{-}2x & \text{-}\mu 2 & 0 & 0\\ \text{-}1 & \text{-}2y & 0 & \text{-}\mu 2 & 0\\ \text{-}1 & 1 & 0 & 0 & 0\end{bmatrix}.$$

At $(\lambda^*, \mu^*, x^*, y^*, z^*) = (4/5, -1/5, 2, 2, 8,)$,

$$\det(\mathbf{H}_3) = \det \begin{bmatrix} 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & -4 & -4 & 1 \\ -1 & -4 & \frac{2}{5} & 0 & 0 \\ -1 & -4 & 0 & \frac{2}{5} & 0 \\ -1 & 1 & 0 & 0 & 0 \end{bmatrix} = -\det \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & -4 & -4 & 1 \\ -1 & -4 & \frac{2}{5} & 0 & 0 \\ -1 & -4 & 0 & \frac{2}{5} & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

$$= -\det \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & -4 & -4 & 1 \\ 0 & -5 & \frac{2}{5} & 0 & 0 \\ 0 & -5 & 0 & \frac{2}{5} & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} = \det \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & -5 & 0 & \frac{2}{5} & 0 \\ 0 & -5 & \frac{2}{5} & 0 & 0 \\ 0 & 0 & -4 & -4 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

$$= \det \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & -5 & 0 & \frac{2}{5} & 0 \\ 0 & 0 & \frac{2}{5} & -\frac{2}{5} & 0 \\ 0 & 0 & -4 & -4 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} = \det \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & -5 & 0 & \frac{2}{5} & 0 \\ 0 & 0 & \frac{2}{5} & -\frac{2}{5} & 0 \\ 0 & 0 & 0 & -8 & 1 \\ 0 & 0 & 0 & 2 & 1 \end{bmatrix}$$

$$= \det \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & -5 & 0 & \frac{2}{5} & 0 \\ 0 & 0 & \frac{2}{5} & -\frac{2}{5} & 0 \\ 0 & 0 & 0 & -8 & 1 \\ 0 & 0 & 0 & 0 & \frac{5}{4} \end{bmatrix} = 1(-5)\left(\frac{2}{5}\right)(-8)\left(\frac{5}{4}\right) = 20 > 0.$$

Since $(-1)^k = (-1)^2 = 1$i, $k+1 = n = 3$, and $(-1)^k \det(D^2 L^*) > 0$, $(\lambda^*, \mu^*, x^*, y^*, z^*) = (4/5, -1/5, 2, 2, 8,)$ is a local minimum.

At $(\lambda^*, \mu^*, x^*, y^*, z^*) = (6/5, 1/5, -3, -3, 18)$,

$$\det(\mathbf{H}_3) = \det \begin{bmatrix} 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 6 & 6 & 1 \\ -1 & 6 & -\frac{2}{5} & 0 & 0 \\ -1 & 6 & 0 & -\frac{2}{5} & 0 \\ -1 & 1 & 0 & 0 & 0 \end{bmatrix} = -\det \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 6 & 6 & 1 \\ -1 & 6 & -\frac{2}{5} & 0 & 0 \\ -1 & 6 & 0 & -\frac{2}{5} & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

$$= -\det \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 6 & 6 & 1 \\ 0 & 5 & -\frac{2}{5} & 0 & 0 \\ 0 & 5 & 0 & -\frac{2}{5} & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} = \det \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 5 & 0 & -\frac{2}{5} & 0 \\ 0 & 5 & -\frac{2}{5} & 0 & 0 \\ 0 & 0 & 6 & 6 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

$$= \det \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 5 & 0 & -\frac{2}{5} & 0 \\ 0 & 0 & -\frac{2}{5} & \frac{2}{5} & 0 \\ 0 & 0 & 6 & 6 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} = \det \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 5 & 0 & -\frac{2}{5} & 0 \\ 0 & 0 & -\frac{2}{5} & \frac{2}{5} & 0 \\ 0 & 0 & 0 & 12 & 1 \\ 0 & 0 & 0 & 2 & 1 \end{bmatrix}$$

$$\det(\mathbf{H}_3) = \det \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 5 & 0 & -\frac{2}{5} & 0 \\ 0 & 0 & -\frac{2}{5} & \frac{2}{5} & 0 \\ 0 & 0 & 0 & 12 & 1 \\ 0 & 0 & 0 & 0 & \frac{5}{6} \end{bmatrix} = -20 < 0.$$

Since $\ell = n = 3$, $(-1)^\ell \det(D^2L^*) > 0$ and $(\lambda^*, \mu^*, x^*, y^*, z^*) = (6/5, 1/5, -3, -3, 18)$ is a local maximum.

These answers are compatible with the values of $f(x, y, z)$ at the two critical points: on the constraint set, $(\lambda^*, \mu^*, x^*, y^*, z^*) = (4/5, -1/5, 2, 2, 8,)$ is a global minimum, and $(\lambda^*, \mu^*, x^*, y^*, z^*) = (6/5, 1/5, -3, -3, 18,)$ is a global maximum. ■

Example 3.38. Consider the problem of finding the extreme point of $f(x, y, z) = x^2+y^2+z^2$ on $2 = z - xy$. The points that satisfy the first order conditions for the method of Lagrange are

$$\begin{aligned} (\lambda^*, x^*, y^*, z^*) &= (4, 0, 0, 2), \\ &= (2, 1, -1, 1), \qquad \text{and} \\ &= (2, -1, 1, 1). \end{aligned}$$

The Lagrangian is $L(\lambda, x, y, z) = x^2 + y^2 + z^2 - \lambda z + \lambda xy + \lambda 2$ with bordered Hessian matrix

$$\mathbf{H}_3 = D^2L = \begin{pmatrix} 0 & y & x & -1 \\ y & 2 & \lambda & 0 \\ x & \lambda & 2 & 0 \\ -1 & 0 & 0 & 2 \end{pmatrix}.$$

At the point $(\lambda^*, x^*, y^*, z^*) = (4, 0, 0, 2)$, expanding on the first row,

$$\det(\mathbf{H}_3) = \det \begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 2 & 4 & 0 \\ 0 & 4 & 2 & 0 \\ -1 & 0 & 0 & 2 \end{pmatrix} = \det \begin{pmatrix} 0 & 2 & 4 \\ 0 & 4 & 2 \\ -1 & 0 & 0 \end{pmatrix} = 12 > 0.$$

Since $n = 3$ and $k = 1$ and both $(-1)^k \det(\mathbf{H}_3) < 0$ and $(-1)^n \det(\mathbf{H}_3) < 0$, this fails the test for either a local minimum or a local maximum, so the point is not a local extremum.

The calculation at the other two points is similar, so we consider the point $(\lambda^*, x^*, y^*, z^*) = (2, 1, -1, 1)$. The partial derivative $g_x(1, -1, 1) = -(-1) \neq 0$, so

$$\mathbf{H}_2 = \begin{pmatrix} 0 & y & x \\ y & 2 & \lambda \\ x & \lambda & 2 \end{pmatrix} = \begin{pmatrix} 0 & -1 & 1 \\ -1 & 2 & 2 \\ 1 & 2 & 2 \end{pmatrix}.$$

Expanding $\det(\mathbf{H}_2)$ on the first row,

$$\begin{aligned} \det(\mathbf{H}_2) &= \det \begin{pmatrix} 0 & -1 & 1 \\ -1 & 2 & 2 \\ 1 & 2 & 2 \end{pmatrix} = \det \begin{pmatrix} -1 & 2 \\ 1 & 2 \end{pmatrix} + \det \begin{pmatrix} -1 & 2 \\ 1 & 2 \end{pmatrix} \\ &= -4 - 4 = -8 < 0. \end{aligned}$$

Expanding $\det(\mathbf{H}_3)$ on the fourth row,

$$\begin{aligned}\det(\mathbf{H}_3) &= \det\begin{pmatrix} 0 & -1 & 1 & -1 \\ -1 & 2 & 2 & 0 \\ 1 & 2 & 2 & 0 \\ -1 & 0 & 0 & 2 \end{pmatrix} = \det\begin{pmatrix} -1 & 1 & -1 \\ 2 & 2 & 0 \\ 2 & 2 & 0 \end{pmatrix} + (2)\det\begin{pmatrix} 0 & -1 & 1 \\ -1 & 2 & 2 \\ 1 & 2 & 2 \end{pmatrix} \\ &= 0 + 2(-8) = -16 < 0.\end{aligned}$$

Since $k = 1$, $(-1)^k \det(\mathbf{H}_3) > 0$ and $(-1)^k \det(\mathbf{H}_2) > 0$ and this point is a local minimum.

A similar calculation at $(\lambda^*, x^*, y^*, z^*,) = (2, -1, 1, 1)$ shows that it is also a local minimum. ■

3.5. Exercises

3.5.1. Find the points satisfying the first order conditions for a constrained extrema and then apply the second order test to determine whether they are local maximum or local minimum.

a. $f(x, y, z) = xyz$ and $g(x, y, z) = 2x + 3y + z = 6$.

b. $f(x, y, z) = 2x + y^2 - z^2$, $g_1(x, y, z) = x - 2y = 0$, and $g_2(x, y, z) = x + z = 0$.

3. Exercises for Chapter 3

3.1. Indicate which of the following statements are *true* and which are *false*. Justify each answer: For a true statement explain why it is true and for a false statement either indicate how to make it true or indicate why the statement is false. In several of these parts, $\mathscr{F}_{\mathbf{g},\mathbf{b}} = \{\, \mathbf{x} \in \mathbb{R}^n : g_i(\mathbf{x}) \leq b_i \ \text{ for } 1 \leq i \leq m \,\}$.

a. If $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ is convex, then each of the g_i must be convex.

b. If $f : \mathscr{D} \subset \mathbb{R}^n$ is continuous and $f(\mathbf{x})$ attains a maximum on $\mathscr{D}$, then $\mathscr{D}$ is compact.

c. If $f, g_i : \mathbb{R}^n \to \mathbb{R}$ are C^1 for $1 \leq i \leq m$, $f(\mathbf{x})$ satisfies KKT-1,2,3 at $\mathbf{x}^* \in \mathscr{F}_{\mathbf{g},\mathbf{b}}$, and the constraint qualification holds at $\mathbf{x}^*$, then $\mathbf{x}^*$ must be a maximizer of $f(\mathbf{x})$ on $\mathscr{F}_{\mathbf{g},\mathbf{b}}$.

d. If $\mathscr{F}_{\mathbf{g},\mathbf{b}}$, $f : \mathbb{R}^n \to \mathbb{R}$ is C^1, $\mathbf{x}^*$ satisfies KKT-1,2,3 and is a maximizer of f, then f must be concave.

e. Assume that f, $g_i : \mathbb{R}^n \to \mathbb{R}$ are C^1 for $1 \leq i \leq m$, the constraint qualification is satisfied at all points of $\mathscr{F}_{\mathbf{g},\mathbf{b}}$, and $\mathbf{p}_1, \ldots, \mathbf{p}_k$ are the set of all the points in $\mathscr{D}_{\mathbf{g},\mathbf{b}}$ that satisfy KKT-,1,2,3. Then, $f(\mathbf{x})$ attains a maximum on $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ and $\max\{f(\mathbf{x}) : \mathbf{x} \in \mathscr{F}_{\mathbf{g},\mathbf{b}}\} = \max\{f(\mathbf{p}_j) : 1 \leq j \leq k\}$.

f. Let f, $g_j : \mathbb{R}^n_+ \to \mathbb{R}$ be C^1 for $1 \leq j \leq m$, $\mathscr{F}_{\mathbf{g},\mathbf{b}} = \{\, \mathbf{x} : g_j(\mathbf{x}) \leq b_j \ \text{ for } 1 \leq j \leq m \,\}$, and $\{x_k^*\}_{k=1}^K$ be the set of points in $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ where either (i) the KKT-1,2,3 conditions hold or (ii) the constraint qualification fails for $\mathscr{F}_{\mathbf{g},\mathbf{b}}$. Then f must have a maximum on $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ at one of the points $\{x_k^*\}_{k=1}^K$.

g. Assume that $g_j : \mathbb{R}^n_+ \to \mathbb{R}$ are continuous and convex for $1 \leq j \leq m$, $f : \mathbb{R}^n \to \mathbb{R}$ is concave, and f has a local maximum on $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ at $\mathbf{x}^*$. Then $\mathbf{x}^*$ is a global maximizer of f on $\mathscr{F}_{\mathbf{g},\mathbf{b}}$.

h. If $\mathscr{F}_{\mathbf{g},\mathbf{b}}$ is convex and f is concave on $\mathscr{F}_{\mathbf{g},\mathbf{b}}$, then f must have a maximum on $\mathscr{F}_{\mathbf{g},\mathbf{b}}$.

i. To find the maximum of $f(x, y, z, w)$ subject to the three constraints $g_i(x, y, z, w) = b_i$ for $i = 1, \ldots, 3$ using the Lagrange Multiplier Theorem, one has to solve a system of 4 equations with 4 unknowns.

j. Consider a level set $\mathscr{F} = \{\, \mathbf{x} \in \mathbb{R}^n : g_i(\mathbf{x}) = b_i \ \text{ for } 1 \leq i \leq k \,\}$. If $\mathbf{x}^*$ is a maximizer of $f(\mathbf{x})$ on the level set $\mathscr{F} = \{\, \mathbf{x} \in \mathbb{R}^n : g_i(\mathbf{x}) = b_i \ \text{ for } 1 \leq i \leq k \,\}$ with $\nabla f(\mathbf{x}^*) = \sum_{i=1}^k \lambda_i^* \nabla g_i(\mathbf{x}^*)$, then all the λ_i^* must satisfy $\lambda_i^* \geq 0$.

k. If f, $g_i : \mathscr{D} \subset \mathbb{R}^n \to \mathbb{R}$ are C^1, $\mathbf{x}^*$ is a maximizer of f on $\mathscr{F}_{\mathbf{g},\mathbf{b}}$, and $(\boldsymbol{\lambda}^*, \mathbf{x}^*)$ satisfies KKT-1, then for $1 \leq i \leq k$, either $\lambda_i^* > 0$ or $g_j(\mathbf{x}^*) < b_j$.

l. For a C^1 function $f(\mathbf{x})$ on $\mathscr{F} = \{\, \mathbf{x} \in \mathbb{R}^2_+ : 1 - x_1 x_2 \leq 0 \,\}$, if $(\mathbf{x}^*, \lambda^*)$ satisfy KKT-1,2,3 with $\lambda^* > 0$, then $\mathbf{x}^*$ must be a maximizer.

Chapter 4

Dynamic Programming

This chapter focuses on maximizing over more than one time period where unused resources are carried forward to the next time period. The easiest case considers a finite number of time periods, which is said to have a finite horizon. This type of problem can be solved as a Karush-Kuhn-Tucker problem with many variables. However, a simpler method is to solve the problem recursively, one period at a time. This approach is called dynamic programming. The harder case considers an infinite number of time periods, or infinite horizon. This type of problem requires solving an equation for a function rather than just the value of the function in some $\mathbb{R}^n$. Although the solution method looks similar to the finite horizon case, proving that a solution exists is much harder.

In both finite- and infinite-horizon dynamic programs, the function maximized is restricted to a domain that usually depends on the state, i.e., depends on a parameter. Such dependence of a set on a parameter is called a *correspondence*. For each parameter, there can be more than one maximizer, so the set of maximizers in general is also a correspondence and not a function. We start the chapter by considering correspondences and various types of continuity as the parameter varies. The Parametric Maximization Theorem indicates how the set of maximizers and the maximal value varies with the parameter. This theorem is at the heart of the solution method of dynamic programming problems.

4.1. Parametric Maximization and Correspondences

In game theory, the best response to the opponents choice is an example of what is called a correspondence because there is a set of possible responses. In this section, we introduce correspondences and indicate how they arise in maximization problems that depend on a parameter.

Example 4.1. The assignment $\mathscr{C}_1(s) = [0, s]$ of an interval that depends on the parameter $0 \le s \le 1$ is an example of a correspondence. Its graph is the set of points

$$\mathrm{Gr}(\mathscr{C}_1) = \{\, (s, x) : 0 \le s \le 1,\ \ x \in [0, s] \,\}$$

and is shown in Figure 4.1.1.

The general definitions are as follows.

Definition. A *correspondence* $\mathscr{C}$ from $\mathbf{S} \subset \mathbb{R}^\ell$ to $\mathbf{X} \subset \mathbb{R}^n$ is a map that associates a nonempty subset $\mathscr{C}(\mathbf{s})$ of $\mathbf{X}$ to each $\mathbf{s} \in \mathbf{S}$. Let $\mathscr{P}(\mathbf{X})$ be the power set of $\mathbf{X}$ of all nonempty subsets of $\mathbf{X}$. Thus, $\mathscr{C}$ takes its values in $\mathscr{P}(\mathbf{X})$, $\mathscr{C} : \mathbf{S} \to \mathscr{P}(\mathbf{X})$.

The *graph* of a correspondence $\mathscr{C} : \mathbf{S} \to \mathscr{P}(\mathbf{X})$ is the set

$$\mathrm{Gr}(\mathscr{C}) = \{\, (\mathbf{s}, \mathbf{x}) : \mathbf{s} \in \mathbf{S},\ \ \mathbf{x} \in \mathscr{C}(\mathbf{s}) \,\} \subset \mathbf{S} \times \mathbf{X}.$$

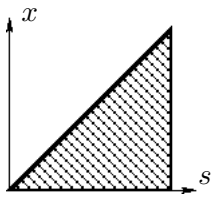


Figure 4.1.1. Graph of correspondence $\mathscr{C}_1(s) = [0, s]$

Definition. Various properties that a correspondence $\mathscr{C} : \mathbf{S} \to \mathscr{P}(\mathbb{R}^n)$ might possess are defined as follows.

$\mathscr{C}$ is *bounded* provided that there is a $K > 0$ such that $\mathscr{C}(\mathbf{s}) \subset \overline{\mathbf{B}}(\mathbf{0}, K)$ for all $\mathbf{s} \in \mathbf{S}$.

$\mathscr{C}$ is *locally bounded* provided that for each $\mathbf{s}_0 \in \mathbf{S}$, there are $\delta > 0$ and $K > 0$, such that $\mathscr{C}(\mathbf{s}) \subset \overline{\mathbf{B}}(\mathbf{0}, K)$ for all $\mathbf{s} \in \mathbf{B}(\mathbf{s}_0, \delta) \cap \mathbf{S}$.

$\mathscr{C}$ is a *closed-graphed* provided that its graph $\mathrm{Gr}(\mathscr{C})$ is a closed subset of $\mathbf{S} \times \mathbb{R}^n$.

$\mathscr{C}$ is *closed-valued* (resp. *compact-valued*, or *convex-valued*) provided that $\mathscr{C}(\mathbf{s})$ is a closed (resp. compact, or convex) subset of $\mathbb{R}^n$ for every $\mathbf{s} \in \mathbf{S}$.

Remark. The terms that include "valued" refer to properties for each parameter $\mathbf{s}$ and not properties for all $\mathbf{s}$ at once. Thus, a closed-valued correspondence is not necessarily closed-graphed, but a closed-graphed correspondence is closed-valued.

We next give several examples of correspondences with various properties.

Example 4.2. With $\mathbf{S} = [0, 2]$ and $\mathbf{X} = \mathbb{R}$, define two correspondences by

$$\mathscr{C}_2(s) = \begin{cases} [1, 2] & \text{for } 0 \le s < 0.5, \ \ 1.5 < s \le 2, \\ [0, 3] & \text{for } 0.5 \le s \le 1.5, \end{cases}$$

$$\mathscr{C}_3(s) = \begin{cases} [1, 2] & \text{for } 0 \le s \le 0.5, \ \ 1.5 \le s \le 2, \\ [0, 3] & \text{for } 0.5 < s < 1.5. \end{cases}$$

Figure 4.1.2 shows their graphs. These correspondence only differ for $s = 0.5$ and 1.5. Both $\mathscr{C}_2$ and $\mathscr{C}_3$ are compact-valued and bounded; $\mathscr{C}_2$ is closed-graphed but not $\mathscr{C}_3$. ■

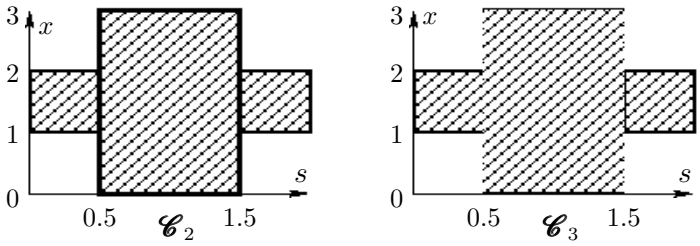


Figure 4.1.2. Graphs of correspondences in Example 4.2

Example 4.3. The correspondence

$$\mathscr{C}_4(s) = \begin{cases} \left\{ \dfrac{1}{s} \right\} & s \neq 0 \\ \{0\} & s = 0 \end{cases}$$

has a single point for each parameter $s \in \mathbb{R}$ and is neither bounded nor locally bounded near $s = 0$. See Figure 4.1.3. This correspondence is (i) closed-graphed and (ii) compact-valued since it is a single point for each s. ■

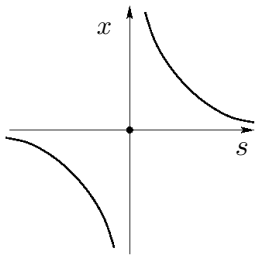


Figure 4.1.3. Graph of correspondence $\mathscr{C}_4$ in Example 4.3

Our main use of correspondences occurs in maximization problems with a parameter that have a general setup is as follows. As $\mathbf{s}$ varies over a parameter space $\mathbf{S}$, assume that the feasible set $\mathscr{F}(\mathbf{s}) \subset \mathbb{R}^n$ can vary with the parameter and is compact for each $\mathbf{s}$. Thus, $\mathscr{F} : \mathbf{S} \to \mathscr{P}(\mathbb{R}^n)$ is a compact-valued correspondence. Assume that $f : \mathrm{Gr}(\mathscr{F}) \subset \mathbf{S} \times \mathbb{R}^n \to \mathbb{R}$ is continuous. For each $\mathbf{s} \in \mathbf{S}$, denote the maximal value of $f(\mathbf{s}, \mathbf{x})$ subject to $\mathbf{x} \in \mathscr{F}(\mathbf{s})$ by

$$f^*(\mathbf{s}) = \max\{\, f(\mathbf{s}, \mathbf{x}) : \mathbf{x} \in \mathscr{F}(\mathbf{s}) \,\}$$

and the set of feasible maximizers by

$$\mathscr{F}^*(\mathbf{s}) = \{\, \mathbf{x} \in \mathscr{F}(\mathbf{s}) : f(\mathbf{s}, \mathbf{x}) = f^*(\mathbf{s}) \,\}.$$

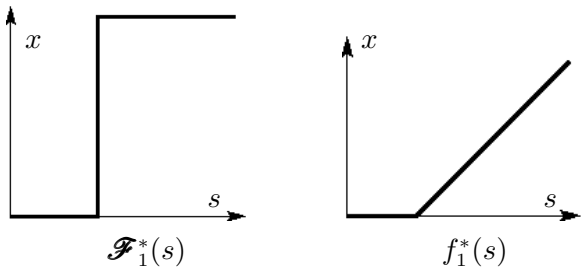
Since $\mathscr{F}(\mathbf{s})$ and $\mathscr{F}^*(\mathbf{s})$, are sets for each $\mathbf{s}$, the assignments from $\mathbf{s}$ to $\mathscr{F}(\mathbf{s})$ or $\mathscr{F}^*(\mathbf{s})$ are examples of correspondences. On the other hand, $f^*(\mathbf{s})$ is a real-valued function. The following examples illustrate how $f^*(\mathbf{s})$ and $\mathscr{F}^*(\mathbf{s})$ can vary with $\mathbf{s}$.

Example 4.4. Consider the example with $f_1(s, x) = (s - 1/3)\, x$ for $s \in [0, 1] = \mathbf{S}_1$ and $x \in [0, 1] = \mathscr{F}_1(s)$. (The feasible set does not vary with s in this example.) Since

$$\frac{\partial f_1}{\partial x}(s, x) = (s - 1/3) \begin{cases} < 0 & \text{for } s < 1/3 \text{ and for all } x \in [0, 1], \\ \equiv 0 & \text{for } s = 1/3 \text{ and for all } x \in [0, 1], \\ > 0 & \text{for } s > 1/3 \text{ and for all } x \in [0, 1], \end{cases}$$

$$\mathscr{F}_1^*(s) = \begin{cases} \{0\} & \text{for } s < 1/3 \\ [0, 1] & \text{for } s = 1/3 \\ \{1\} & \text{for } s > 1/3. \end{cases} \quad \text{and} \quad f_1^*(s) = \begin{cases} 0 & \text{for } s \le 1/3 \\ s - 1/3 & \text{for } s > 1/3. \end{cases}$$

See Figure 4.1.4. The set-valued correspondence $\mathscr{F}_1^*(s)$ changes from $\{0\}$ to $\{1\}$ as s crosses $1/3$, while the maximal value $f_1^*(s)$ is continuous. Also, note that $\mathscr{F}_1^*(s)$ is (i) bounded, (ii) compact-valued, and (iii) closed-graphed. In a strategic game from game theory, if $f_1(s, x)$ is your payoff for mixed strategies of s by other player and x by you, then $\mathscr{F}_1^*(s)$ is called the *best response correspondence*. ■



$\mathscr{F}_1^*(s)$ $\qquad$ $f_1^*(s)$

Figure 4.1.4. Graphs of maximizer set and maximal-value function for Example 4.4

Example 4.5. Let $f_2(s,x) = -\frac{1}{4}x^4 + \frac{1}{3}sx^3 + \frac{1}{2}x^2$ for $s \in \mathbb{R} = \mathbf{S}_2$ and $x \in \mathbb{R} = \mathscr{F}_2(s)$. Its graph for different signs of s is given in Figure 4.1.5. Its partial derivative is $f_{2x}(s,x) = -x^3 + sx^2 + x$, so the critical points are 0 and $x_s^{\pm} = \frac{1}{2}(s \pm \sqrt{s^2+4})$. The second partial derivative is $f_{2xx}(s,x) = -3x^2 + 2sx + 1$. At $x = 0$, $f_{2xx}(s,0) = 1 > 0$, so 0 is a local minimum. At $x_s^{\pm}$, $f_{2xx}(s,x_s^{\pm}) = -(x_s^{\pm})^2 + 2\left[-(x_s^{\pm})^2 + sx_s^{\pm} + 1\right] - 1 = -(x_s^{\pm})^2 - 1 < 0$, and f attains a maximum at x_s^+ or x_s^-. For $s = 0$, $x_0^{\pm} = \pm 1$, $f_2(0,\pm 1) = 1/4 > 0 = f(0,0)$, and $\mathscr{F}_2^*(0) = \{-1, 1\}$.

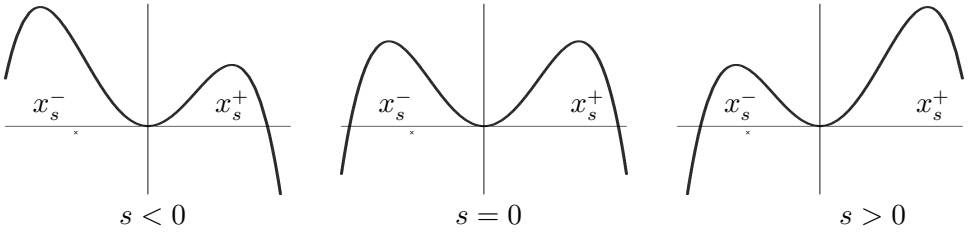


Figure 4.1.5. Graph of $f_2(s,x)$ for Example 4.5

Rather than calculate $f_2(s,x_s^{\pm})$ as a function of s, we determine the sign of its (total) derivative with respect to s at these points:

$$\frac{d}{ds}f_2(s,x_s^{\pm}) = f_{2x}(s,x_s^{\pm})\frac{dx_s^{\pm}}{ds} + \tfrac{1}{3}\left(x_s^{\pm}\right)^3 = \tfrac{1}{3}\left(x_s^{\pm}\right)^3.$$

This derivative has the same sign as $x_s^{\pm}$, so

$$\begin{aligned} f_2(s,x_s^-) > f(0,\pm 1) > f_2(s,x_s^+) \qquad & \text{for } s < 0, \\ f_2(s,x_s^-) < f(0,\pm 1) < f_2(s,x_s^+) \qquad & \text{for } s > 0. \end{aligned}$$

Thus,

$$\mathscr{F}_2^*(s) = \begin{cases} \{x_s^-\} & \text{for } s < 0 \\ \{x_0^-, x_0^+\} = \{-1, 1\} & \text{for } s = 0 \\ \{x_s^+\} & \text{for } s > 0. \end{cases}$$

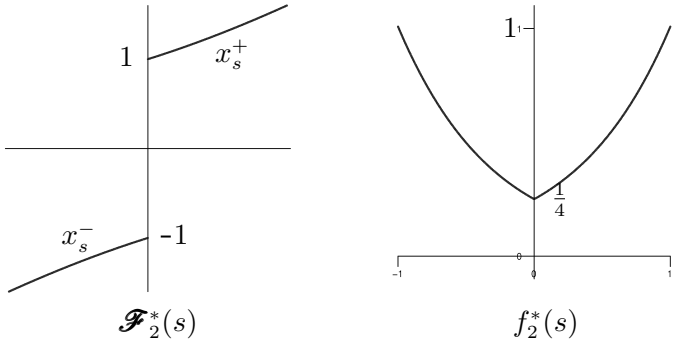


Figure 4.1.6. Graphs of maximizer set and maximal-value function for Example 4.5

Theorem 4.8 shows that f_2^* must be continuous, and Figure 4.1.6(b) gives its numerically calculated graph. The correspondence $\mathscr{F}_2^*$ switches the maximizer from x_s^- to x_s^+ at $s = 0$ but is (i) compact-valued, (ii) locally bounded, and (iii) closed-graphed. (See Theorem 4.8.) ■

For correspondences, we consider not only continuity but also two weaker properties called hemicontinuity. The precise conditions involve the neighborhood of a set that we define first.

Definition. For a set $\mathbf{A} \subset \mathbb{R}^n$, the *$\epsilon$-neighborhood of* $\mathbf{A}$ is the set

$$\mathbf{B}(\mathbf{A}, \epsilon) = \{\, \mathbf{x} \in \mathbb{R}^n : \text{ there is a } \mathbf{y} \in \mathbf{A} \text{ with } \|\mathbf{x} - \mathbf{y}\| < \epsilon \,\}.$$

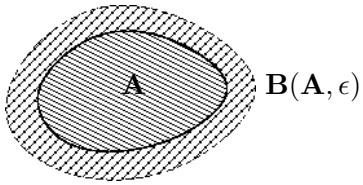


Figure 4.1.7. Neighborhood of a set

A function $f(x)$ is continuous at x_0 provided that $\lim_{x \to x_0} f(x) = f(x_0)$, i.e., for all $\epsilon > 0$, there exists a $\delta > 0$ such that if $|x - x_0| < \delta$, then $|f(x) - f(x_0)| < \epsilon$. The conclusion can be written as $f(x) \in \mathbf{B}(f(x_0), \epsilon)$. If we apply a similar condition to a correspondence, it restricts the extent that the correspondence can expand or implode (vacate regions).

Definition. A compact-valued correspondence $\mathscr{C} : \mathbf{S} \subset \mathbb{R}^\ell \to \mathscr{P}(\mathbf{X})$ is *upper-hemicontinuous* (uhc) at $\mathbf{s}_0 \in \mathbf{S}$ provided that $\mathscr{C}(\mathbf{s})$ must remain inside a small neighborhood of $\mathscr{C}(\mathbf{s}_0)$ for small changes of $\mathbf{s}$ away from $\mathbf{s}_0$, i.e., any $\epsilon > 0$ there exists $\delta > 0$ such that if $\mathbf{s} \in \mathbf{B}(\mathbf{s}_0, \delta) \cap \mathbf{S}$ then $\mathscr{C}(\mathbf{s}) \subset \mathbf{B}(\mathscr{C}(\mathbf{s}_0), \epsilon)$. This restricts the amount the correspondence can expand.

We say that $\mathscr{C}$ is *upper-hemicontinuous on* $\mathbf{S}$ if it is upper-hemicontinuous at each $\mathbf{s} \in \mathbf{S}$.

Definition. A compact-valued correspondence $\mathscr{C} : \mathbf{S} \to \mathscr{P}(\mathbf{X})$ is *lower-hemicontinuous* (lhc) at $\mathbf{s}_0 \in \mathbf{S}$ provided that for each $\mathbf{x}_0 \in \mathscr{C}(\mathbf{s}_0)$, there is a point of $\mathscr{C}(\mathbf{s})$ nearby, i.e., for any point $\mathbf{x}_0 \in \mathscr{C}(\mathbf{s}_0)$ and any $\epsilon > 0$, there exists $\delta > 0$ such that if $\mathbf{s} \in \mathbf{B}(\mathbf{s}_0, \delta) \cap \mathbf{S}$ then $\mathbf{B}(\mathbf{x}_0, \epsilon) \cap \mathscr{C}(\mathbf{s}) \neq \emptyset$ or $\mathbf{x}_0 \in \mathbf{B}(\mathscr{C}(\mathbf{s}), \epsilon)$. This is equivalent to saying that for any $\epsilon > 0$, there exists $\delta > 0$ such that if $\mathbf{s} \in \mathbf{B}(\mathbf{s}_0, \delta) \cap \mathbf{S}$ then $\mathscr{C}(\mathbf{s}_0) \subset \mathbf{B}(\mathscr{C}(\mathbf{s}), \epsilon)$. This restricts the amount the correspondence can implode or vacate the region $\mathscr{C}(\mathbf{s}_0)$.

We say that $\mathscr{C}$ is *lower-hemicontinuous on* $\mathbf{S}$ if it is lower-hemicontinuous at each $\mathbf{s} \in \mathbf{S}$.

Definition. A compact-valued correspondence is said to be *continuous* provided that it is both upper- and lower-hemicontinuous. Thus, it is continuous provided that for any $\epsilon > 0$, there exists $\delta > 0$ such that if $\mathbf{s} \in \mathbf{B}(\mathbf{s}_0, \delta) \cap \mathbf{S}$ then $\mathscr{C}(\mathbf{s}_0) \subset \mathbf{B}(\mathscr{C}(\mathbf{s}), \epsilon)$ and $\mathscr{C}(\mathbf{s}) \subset \mathbf{B}(\mathscr{C}(\mathbf{s}_0), \epsilon)$. Thus for small $\mathbf{s} - \mathbf{s}_0$ and for each point in $\mathscr{C}(\mathbf{s}_0)$ or $\mathscr{C}(\mathbf{s})$, there is a point nearby in the other set and the two sets are close to each other.

Although we give examples of lhc correspondences, we mainly use uhc and continuous correspondences.

Example 4.6. The continuity of the correspondences given in the preceding examples is as follows: $\mathscr{C}_1$ is continuous; $\mathscr{C}_2$, $\mathscr{F}_1^*$, and $\mathscr{F}_2^*$ are upper-hemicontinuous but not lower-hemicontinuous nor continuous at $\mathbf{s} = 0.5$ or 1.5; $\mathscr{C}_3$ is lower-hemicontinuous but not upper-hemicontinuous nor continuous at $\mathbf{s} = 0.5$ or 1.5; finally, $\mathscr{C}_4$ is neither lower-hemicontinuous nor upper-hemicontinuous at $\mathbf{s} = 0$,

$$\mathscr{C}_4(s) = \left\{\frac{1}{s}\right\} \not\subseteq \mathbf{B}(\mathscr{C}_4(0), \epsilon) = (-\epsilon, \epsilon) \qquad \text{and}$$

$$\mathscr{C}_4(0) = \{0\} \not\subseteq \mathbf{B}\left(\mathscr{C}_4(s), \epsilon\right) = \left(-\epsilon + \frac{1}{s}, \epsilon + \frac{1}{s}\right).$$

Remark. There is a related but different concept of upper- or lower-semicontinuous function. Unfortunately, the graph of an upper-semicontinuous function is not the graph of an upper-hemicontinuous correspondence. Because of this confusion, we will not consider the semi-continuity of functions.

Proposition 4.7. *Let $\mathscr{C} : \mathbf{S} \to \mathscr{P}(\mathbf{X})$ be a compact-valued correspondence and locally bounded. Then, $\mathscr{C}$ is upper-hemicontinuous iff $\mathscr{C}$ is a closed-graphed correspondence.*

Proof. Assume that $\mathscr{C}$ is not closed-graphed. Then there exists a point $(\mathbf{s}_0, \mathbf{x}_0) \in \mathrm{cl}(\mathrm{Gr}(\mathscr{C})) \setminus \mathrm{Gr}(\mathscr{C})$, so $\mathbf{x}_0 \notin \mathscr{C}(\mathbf{s}_0)$. Since $\mathscr{C}(\mathbf{s}_0)$ is compact, there exists $\epsilon > 0$ such that $\mathbf{x}_0 \notin \mathbf{B}(\mathscr{C}(\mathbf{s}_0), \epsilon)$. But since $(\mathbf{s}_0, \mathbf{x}_0)$ is in the closure of the graph, for every $\delta > 0$, there is a point $(\mathbf{s}_\delta, \mathbf{x}_\delta) \in \mathrm{Gr}(\mathscr{C})$ with $\|\mathbf{s}_\delta - \mathbf{s}_0\| < \delta$ and $\|\mathbf{x}_\delta - \mathbf{x}_0\| < \delta$. By taking $\delta < \epsilon/2$, $\mathbf{x}_\delta \notin \mathbf{B}(\mathscr{C}(\mathbf{s}_0), \epsilon - \delta) \supset \mathbf{B}(\mathscr{C}(\mathbf{s}_0), \epsilon/2)$, so $\mathscr{C}(\mathbf{s}_\delta) \not\subseteq \mathbf{B}(\mathscr{C}(\mathbf{s}_0), \epsilon/2)$. Since $\|\mathbf{s}_\delta - \mathbf{s}_0\| < \delta$, and $\delta > 0$ is arbitrarily small, $\mathscr{C}$ is not upper-hemicontinuous at $\mathbf{s}_0$.

If $\mathscr{C}$ is not upper-hemicontinuous at s_0, then there exists some $\epsilon > 0$ such that for any $\delta > 0$, there exists some $(\mathbf{s}_\delta, \mathbf{x}_\delta)$ with $\|\mathbf{s}_\delta - \mathbf{s}_0\| < \delta$ and $\mathbf{x}_\delta \in \mathscr{C}(\mathbf{s}_\delta) \setminus \mathbf{B}(\mathscr{C}(\mathbf{s}_0), \epsilon)$. If $\mathscr{C}$ is locally bounded, the $(\mathbf{s}_\delta, \mathbf{x}_\delta)$ must accumulate to some point not on the graph, so the graph is not closed. □

Remark. The example $\mathscr{C}_4$ given above shows why the correspondence must be locally bounded in this proposition.

We can now state the principal result of the section which will be used in the rest of the chapter.

Theorem 4.8 (Parametric Maximization Theorem). *Assume that the feasible set $\mathscr{F} : \mathbf{S} \to \mathscr{P}(\mathbf{X})$ is a compact-valued and continuous correspondence and $f : \mathrm{Gr}(\mathscr{F}) \subset \mathbf{S} \times \mathbf{X} \to \mathbb{R}$ is a continuous function.*

Then, $f^*(\mathbf{s}) = \max\{ f(\mathbf{s}, \mathbf{x}) : \mathbf{x} \in \mathscr{F}(\mathbf{s}) \}$ is a continuous function and $\mathscr{F}^*(\mathbf{s}) = \{ \mathbf{x} \in \mathscr{F}(\mathbf{s}) : f(\mathbf{s}, \mathbf{x}) = f^*(\mathbf{s}) \}$ is a compact-valued upper-hemicontinuous correspondence on $\mathbf{S}$.

If $\mathscr{F}^*(\mathbf{s})$ is a single point for each $\mathbf{s}$, then this correspondence is continuous and defines a continuous function.

Remark. If $f(\mathbf{s}, \mathbf{x})$ is strictly concave as a function of $\mathbf{x}$ for each $\mathbf{s}$, then each $\mathscr{F}^*(\mathbf{s})$ is a single point and so $\mathscr{F}^*$ is continuous.

Remark. Examples 4.4 and 4.5 both satisfy the assumptions of the theorem and have sets of maximizers that are upper-hemicontinuous but not continuous.

Example 4.9. Let $\mathbf{S} = \mathbf{X} = \mathbb{R}_+$, $\mathscr{F}(s) = [0, s]$, and $h : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ be defined by

$$h(s, x) = x^{\frac{1}{2}} + (s - x)^{\frac{1}{2}}.$$

The function h is continuous and the feasible set $\mathscr{F}(s) = [0, s]$ is a continuous correspondence, so the Parametric Maximization Theorem applies. The function h is differentiable for positive values of x and s, and the critical point satisfies

$$\begin{aligned} 0 = \frac{\partial h}{\partial x} &= \frac{1}{2} x^{-\frac{1}{2}} - \frac{1}{2} (s - x)^{-\frac{1}{2}}, \\ x^{-\frac{1}{2}} &= (s - x)^{-\frac{1}{2}}, \\ s - x &= x, \\ s &= 2x, \\ \bar{x} &= \tfrac{1}{2} s \in [0, s]. \end{aligned}$$

Since $\dfrac{\partial^2 h}{\partial x^2} = -\frac{1}{4}x^{-\frac{3}{2}} - \frac{1}{4}(s-x)^{-\frac{3}{2}} < 0$ for all $x \geq 0$, $h(x,s)$ is a concave function of x, and $\bar{x}$ is the unique maximizer on $[0,s]$. Also note that $\mathscr{F}^*(s) = \{\frac{1}{2}s\}$ is a continuous correspondence and $h^*(s) = (s/2)^{\frac{1}{2}} + (s/2)^{\frac{1}{2}} = 2^{\frac{1}{2}}s^{\frac{1}{2}}$ is continuous. ■

4.1.1. Budget Correspondence for Commodity Bundles

In consumer theory, a commodity bundle of n items is an element of $\mathbb{R}^n_+$. Often, it is restricted to a set based on prices of the commodities and wealth of the individual. We show explicitly that the set of allowable commodity bundles is a continuous correspondence.

Theorem 4.10. *Let the budget correspondence $\mathscr{B} : \mathbb{R}^{n+1}_{++} \to \mathscr{P}(\mathbb{R}^n_+)$ be defined by*

$$\mathscr{B}(\mathbf{p}, w) = \{\, \mathbf{x} \in \mathbb{R}^n_+ : \mathbf{p} \cdot \mathbf{x} \leq w \,\},$$

where the prices $p_i > 0$ for $1 \leq i \leq n$, the wealth $w > 0$, and the parameter space is $\mathbf{S} = \{\, (\mathbf{p}, w) \in \mathbb{R}^{n+1}_{++} \,\}$. Then, $\mathscr{B}$ is a continuous, compact-valued correspondence.

The following corollary follows immediately from the previous theorem and the Parametric Maximization Theorem.

Corollary 4.11. *Assume $u : \mathbb{R}^n_+ \to \mathbb{R}$ is a continuous utility function. Let $v : \mathbb{R}^{n+1}_{++} \to \mathbb{R}$ be the indirect utility function that is the maximal value of utility on the budget constraint,*

$$v(\mathbf{p}, w) = u^*(\mathbf{p}, w) = \max\{\, u(\mathbf{x}) : \mathbf{x} \in \mathscr{B}(\mathbf{p}, w) \,\},$$

and let $\mathbf{d} : \mathbb{R}^{n+1}_{++} \to \mathscr{P}(\mathbb{R}^n_+)$ be the demand correspondence which achieve this maximum,

$$\mathbf{d}(\mathbf{p}, w) = \mathscr{B}^*(\mathbf{p}, w) = \{\, \mathbf{x} \in \mathscr{B}(\mathbf{p}, w) : u(\mathbf{x}) = v(\mathbf{p}, w) \,\}.$$

Then the indirect utility function v is a continuous function and the demand correspondence $\mathbf{d}$ is a compact-valued upper-hemicontinuous correspondence.

Proof of Theorem 4.10. Since it is defined by linear constraints that allow equality, it is closed-valued. Since every $x_i \leq w/p_i$, each $\mathscr{B}(\mathbf{p}, w)$ is bounded and so compact-valued. Intuitively, it is a continuous correspondence, but we prove it explicitly by means of the following two lemmas. □

Lemma 4.12. *$\mathscr{B}$ is upper-hemicontinuous.*

Proof. Since $\mathscr{B}$ is compact-valued, by Proposition 4.7, it suffices to prove that $\mathscr{B}$ is locally bounded and closed-graphed.

Take an allowable $(\bar{\mathbf{p}}, \bar{w}) \in \mathbb{R}^{n+1}_{++}$. The prices $\bar{\mathbf{p}} \gg 0$, so there exists $\delta > 0$ such that $\bar{p}_i \geq 2\delta$ for all i. If $\|\mathbf{p} - \bar{\mathbf{p}}\| < \delta$, $|w - \bar{w}| < \delta$, and $\mathbf{x} \in \mathscr{B}(\mathbf{p}, w)$, then for any fixed $1 \leq i \leq n$, $p_i > \bar{p}_i - \delta \geq \delta$,

$$\delta\, x_i \leq p_i x_i \leq \sum\nolimits_j p_j x_j \leq w \leq \bar{w} + \delta, \qquad \text{and}$$

$$x_i \leq \frac{\bar{w} + \delta}{\delta}.$$

This shows that there is one bound on points $\mathbf{x} \in \mathscr{B}(\mathbf{p}, w)$ for $\|\mathbf{p} - \bar{\mathbf{p}}\| < \delta$ and $|w - \bar{w}| < \delta$, i.e., $\mathscr{B}(\mathbf{p}, w) \subset [0, C]^n$ where $C = \dfrac{\bar{w} + \delta}{\delta}$ and $\mathscr{B}$ is locally bounded.

The function $h(\mathbf{p}, w, \mathbf{x}) = \min\{\, 0,\ w - \mathbf{p} \cdot \mathbf{x} \,\}$ is easily seen to be continuous. Restricting h to $\mathscr{F} = [0, C]^n$ for $\|\mathbf{p} - \bar{\mathbf{p}}\| < \delta$ and $|w - \bar{w}| < \delta$, it follows that $h^*(\mathbf{p}, w)$ is continuous and $\mathscr{F}^*(\mathbf{p}, w)$ is upper-hemicontinuous. Since $h(\mathbf{p}, w, \mathbf{0}) = 0$ and $h(\mathbf{p}, w, \mathbf{x}) \leq 0$, $h^*(\mathbf{p}, w) = 0$ or $w - \mathbf{p} \cdot \mathbf{x} \geq 0$ for $\mathbf{x} \in \mathscr{F}^*(\mathbf{p}, w)$. Thus, $\mathscr{F}^*(\mathbf{p}, w) = \mathscr{B}(\mathbf{p}, w)$ and the set of commodity bundles is upper-hemicontinuous as desired. □

Lemma 4.13. *$\mathscr{B}$ is lower-hemicontinuous.*

Proof. Fix $(\mathbf{p}_0, w_0) \in \mathbb{R}^{n+1}_{++}$, $\mathbf{x}_0 \in \mathscr{B}(\mathbf{p}_0, w_0)$, and $\epsilon > 0$. Then there exists $\bar{x} \in \mathbf{B}(\mathbf{x}_0, \epsilon)$ such that $\bar{\mathbf{x}} \gg 0$ and $\bar{w} - \bar{\mathbf{p}} \cdot \bar{\mathbf{x}} > 0$. The function $g(\mathbf{p}, w, \mathbf{x}) = w - \mathbf{p} \cdot \mathbf{x}$ is continuous and $g(\bar{\mathbf{p}}, \bar{w}, \bar{\mathbf{x}}) > 0$, so there exists $\delta > 0$ such that $g(\mathbf{p}, w, \bar{\mathbf{x}}) > 0$ for $(\mathbf{p}, w)$ within ϵ of $(\bar{\mathbf{p}}, \bar{w})$. Therefore, $\bar{\mathbf{x}} \in \mathscr{B}(\mathbf{p}, w)$ and $\mathbf{x}_0 \in \mathbf{B}\left(\mathscr{B}(\mathbf{p}, w), \epsilon\right)$. Since this is possible for any $(\mathbf{p}_0, w_0) \in \mathbb{R}^{n+1}_{++}$, $\mathbf{x}_0 \in \mathscr{B}(\mathbf{p}_0, w_0)$, and $\epsilon > 0$, $\mathscr{B}$ is lower-hemicontinuous. □

4.1.2. Existence of a Nash Equilibrium

Consider a two player strategic game where each player has a finite number of pure choices, n_i for the i^{th} player. We label the choices of the i^{th} player by integers $1 \leq j \leq n_i$. Each player has a payoff $u_i(j, k)$ that depends on the pure choices of both players. For the i^{th} player, a mixed strategy is a distribution $(s_{ij})_{j=1}^{n_i}$ such that each $s_{ij} \geq 0$ and $\sum_{j=1}^{n_i} s_{ij} = 1$, where s_{ij} is the probability of playing strategy j. The set of all such mixed strategies $\mathbf{S}_i$ is a compact, convex subset of $\mathbb{R}^{n_i}$, a simplex. The payoff on pure strategies induces a Bernoulli payoff function on mixed strategies

$$U_i(\mathbf{s}_1, \mathbf{s}_2) = \sum_{1 \leq j \leq n_1,\, 1 \leq k \leq n_2} s_{1j}\, s_{2k}\, u_i(j, k).$$

The functions U_1 and U_2 are continuous functions on $\mathbf{S}_1 \times \mathbf{S}_2$. Denote the maximal payoff for the i^{th} player in response to a mixed strategy by of $\mathbf{s}_{\text{-}i}$ for the other player by

$$m_i(\mathbf{s}_{\text{-}i}) = \max\{\, U_i(\mathbf{s}_i, \mathbf{s}_{\text{-}i}) : \mathbf{s}_i \in \mathbf{S}_i \,\},$$

and the *best response correspondence* for player i by

$$\mathbf{b}_i(\mathbf{s}_{\text{-}i}) = \{\, \mathbf{s}_i : U_i(\mathbf{s}_i, \mathbf{s}_{\text{-}i}) = m_i(\mathbf{s}_{\text{-}i}) \,\}.$$

A *Nash equilibrium* is a pair of mixed strategies $(\mathbf{s}_1^*, \mathbf{s}_2^*)$ such that $\mathbf{s}_1^* \in \mathbf{b}_1(\mathbf{s}_2^*)$ is a best response to $\mathbf{s}_2^*$ and $\mathbf{s}_2^* \in \mathbf{b}_2(\mathbf{s}_1^*)$ is a best response to $\mathbf{s}_1^*$, $(\mathbf{s}_1^*, \mathbf{s}_2^*) \in \mathbf{b}_1(\mathbf{s}_2^*) \times \mathbf{b}_2(\mathbf{s}_1^*)$.

Given $\mathbf{s}_{\text{-}i}$, there are a finite number of pure strategies that realize $m_i(\mathbf{s}_{\text{-}i})$ and $\mathbf{b}_i(\mathbf{s}_{\text{-}i})$ is the set of all convex combinations of these pure strategies. Therefore, the correspondence

$$(\mathbf{s}_1, \mathbf{s}_2) \in \mathbf{S}_1 \times \mathbf{S}_2 \mapsto \mathbf{b}(\mathbf{s}_1, \mathbf{s}_2) = \mathbf{b}_1(\mathbf{s}_2) \times \mathbf{b}_2(\mathbf{s}_1) \subset \mathbf{S}_1 \times \mathbf{S}_2$$

is convex valued. Since U_1 and U_2 are continuous and the feasible set is the same for all strategies, $\mathbf{b}_1 s_2)$ and $\mathbf{b}_2(\mathbf{s}_1)$ are each upper-hemicontinuous and so is $\mathbf{b}(\mathbf{s}_1, \mathbf{s}_2)$. The existence of a Nash equilibrium in mixed strategies then follows from the Kakutani Fixed Point Theorem.

Theorem 4.14 (Kakutani). *Let $\mathbf{S}$ be a non-empty, compact, and convex subset of some Euclidean space $\mathbb{R}^n$ and $\mathscr{C} : \mathbf{S} \to \mathscr{P}(\mathbf{S})$ be a upper-hemicontinuous and convex valued correspondence. Then, there exists a $\mathbf{p}^* \in \mathbf{S}$ such that $\mathbf{p}^* \in \mathscr{C}(\mathbf{p}^*)$.*

See [**14**] for more details and examples. The book [**1**] by Arrow and Hahn has a proof and applications to economics.

4.1. Exercises

4.1.1. Let $\mathbf{S} = [0, 1]$ and $\mathbf{S} = \mathbb{R}$. For each of the following correspondences $\mathscr{C} : \mathbf{S} \to \mathscr{P}(\mathbb{R})$, (i) draw its graph and (ii) determine whether it is uhc, and/or continuous. *Hint:* By Proposition 4.7, the correspondence is upper-hemicontinuous if and only if it is closed-graphed. (They satisfy the other assumptions of the proposition.)

a. $\mathscr{C}(s) = \begin{cases} [0, 2s] & \text{for } s \in [0, 1/2)\,, \\ [0, 2-2s] & \text{for } s \in [1/2, 1]\,. \end{cases}$

b. $\mathscr{C}(s) = \begin{cases} [0, 1-2s] & \text{for } s \in [0, 1/2]\,, \\ [0, 2-2s] & \text{for } s \in (1/2, 1]\,. \end{cases}$

c. $\mathscr{C}(s) = \begin{cases} [0, 1-2s] & \text{for } s \in [0, 1/2)\,, \\ [0, 2-2s] & \text{for } s \in [1/2, 1]\,. \end{cases}$

d. $\mathscr{C}(s) = \{0, s\}$ for $s \in [0, 1]$ (two points for each s).

e. $\mathscr{C}(s) = \begin{cases} \{0\} & \text{for } s < 0 \quad \text{(one point for each } s\text{)}, \\ \{-1, 1\} & \text{for } s \geq 0 \quad \text{(two points for each } s\text{)}. \end{cases}$

4.1.2. Let $\mathbf{X} = [0, 1] = \mathbf{S}$, and $f : \mathbf{S} \times \mathbf{X} \to \mathbb{R}$ be defined by $f(s, x) = 3 + 2x - 3s - 5xs$. Here, $\mathscr{F}(s) = [0, 1]$ for all s. Find $f^*(s)$ and $\mathscr{F}^*(s)$ for each value of s. Using the f^* and $\mathscr{F}^*$ you have found, discuss why $f^*(s)$ is a continuous function and $\mathscr{F}^*(s)$ is a uhc correspondence. (Do not just quote a theorem.) *Hint*: If $f_x(s, x) > 0$ for all $x \in [0, 1]$, the then maximum occurs for $x = 1$. If $f_x(s, x) < 0$ for all $x \in [0, 1]$, then the maximum occurs for $x = 0$.

4.1.3. Let $f : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ be defined by $f(s, x) = (x - 1) - (x - s)^2$. Define the correspondence $\mathscr{F} : \mathbb{R}_+ \to \mathscr{P}(\mathbb{R}_+)$ by $\mathscr{F}(s) = [0, 1]$ for $s \geq 0$. Do the hypotheses of the Parametric Maximum Theorem 2 hold for this problem? Verify, through direct calculation whether the conclusions of the Parametric Maximum Theorem hold for $\mathscr{F}^*(s)$ and $f^*(s)$.

Hint: Find the critical point, x_s and verify that $\dfrac{\partial^2 f}{\partial x^2} < 0$. If $x_s \in \mathscr{F}(s)$ then it is the maximizer.

If $x_s \notin \mathscr{F}(s)$ is $\dfrac{\partial f}{\partial x}$ always positive or always negative on $[0, 1]$? Is the maximizer the right or left end point.

4.1.4. Let $f(s, x) = \sin(x) + sx$, for $s \in \mathbf{S} = [-1, 1]$ and $x \in \mathscr{F}(s) = [0, 3\,\pi]$.

a. Discuss why the Maximum Theorem applies.

b. Without finding explicit values, sketch the graph of f^* and $\mathscr{F}^*$. Discuss why these graphs look as they do and how they satisfy the conclusion of the Maximum Theorem.

Hint: Draw the graph of $f(s, x)$ as a function of x for three cases of s: (i) $s < 0$, (ii) $s = 0$, and (iii) $s > 0$.

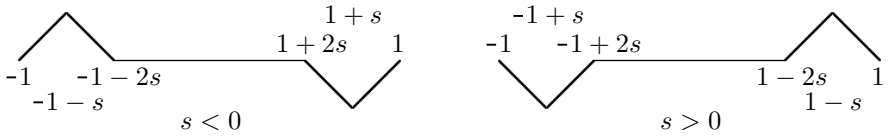
4.1.5. Let $\mathbf{S} = [0, 2]$, $\mathbf{X} = [0, 1]$, the function $f : \mathbf{S} \times \mathbf{X} \to \mathbb{R}$ be defined by $f(s, x) = -(x + s - 1)^2$, and the feasible correspondence by $\mathscr{F}(s) = [0, s]$. Find $f^*(s)$ and $\mathscr{F}^*(s)$ for each value of s. Draw the graphs of f^* and $\mathscr{F}^*$

4.1.6. Let $\mathbf{S} = [\text{-}1/2, 1/2]$ and $\mathbf{X} = \mathbb{R}$. Let the function $f : \mathbf{S} \times \mathbb{R} \to \mathbb{R}$ be defined by

$$\text{for } s < 0, \qquad f(s,x) = \begin{cases} x+1 & \text{for } \text{-}1 \le x \le \text{-}1 - s, \\ \text{-}x - 1 - 2s & \text{for } \text{-}1 - s \le x \le \text{-}1 - 2s, \\ 0 & \text{for } \text{-}1 - 2s \le x \le 1 + 2s \\ \text{-}x + 1 + 2s & \text{for } 1 + 2s \le x \le 1 + s, \\ x - 1 & \text{for } 1 + s \le x \le 1. \end{cases}$$

$$f(0,x) = 0,$$

$$\text{for } s > 0, \qquad f(s,x) = \begin{cases} \text{-}x - 1 & \text{for } \text{-}1 \le x \le \text{-}1 + s, \\ x + 1 - 2s & \text{for } \text{-}1 + s \le x \le \text{-}1 + 2s, \\ 0 & \text{for } \text{-}1 + 2s \le x \le 1 - 2s, \\ x - 1 + 2s & \text{for } 1 - 2s \le x \le 1 - s, \\ \text{-}x + 1 & \text{for } 1 - s \le x \le 1. \end{cases}$$



Let the feasible correspondence $\mathscr{F} : \mathbf{S} \to \mathscr{P}(\mathbf{X})$ be defined by

$$\mathscr{F}(s) = \begin{cases} [\text{-}1, 1 + 4s] & \text{for } s < 0, \\ [\text{-}1, 1] & \text{for } s = 0, \\ [\text{-}1 + 4s, 1] & \text{for } s > 0. \end{cases}$$

a. Sketch the graph of $\mathscr{F}$. Do f and $\mathscr{F}$ meet all the conditions of the Maximum Theorem? If yes, justify your claim. If no, list all the conditions you believe are violated and explain why you believe each of them is violated.

b. For each s, determine the value of $f^*(s)$ and the set $\mathscr{F}^*(s)$, and sketch the graphs of f^* and $\mathscr{F}^*$.
Hint: Consider $s > 0$, $s = 0$, and $s < 0$ separately. Also, you may have to split up $\mathbf{S}$ into subintervals where $\mathscr{F}(s)$ contains the point that maximizes $f(s,x)$ on $[\text{-}1, 1]$ and where it does not.

c. Is f^* continuous? Is $\mathscr{F}^*(s) \neq \emptyset$ for each s? If so, determine whether $\mathscr{F}^*$ is uhc and/or continuous on $\mathbf{S}$.

4.1.7. Let $\mathbf{S} = [0,1]$ and $\mathbf{X} = [0,2]$. Let the feasible correspondence $\mathscr{F} : \mathbf{S} \to \mathscr{P}(\mathbf{X})$ be defined by

$$\mathscr{F}(s) = \begin{cases} [0, 1 - 2s] & \text{for } s \in [0, 1/2), \\ [0, 2 - 2s] & \text{for } s \in [1/2, 1]. \end{cases}.$$

Let the function $f : \mathbf{S} \times \mathbf{X} \to \mathbb{R}$ be defined by

$$f(s,x) = \begin{cases} 0 & \text{if } s = 0, \ \ x \in [0,2], \\ \dfrac{x}{s} & \text{if } s > 0, \ \ x \in [0,s), \\ 2 - \left(\dfrac{x}{s}\right) & \text{if } s > 0, \ \ x \in [s,2s], \\ 0 & \text{if } s > 0, \ \ x \in (2s,2]. \end{cases}.$$

a. Sketch the graph of f for $s > 0$. Sketch the graph of $\mathscr{F}$. Do f and $\mathscr{F}$ meet all the conditions of the Maximum Theorem? If yes, justify your claim. If no, list all the conditions you believe are violated and explain why you believe each of them is violated. (Is f continuous at $s = 0$?)

b. For each s, determine the value of $f^*(s)$ and the set $\mathscr{F}^*(s)$, and sketch the graphs of f^* and $\mathscr{F}^*$.
Hint: You may have to split up $\mathbf{S}$ into subintervals where $\mathscr{F}(s)$ contains the point that maximizes $f(s, x)$ on $[0, 2]$ and where it does not.

c. Is f^* continuous? Is $\mathscr{F}^*(s) \neq \emptyset$ for each s? If so, determine whether $\mathscr{F}^*$ is uhc and/or continuous on $\mathbf{S}$.

4.2. Finite-Horizon Dynamic Programming

This section and the next considers maximization over discrete time periods: This section considers the case of a finite number of time periods and the next when there are infinity many periods. We start with a specific model problem of a one-sector economy rather than the general situation and use it to introduce the solution method and definitions of the key concepts.

Example 4.15 (Consumption and Savings). We consider a model of a one-sector economy where a given amount of wealth can either be consumed in this time period with an immediate reward or be invested and carried forward to the next time period. The problem is to maximize the total reward over all periods.

Fix $T \geq 1$ and consider time periods $0 \leq t \leq T$, where t is an integer.

The initial wealth $w_0 \in \mathbb{R}_+$ at period-0 is given. The wealth at period-t is derived by choices of consumption at previous time periods and is denoted by w_t. In a general situation, w_t is called the *state*.

If we have wealth $w_t \geq 0$ at time period-t, then we can choose a consumption c_t with $0 \leq c_t \leq w_t$, called the *action* at period-t. The interval $\mathscr{F}(w_t) = [0, w_t]$ is called the *feasible action correspondence*.

A *transition function* $w_{t+1} = f(w_t, c_t) = k(w_t - c_t)$ with $k \geq 1$ is given and fixed and determines the wealth at the next time period in terms of the wealth and consumption at the present time period. It can be thought of as due to production or interest on the capital.

A *utility function* $u(c) = \sqrt{c}$ gives the immediate value or payoff for the consumption at any one period. Because of a psychological factor of impatience, the period-t consumption valued back at the initial period is discounted by a factor of δ^t, where $0 < \delta \leq 1$. Thus the *period-t reward function* is $r_t(w_t, c_t) = \delta^t u(c_t) = \delta^t \sqrt{c_t}$.

Problem: Given T, u, k, δ, and w_0, maximize $\sum_{t=0}^{T} r_t(w_r, c_t) = \sum_{t=0}^{T} \delta^t u(c_t)$, the total reward over all periods, contingent on $0 \leq c_t \leq w_t$ and $w_{t+1} = k(w_t - c_t)$.

It is possible to solve this problem using the KKT approach, but this involves many variables and equations. It is easier to break up the problem into simpler problems at each time period, starting at $t = T$ and working backward. Solving the problem by considering successive time periods is the method of treating the problem as a *dynamic programming problem*.

A *Markovian strategy profile* $\boldsymbol{\sigma} = (\sigma_0, \ldots, \sigma_T)$ is a rule σ_t for each period-t of a choice $c_t = \sigma_t(w_t)$ as function of only w_t and σ_t does not depend on the other $w_{t'}$ for $t' < t$. We can pick recursively a Markovian strategy that maximizes the sum by backward induction.

(T) For $t = T$, we want to maximize $r_t(c) = \delta^T u(c) = \delta^T c^{\frac{1}{2}}$ for $0 \leq c = c_T \leq w_T$. The payoff is strictly increasing, so the choice that maximizes the payoff is $\bar{c}_T = w_T$. We denote this choice by

$$\bar{c}_T = \sigma_T^*(w_T) = w_T,$$

and it is called the *optimal strategy at period-T*. The *value function at the period-T* is the maximal payoff at the period-T and is given by

$$V_T(w_T) = r_T(\sigma_T^*(w_T)) = \delta^T w_T^{\frac{1}{2}}.$$

(T-1) Let $w = w_{T\text{-}1}$ be the wealth at period $t = T\text{-}1$. For an action c, the immediate payoff is $r_{T\text{-}1}(c) = \delta^{T\text{-}1}\, c^{\frac{1}{2}}$. The wealth carried forward to the next period is $w_T = f(w,c) = k(w-c)$, with maximal payoff at the period-T of $V_T(k(w-c)) = \delta^T k^{\frac{1}{2}}(w-c)^{\frac{1}{2}}$. Thus, the for any choice of consumption $0 \le c \le w$, the sum of immediate payoff and the optimal pay of the wealth carried forward to the period-T is

$$h_{T\text{-}1}(w,c) = \delta^{T\text{-}1}u(c) + V_T(f_{T\text{-}1}(w,c)) = \delta^{T\text{-}1}\, c^{\frac{1}{2}} + \delta^T k^{\frac{1}{2}}(w-c)^{\frac{1}{2}}.$$

To maximize $h_{T\text{-}1}$ as a function of c with $w_{T\text{-}1}$ as a parameter, we find a critical point:

$$\begin{aligned}
0 = \frac{\partial h_{T\text{-}1}}{\partial c}(w,c) &= \delta^{T\text{-}1}\tfrac{1}{2}\, c^{-\frac{1}{2}} + \delta^T k^{\frac{1}{2}}\frac{1}{2}\,(w-c)^{-\frac{1}{2}}\,(\text{-}1),\\
c^{-\frac{1}{2}} &= \delta\, k^{\frac{1}{2}}\,(w-c)^{-\frac{1}{2}}\,,\\
w - c &= \delta^{\,2} k\, c,\\
w &= (1+\delta^{\,2}k)\, c,\\
\bar{c} &= \sigma^*_{T\text{-}1}(w_{T\text{-}1}) = \frac{w_{T\text{-}1}}{1+\delta^{\,2}k}.
\end{aligned}$$

Since $\dfrac{\partial^2 h_{T\text{-}1}}{\partial c^2} < 0$ and $0 \le \bar{c} = \sigma^*_{T\text{-}1}(w_{T\text{-}1}) \le w_{T\text{-}1}$, this critical point is a maximum. Thus, the optimal strategy is the choice $\bar{c}_{T\text{-}1} = \sigma^*_{T\text{-}1}(w_{T\text{-}1}) = \dfrac{w_{T\text{-}1}}{1+\delta^{\,2}k}$. The value function at period $T-1$ is the maximal payoff for periods $T-1 \le t \le T$,

$$\begin{aligned}
V_{T\text{-}1}(w_{T\text{-}1}) = h^*_{T\text{-}1}(w_{T\text{-}1}) &= h_{T\text{-}1}(w_{T\text{-}1},\bar{c}) = \delta^{T\text{-}1}\bar{c}^{\frac{1}{2}} + \delta^T\,[k(w_{T\text{-}1}-\bar{c})]\\
&= \delta^{T\text{-}1}\bar{c}^{\frac{1}{2}} + \delta^T\left(\delta^{\,2}k^2\bar{c}\right)^{\frac{1}{2}}\\
&= \delta^{T\text{-}1}(1+\delta^{\,2}k)\,\bar{c}^{\frac{1}{2}}\\
&= \delta^{T\text{-}1}(1+\delta^{\,2}k)\,\frac{w_{T\text{-}1}^{\frac{1}{2}}}{(1+\delta^{\,2}k)^{\frac{1}{2}}}\\
&= \delta^{T\text{-}1}(1+\delta^{\,2}k)^{\frac{1}{2}}\, w_{T\text{-}1}^{\frac{1}{2}}.
\end{aligned}$$

We show by backward induction that

$$V_j(w_j) = \delta^j\left(1+\delta^{\,2}k+\cdots+\delta^{2T-2j}k^{T-j}\right)^{\frac{1}{2}} w_j^{\frac{1}{2}}.$$

We have shown that this formula for the value function is valid for $j = T$ and $T\text{-}1$. Assume it is valid for $j = t+1$, and we show the best choice of c_t at time t gives a similar expression for V_t. For a fixed $w = w_t$, we want to maximize

$$\begin{aligned}
h_t(w,c) &= r_t(w,c) + V_{t+1}(k(w-c))\\
&= \delta^{\,t}\, c^{\frac{1}{2}} + \delta^{\,t+1}\left(1+\cdots+\delta^{\,2T-2t-2}k^{T-t-1}\right)^{\frac{1}{2}} k^{\frac{1}{2}}(w-c)^{\frac{1}{2}}
\end{aligned}$$

for $0 \le c \le w$: The immediate payoff at period t is $r_t(w,c) = \delta^t c^{\frac{1}{2}}$ and the maximum

payoff for periods $t+1$ to T is $V_{t+1}(k(w-c))$. The critical point satisfies

$$\begin{aligned}
0 = \frac{\partial h_t}{\partial c} &= \delta^t \tfrac{1}{2} c^{-\frac{1}{2}} + \delta^{t+1} \left(1 + \cdots + \delta^{2T-2t-2} k^{T-t-1}\right)^{\frac{1}{2}} k^{\frac{1}{2}} \tfrac{1}{2} (w-c)^{-\frac{1}{2}} (-1), \\
c^{-\frac{1}{2}} &= \delta\, k^{\frac{1}{2}} \left(1 + \cdots + \delta^{2T-2t-2} k^{T-t-1}\right)^{\frac{1}{2}} (w-c)^{-\frac{1}{2}}, \\
w - c &= \delta^2 k \left(1 + \cdots + \delta^{2T-2t-2} k^{T-t-1}\right) c, \\
w &= \left(1 + \delta^2 k + \cdots + \delta^{2T-2t} k^{T-t}\right) c, \\
\bar{c} &= \frac{w}{1 + \cdots + \delta^{2T-2t} k^{T-t}} = \sigma_t^*(w_t) \le w.
\end{aligned}$$

Since $\dfrac{\partial^2 h_t}{\partial c^2} < 0$ and $0 \le \sigma_t^*(w_t) \le w_t$, this critical point is the maximizer. Thus, the optimal strategy is $\bar{c}_t = \sigma_t^*(w_t) = \dfrac{w_t}{1 + \cdots + \delta^{2T-2t} k^{T-t}} < w_t$. In the calculation of the maximal value, we use that

$$[k(w_t - \bar{c})]^{\frac{1}{2}} = \delta\, k \left(1 + \cdots + \delta^{2T-2t-2} k^{T-t-1}\right)^{\frac{1}{2}} \bar{c}^{\frac{1}{2}}.$$

The maximal payoff from period t onward is

$$\begin{aligned}
V_t(w_t) = h_t^*(w_t) &= \delta^t \bar{c}^{\frac{1}{2}} + \delta^{t+1} \left(1 + \cdots + \delta^{2T-2t-2} k^{T-t-1}\right)^{\frac{1}{2}} [k(w_t - \bar{c})]^{\frac{1}{2}} \\
&= \delta^t \bar{c}^{\frac{1}{2}} + \delta^{t+1} \left(1 + \cdots + \delta^{2T-2t-2} k^{T-t-1}\right)^{\frac{1}{2}} \delta\, k \left(1 + \cdots + \delta^{2T-2t-2} k^{T-t-1}\right)^{\frac{1}{2}} \bar{c}^{\frac{1}{2}} \\
&= \delta^t \bar{c}^{\frac{1}{2}} + \delta^t \left[\delta^2 k + \cdots + \delta^{2T-2t} k^{T-t}\right] \bar{c}^{\frac{1}{2}} \\
&= \delta^t \left[1 + \delta^2 k + \cdots + \delta^{2T-2t} k^{T-t}\right] \bar{c}^{\frac{1}{2}} \\
&= \frac{\delta^t \left(1 + \delta^2 k + \cdots + \delta^{2T-2t} k^{T-t}\right)}{\left(1 + \delta^2 k + \cdots + \delta^{2T-2t} k^{T-t}\right)^{\frac{1}{2}}} w_t^{\frac{1}{2}} \\
&= \delta^t (1 + \delta^2 k + \cdots + \delta^{2T-2t} k^{T-t})^{\frac{1}{2}} w_t^{\frac{1}{2}}.
\end{aligned}$$

This verifies the induction hypothesis for period t, so valid for all $T \ge t \ge 0$.

By induction, for each t with $0 \le t \le T$, the optimal strategy is

$$\bar{c}_t = \sigma_t^*(w_t) = \frac{w_t}{1 + \delta^2 k + \cdots + \delta^{2T-2t} k^{T-t}},$$

and the maximal payoff for all periods $t = 0$ to T is given by

$$V_0(w_0) = (1 + \delta^2 k + \cdots + \delta^{2T} k^T)^{\frac{1}{2}} w_0^{\frac{1}{2}}.$$

Thus, we have completely solved this problem. ■

Example 4.16. This example incorporates production to determine the feasible consumption for each period. The labor force is assumed fixed, and the production is assumed to be w_t^β for capital w_t with $0 < \beta < 1$. The consumption satisfies $0 \le c_t \le w_t^\beta$ and the transition function satisfies $w_{t+1} = f_t(w_t, c_t) = w_t^\beta - c_t$. Assume the utility is $u(c) = \ln(c)$ and the discounted reward is $r_t(w, c) = \delta^t \ln(c)$, where $0 < \delta \le 1$. We also assume that $T > 0$ is given.

($\mathbf{t = T}$) For $0 \le c \le w^\beta$, the maximum of $\delta^T \ln(c)$ occurs for $\bar{c}_T = w^\beta = \sigma_T^*(w)$, with maximal value $V_T(w) = \delta^T \beta \ln(w)$.

($\mathbf{t = T - 1}$) Let

$$h_{T-1}(w, c) = \delta^{T-1} \ln(c) + V_T(w^\beta - c) = \delta^{T-1} \ln(c) + \delta^T \beta \ln(w^\beta - c).$$

The critical point satisfies

$$\begin{aligned} 0 = \frac{\partial h_{T-1}}{\partial c} &= \frac{\delta^{T-1}}{c} - \frac{\delta^T \beta}{w^\beta - c}, \\ w^\beta - c &= \delta\,\beta\,c, \\ w^\beta &= (1+\delta\,\beta)\,c, \\ \bar{c}_{T-1} = \sigma_{T-1}(w) &= \frac{w^\beta}{1+\delta\,\beta} \leq w^\beta. \end{aligned}$$

The value function is given by the maximal value,

$$\begin{aligned} V_{T-1}(w) &= \delta^{T-1}\ln(\bar{c}) + V_t(w^\beta - \bar{c}) = \delta^{T-1}\ln(\bar{c}) + \delta^T\beta\ln(\delta\,\beta\,\bar{c}) \\ &= \delta^{T-1}\left[1+\delta\beta\right]\left[\beta\ln(w) - \ln(1+\delta\,\beta)\right] + \delta^T\beta\ln(\delta\,\beta) \\ &= \delta^{T-1}\beta\left[1+\delta\beta\right]\,\ln(w) + v_{T-1}, \end{aligned}$$

where the constant v_{T-1} includes all the terms not involving w only involving the parameters δ, β, and T.

For the induction hypothesis, assume that

$$V_j(w) = \delta^j\beta\left[1+\delta\beta+\cdots+\delta^{T-j}\beta^{T-j}\right]\ln(w) + v_j,$$

where v_j is a constant involving only the parameters δ, β, T, and j. Assume this is valid for $j = t+1$ and verify it for t. Let

$$\begin{aligned} h_t(w,c) &= \delta^t\ln(c) + V_{t+1}(w^\beta - c) \\ &= \delta^t\ln(c) + \delta^{t+1}\beta[1+\cdots+\delta^{T-t-1}\beta^{T-t-1}]\ln(w^\beta - c) + v_{t+1}. \end{aligned}$$

The critical point, $\dfrac{\partial h_t}{\partial c} = 0$, satisfies

$$\begin{aligned} 0 &= \frac{\delta^t}{c} - \delta^{t+1}\,\beta\left[1+\cdots+\delta^{T-t-1}\beta^{T-t-1}\right]\frac{1}{w^\beta - c} \\ w^\beta - c &= \left[\delta\,\beta+\cdots+\delta^{T-t}\beta^{T-t}\right]c \\ w^\beta &= \left[1+\cdots+\delta^{T-t}\beta^{T-t}\right]c \\ \bar{c}_t = \sigma_t^*(w) &= \frac{w^\beta}{1+\cdots+\delta^{T-t}\beta^{T-t}} \leq w^\beta. \end{aligned}$$

The value function is given by the maximal value,

$$\begin{aligned} V_t(w) &= \delta^t\ln(\bar{c}) + V_{t+1}(w^\beta - \bar{c}) = \delta^t\ln(\bar{c}) + V_{t+1}\left(\left[\delta\,\beta+\cdots+\delta^{T-t}\beta^{T-t}\right]\bar{c}\right) \\ &= \delta^t\ln(\bar{c}) + \delta^t\left[\delta\,\beta+\cdots+\delta^{T-t}\beta^{T-t}\right]\left[\ln(\bar{c}) + \ln\left(\delta\,\beta+\cdots+\delta^{T-t}\beta^{T-t}\right)\right] + v_{t+1} \\ &= \delta^t\left[1+\cdots+\delta^{T-t}\beta^{T-t}\right]\left[\beta\ln(w) - \ln(1+\cdots+\delta^{T-t}\beta^{T-t})\right] \\ &\quad + \delta^t\left[\delta\,\beta+\cdots+\delta^{T-t}\beta^{T-t}\right]\ln\left(\delta\,\beta+\cdots+\delta^{T-t}\beta^{T-t}\right) + v_{t+1} \\ &= \delta^t\beta\left[1+\cdots+\delta^{T-t}\beta^{T-t}\right]\,\ln(w) + v_t, \end{aligned}$$

where the constant v_{T-1} includes all the terms involving only the parameters. This proves the induction step. ■

4.2.1. Supremum and Infimum

Before discussing general finite-horizon dynamic programming, we generalize the concept of maximum and minimum. When we discuss maximizing a function on a domain, we want to use the value that is a possible maximum even before we know that it is attained. For a function $f : \mathbf{X} \to \mathbb{R}$ the *supremum* or *least upper bound* is the number M such that M is an upper bound, $f(\mathbf{x}) \leq M$ for all $\mathbf{x} \in \mathbf{X}$, and there is no upper bound that is less than M. The supremum is infinity if $f(\mathbf{x})$ is not bounded above. The supremum is denoted by $\sup\{ f(\mathbf{x}) : \mathbf{x} \in \mathbf{X} \}$. Thus, the function is bounded above if and only if it has a finite supremum. Note that the supremum is above the values of the function.

In the same way, the *infimum* or *greatest lower bound* is the number m such that m is a lower bound, $f(\mathbf{x}) \geq m$ for all $\mathbf{x} \in \mathbf{X}$, and there is no lower bound that is greater than m. The infimum is minus infinity iff $f(\mathbf{x})$ is not bounded below. The infimum is denoted by $\inf\{ f(\mathbf{x}) : \mathbf{x} \in \mathbf{X} \}$. Thus, the function is bounded below if and only if it has a finite infimum. Note that the infimum is below the values of the function.

Example 4.17. The function $\arctan(x)$ is bounded on $\mathbb{R}$ but does not attain a maximum nor a minimum. However, $\inf\{ \arctan(x) : x \in \mathbb{R} \} = -\pi/2$ and $\sup\{ \arctan(x) : x \in \mathbb{R} \} = \pi/2$ are both finite.

The function $f(x) = 1/x$ for $x \neq 0$ has $\sup\{ 1/x : x > 0 \} = \infty$, $\inf\{ 1/x : x > 0 \} = 0$, $\sup\{ 1/x : x < 0 \} = 0$, and $\inf\{ 1/x : x < 0 \} = -\infty$. ■

4.2.2. General Theorems

Definition. A *finite-horizon dynamic programming problem*, FHDP, is specified as follows.

FH1. T is a positive integer, the *horizon*. The periods t are integers with $0 \leq t \leq T$.

FH2. $\mathbf{S}$ is the *state space* , with the *state* at period t given by $s_t \in \mathbf{S}$.
(In the C-S problem, $\mathbf{S} = [0, \infty)$ and $s_t = w_t \in [0, \infty)$.)

FH3. $\mathbf{A}$ is the action space, with the action at period t given by $a_t \in \mathbf{A}$.
(In the C-S problem, $a_t = c_t \in [0, \infty) = \mathbf{A}$.)

FH4. For each integer $0 \leq t \leq T$, two functions and a correspondence are given as follows.

- **i.** $r_t : \mathbf{S} \times \mathbf{A} \to \mathbb{R}$ is the continuous period-t *reward function*.
  (In the C-S problem, $r_t(w_t, c_t) = \delta^t c_t^{\frac{1}{2}}$.)
- **ii.** $f_t : \mathbf{S} \times \mathbf{A} \to \mathbf{S}$ is the continuous period-t *transition function*, $s_{t+1} = f_t(s_t, a_t)$.
  (In the C-S problem, $f_t(w_t, c_t) = k(w_t - c_t)$.)
- **iii.** $\mathscr{F}_t : \mathbf{S} \to \mathscr{P}(\mathbf{A})$ is the *feasible action correspondence*, and is assumed to be a continuous and compact-valued correspondence. Only $a_t \in \mathscr{F}_t(s_t)$ are allowed.
  (In the C-S problem, $c_t \in [0, w_t] = \mathscr{F}_t(w_t)$.)

The *total reward* for initial state s_0 and allowable actions $\{a_t\}_{t=0}^T$, with $a_t \in \mathscr{F}_t(s_t)$ and $s_{t+1} = f_t(s_t, a_t)$ is

$$W\left(s_0, \{a_t\}_{t=0}^T\right) = \sum_{t=0}^{T} r_t(s_t, a_t).$$

The *value function* of the continuation FHDP starting with state s_t at period t is defined as

$$\begin{aligned} V_t(s_t) &= \sup\left\{ \textstyle\sum_{j=t}^T r_j(s_j, a_j) : a_j \in \mathscr{F}_j(s_j),\ s_{j+1} = f_i(s_j, a_j) \text{ for } j = t, \ldots, T \right\} \\ &= \sup\left\{ W\left(s_t, \{a_j\}_{j=t}^T\right) : \{a_j\}_{j=t}^T \text{ is allowable} \right\}, \end{aligned}$$

and $V(s_0) = V_0(s_0)$ is the value function for the whole FHDP. The value function is the maximal payoff for any choice of allowable actions.

We show that the total reward for various choice of actions does in fact attain a finite maximal value. The problem is to find this maximal value and actions that realize this maximum.

Definition. A *Markovian strategy profile* is a collection of (choice) functions $\boldsymbol{\sigma} = (\sigma_0, \ldots, \sigma_T)$ with $\sigma_t : \mathbf{S} \to \mathbf{A}$ so $a_t = \sigma_t(s_t) \in \boldsymbol{\mathcal{F}}_t(s_t)$ for $0 \le t \le T$. So, each σ_t is a function of only s_t. For a *non-Markovian strategy profile*, each σ_t can be a function of $(s_0, \ldots, s_t)$ and not just s_t. For a Markovian strategy profile $\boldsymbol{\sigma}$ and initial state s_0, the actions and states at all periods are determined by induction as follows: $s_0(s_0, \boldsymbol{\sigma}) = s_0$; for $0 \le t \le T$, given $s_t = s_t(s_0, \boldsymbol{\sigma})$,

$$\begin{aligned} a_t = a_t(s_0, \boldsymbol{\sigma}) &= \sigma_t(s_t), \\ r_t(s_0, \boldsymbol{\sigma}) = r_t(s_t, a_t), &\qquad \text{and} \\ s_{t+1} = s_{t+1}(s_0, \boldsymbol{\sigma}) &= f_t(s_t, a_t). \end{aligned}$$

The *total reward* for a strategy profile $\boldsymbol{\sigma}$ and initial state s_0 is given by

$$W(s_0, \boldsymbol{\sigma}) = \sum\nolimits_{t=0}^{T} r_t(s_0, \boldsymbol{\sigma}).$$

A strategy profile $\boldsymbol{\sigma}^*$ is called an *optimal strategy profile* provided that $W(s_0, \boldsymbol{\sigma}^*) = V(s_0)$ for all $s_0 \in \mathbf{S}$, i.e., it attains the maximal value of the value function.

Theorem 4.18 (FHDP Bellman Equation and Optimal Strategy). *If a FHDP satisfies FH1 – FH4, then the following hold.*

a. *For $0 \le t \le T$, V_t attains a finite maximal value $V_t(s_t) < \infty$ for each $s_t \in \mathbf{S}$, is continuous, and satisfies*

$$V_t(s) = \max\{\, r_t(s, a) + V_{t+1}(f_t(s, a)) : a \in \boldsymbol{\mathcal{F}}_t(s) \,\}. \tag{10}$$

We take $V_{T+1}(f_T(s, a)) \equiv 0$, so the equation for $t = T$ becomes $V_T(x) = \max\{\, r_T(s, a) : a \in \boldsymbol{\mathcal{F}}_T(s) \,\}$.

b. *There exists a Markovian optimal strategy profile $\boldsymbol{\sigma}^* = (\sigma_0^*, \ldots, \sigma_T^*)$ such that $W(s_0, \boldsymbol{\sigma}^*) = V(s_0)$ for all s_0.*

Remark. Equation (10) is call the *Bellman equation* and determines the solution method for a FHDP. First, the strategy $a_T^* = \sigma_T^*(s_T)$ is determined that maximizes $r_T(s_T, a_T)$. This action determines the value function $V_T(s_T) = r_T(s_T, \sigma_T^*(s_T))$. By backward induction, once the strategies σ_j^* and value functions V_j have been determined for $T \ge j \ge t+1$, then the strategy $a_t^* = \sigma_t^*(s_t)$ is determined that maximizes $h_t(s_t, a_t) = r_t(s_t, a_t) + V_{t+1}(f_t(s_t, a_t))$ and the next value function is set equal to the maximal value $V_t(s_t) = h_t^*(s_t) = r_t(s_t, \sigma_t^*(s_t)) + V_{t+1}(f_t(s_t, \sigma_t^*(s_t)))$. By induction, we get back to $V(s_0) = V_0(s_0)$.

Proof. We prove the theorem by backward induction, starting at $t = T$ and going down to $t = 0$.

For $t = T$, there is only one term in the definition of the value function and $V_T(s_T) = \max\{\, r_T(s, a) : a \in \boldsymbol{\mathcal{F}}_T(s) \,\}$. The function r_T is continuous and $\boldsymbol{\mathcal{F}}_T$ is continuous and compact-valued, so by the Parametric Maximization Theorem,

$$V_T(s_T) = \max\{r_T(s_T, a) : a \in \boldsymbol{\mathcal{F}}_T(s_T)\,\} < \infty$$

exists for each $s_T \in \mathbf{S}$ and is continuous, and $\boldsymbol{\mathcal{F}}_T^*(s_T) = \{\, a \in \boldsymbol{\mathcal{F}}_T(s_T) : f(s_t, a) = V_T(s_T)\,\}$ is nonempty upper-hemicontinuous correspondence. Pick $\sigma_T^*(s_T) \in \boldsymbol{\mathcal{F}}_T^*(s_T)$ for each $s_T \in \mathbf{S}$, $\sigma_T^* : \mathbf{S_T} \to \mathbf{A}$. Then, $r_T(s_T, \sigma_T^*(s_T)) = W(s, \sigma_T^*) = V_T(s)$, so σ_T^* is an optimal strategy. This proves the result for $t = T$.

The following lemma is the induction step.

Lemma 4.19 (Induction Step). *For $0 \le t < T$, suppose that the value function V_{t+1} is a continuous function of s_{t+1} and takes a finite value for each s_{t+1} and that the continuation FHDP starting a period $t+1$ admits a Markovian optimal strategy profile $(\sigma^*_{t+1}, \ldots, \sigma^*_T)$, so that $V_{t+1}(s_{t+1}) = W(s_{t+1}, (\sigma^*_{t+1}, \ldots, \sigma^*_T))$. Then the following hold.*

- **a.** *For each $s_t \in \mathbf{S}$, the value function $V_t(s_t)$ attains a finite maximal value, is continuous, and satisfies*
$$V_t(s_t) = \max\{\, r_t(s_t, a_t) + V_{t+1}(f_t(s_t, a_t)) : a_t \in \boldsymbol{\mathscr{F}}_t(s_t) \,\}.$$
- **b.** *There exists a strategy σ^*_t, such that $(\sigma^*_t, \ldots, \sigma^*_T)$ is a Markovian optimal strategy profile for the continuation FHDP starting at period t, $W(s_t, (\sigma^*_t, \ldots, \sigma^*_T)) = V_t(s_t)$ for all s_t.*

Proof. We start by considering the right hand side of the Bellman equation $h_t(s_t, a_t) = r_t(s_t, a_t) + V_{t+1}(f_t(s_t, a_t))$. Since f_t and r_t are continuous by assumptions of the theorem and V_{t+1} is continuous by the induction assumption of the lemma, $h_t(s_t, a_t)$ is continuous. The set correspondence $\boldsymbol{\mathscr{F}}_t$ is continuous and compact-valued. By the Parametric Maximization Theorem, the maximal value $h^*_t(s_t)$ is continuous and set of points that realized the maximum $\boldsymbol{\mathscr{F}}^*(s_t)$ is a nonempty set. If $\sigma^*_t(s_t)$ is any selection of a point in $\boldsymbol{\mathscr{F}}^*(s_t)$, then $h(s_t, \sigma^*_t(s_t)) = h^*_t(s_t)$ is a Markovian strategy that we show satisfies the lemma.

For any s_t and any allowable sequence with $a_i \in \boldsymbol{\mathscr{F}}(s_i)$ and $s_{i+1} = f_i(s_i, a_i)$ for $i \ge t$,

$$\begin{aligned}
\sum_{i=t}^{T} r_i(s_i, a_i) &= r_t(s_t, a_t) + \sum_{i=t+1}^{T} r_i(s_i, a_i) \\
&\le r_t(s_t, a_t) + \max \Bigg\{ \sum\nolimits_{i=t+1}^{T} r_i(s'_i, a'_i) : s'_{t+1} = s_{t+1},\ a'_i \in \boldsymbol{\mathscr{F}}_i(s'_i), \\
&\qquad\qquad s'_{i+1} = f_i(s'_i, a'_i) \text{ for } i \ge t+1 \Bigg\} \\
&= r_t(s_t, a_t) + V_{t+1}(f(s_t, a_t)) = h_t(s_t, a_t) \\
&\le \max\{\, h_t(s_t, a'_t) : a'_t \in \boldsymbol{\mathscr{F}}_t(s_t) \,\} \\
&= h^*_t(s_t).
\end{aligned}$$

Taking the supremum over all allowable choices yields

$$\begin{aligned}
V_t(s_t) &= \sup \left\{ \sum_{i=t}^{T} r_i(s_i, a_i) : a_i \in \boldsymbol{\mathscr{F}}_t(s_t),\ s_{i+1} = f_i(s_i, a_i) \text{ for } t \le i < T \right\} \\
&\le h^*_t(s_t) < \infty.
\end{aligned}$$

For the other inequality,

$$\begin{aligned}
h^*_t(s_t) &= h(s_t, \sigma^*_t(s_t)) \\
&= r_t(s_t, \sigma^*_t(s_t)) + V_{t+1}(f_t(s_t, \sigma^*_t(s_t)) \\
&= r_t(s_t, \sigma^*_t(s_t)) + \sum_{i=t+1}^{T} r_i(s^*_i, \sigma^*_i(s^*_i)) \le V_t(s_t),
\end{aligned}$$

where $s^*_t = s_t$ and $s^*_{i+1} = f_i(s^*_i, \sigma^*(s^*_i))$. Combining the two inequalities, $V_t(s_t) = h^*_t(s_t)$. So V_t in finite, continuous, and satisfies the Bellman equation. By the induction hypothesis,

$$\begin{aligned}
V_t(s_t) &= r_t(s_t, \sigma^*_t(s_t)) + V_{t+1}(f_t(s_t, \sigma^*_t(s_t)) \\
&= r_t(s_t, \sigma^*_t(s_t)) + W\left(f_t(s_t, \sigma^*_t(s_t), (\sigma^*_{t+1}, \ldots, \sigma^*_T)\right) \\
&= W\left(s_t, (\sigma^*_t, \ldots, \sigma^*_T)\right),
\end{aligned}$$

and we have found an optimal Markovian strategy profile as claimed. □

By induction, we have found a strategy $\boldsymbol{\sigma}^* = (\sigma_0^*, \ldots, \sigma_T^*)$ that satisfies the Bellman equation and $W(s_0, \boldsymbol{\sigma}^*) = V_0(s_0)$. Thus, $\boldsymbol{\sigma}^*$ is an optimal strategy. □

4.2. Exercises

4.2.1. Consider the Consumption-Savings FHDP with $\delta = 1$, $r_t(w, c) = c^{\frac{1}{3}}$, transition function $f_t(c, w) = (w - c)$, $\boldsymbol{\mathscr{F}}_t(w_t) = [0, w_t]$, and $T = 2$. Find the value functions and optimal strategy for each stage.

4.2.2. Consider the Consumption-Savings FHDP with $T > 0$, $r_t(w, c) = \ln(c)$ ($\delta = 1$), transition function $f_t(w, c) = w - c$, and $\boldsymbol{\mathscr{F}}_t(w) = [0, w]$ for all periods. Find the value functions and optimal strategy for each stage. *Remark*: The reward function equals minus infinity for $c = 0$, but this just means that it is very undesirable.
Hint: Compute, $V_T(w_T)$ and $V_{T-1}(w_{T-1})$. Then guess the form of V_j, and prove it is valid by induction.

4.2.3. Consider the Consumption-Savings FHDP with $T > 0$, $r(w, c) = 1 - e^{-c}$, transition function $f_t(w_t, c) = w_t - c$, and $\boldsymbol{\mathscr{F}}_t(w_t) = [0, w_t]$. Find the value functions and optimal strategy for each stage.

4.2.4. Consider the FHDP with $\delta = 1$, $r_t(s, c) = 1 - \dfrac{1}{1 + c}$, transition function $f_t(s, c) = (s - c)$, $\boldsymbol{\mathscr{F}}_t(s_t) = [0, s_t]$, and $T \geq 2$.

a. Find the value function and optimal strategy for $t = T$ and $T - 1$.

b. Using backward induction, verify that $V_t(s) = 1 + t - \dfrac{(1 + t)^2}{1 + t + s}$. Also, determine the optimal strategy for each t.

4.2.5. Consider the Consumption-Savings FHDP with $T > 0$, $r_t(w, c) = \delta^t \ln(c)$ with $0 < \delta \leq 1$, transition function $f_t(w, c) = A\, w^\beta - c$ with $A > 0$ and $\beta > 0$, and $\boldsymbol{\mathscr{F}}_t(w) = [0, A\, w^\beta]$ for all periods. Verify that the value function is $V_j(w) = \delta^j \ln(w)\, \beta(1 + \beta\delta + \cdots + \beta^{T-j}\delta^{T-j}) + v_j$ for correctly chosen constants v_j (that can depend on δ, β, and other parameters). Also find the optimal strategy for each stage.
Remark: The reward function equals minus infinity for $c = 0$, but this just means that small values of c are very undesirable.

4.3. Infinite-Horizon Dynamic Program

In this section, we discuss problems with an infinite horizon where the process can go on all future periods, and there is an infinite sum of rewards. In order for the total reward to be finite, we need to discount the reward at time t in the future by a factor δ^t, where $0 < \delta < 1$. Also, there is no final time period at which to start a backward induction to find a value function. However, starting the process one stage later forms an equivalent dynamic program, so we can show that the value function satisfies a type of Bellman equation. Since we can show there is a unique function satisfying the Bellman equation, the solution method is to solve this functional equation for what must be the value function. A more thorough treatment of stationary dynamic programming is given in [**12**] by Stokey and Lucas and [**14**] by Sundaram.

Definition. A *stationary dynamic programming problem* with infinite horizon, SDP, is specified as follows.

SD1. $\mathbf{S} \subset \mathbb{R}^n$ is the *state space* with s_t the state at period-t.

SD2. $\mathbf{A} \subset \mathbb{R}^k$ is the *action space* with the *action* a_t at period-t.

SD3. There is a *feasible action correspondence* $\mathscr{F} : \mathbf{S} \to \mathscr{P}(\mathbf{A})$ that is a compact-valued, nonempty, continuous correspondence. For each $s \in \mathbf{S}$, the set $\mathscr{F}(s) \subset \mathbf{A}$ specifies the allowable actions.

SD4. There is a continuous *transition function* $f : \mathbf{S} \times \mathbf{A} \to \mathbf{S}$ that specifies the state at the next period in terms of the current state and action taken, $s_{t+1} = f(s_t, a_t)$ for $t \geq 0$.

SD5. There is a continuous one-period *reward function* $r : \mathbf{S} \times \mathbf{A} \to \mathbb{R}$ that specifies an immediate reward $r(s, a)$ for an action a taken at state s.

SD6. There is a *discount factor* $\delta \in (0, 1)$, so that $\delta^t r(s_t, a_t)$ is the reward at period-t discounted back to period-0. This psychological factor represents the impatience for the reward.

The *total reward* for an allowable sequence of actions $a_t \in \mathscr{F}(s_t)$ and states $s_{t+1} = f(s_t, a_t)$ for $t \geq 0$ is

$$W\left(s_0, \{a_t\}_{t=0}^{\infty}\right) = \sum_{t=0}^{\infty} \delta^t r(s_t, a_t).$$

Remark. The dynamic program is called *stationary* because the same r, f, and $\mathscr{F}$ are used for every period t. The discount factor allows the possibility that the total reward is finite.

Definition. The *value function* $V : \mathbf{S} \to \mathbb{R}$ is defined as the supremum of the total reward over all possible sequences of allowable actions and states,

$$V(s_0) = \sup\left\{ W\left(s_0, \{a_t\}\right) : \{a_t\} \text{ is an allowable sequence} \right\}.$$

Problem: The problem is to find an allowable sequence of actions that realizes the value function as a maximum, i.e., that maximizes the total reward $W(s_0, \{a_t\})$ for allowable sequences of actions $\{a_t\}$ and $s_{t+1} = f(s_t, a_t)$ $t \geq 0$.

Definition. A *stationary strategy* σ is a choice $\sigma(s) \in \mathscr{F}(s) \subset \mathbf{A}$ for each $s \in \mathbf{S}$ that is the same for all periods.

Definition. For a SDP, given a stationary strategy σ and an initial state s_0, we can determine the actions and the states at all periods by induction, and so the *total reward*: $a_t = a_t(s_0, \sigma) = \sigma(s_t)$, $s_{t+1} = s_{t+1}(s_0, \sigma) = f(s_t, a_t)$, and

$$W(s_0, \sigma) = \sum_{t=0}^{\infty} \delta^t r(s_t, a_t).$$

Definition. An *optimal stationary strategy* σ^* is a stationary strategy such that

$$W(s_0, \sigma^*) = V(s_0) \qquad \text{for all } s_0 \in \mathbf{S}.$$

Theorem 4.20 (SDP Bellman Equation). *For a SDP, the value function $V(s)$ satisfies the following equation, called the* Bellman equation*:*

$$V(s) = \sup\{\, r(s, a) + \delta\, V(f(s, a)) : a \in \mathscr{F}(s) \,\}. \tag{11}$$

Remark. Note that for a SDP, the same function V is on both sides of Bellman equation. Thus, it is not possible to just find the maximum value of the right hand side as is done for finite-horizon dynamic programming. Instead, it is necessary to solve the equation for a value function that is the same on both sides of the equation.

Proof. Define $h(s,a) = r(s,a) + \delta\, V(f(s,a))$ to be the function on the right-hand side of Bellman's equation. To show that the Bellman equation holds, we first show that $V(s)$ is less than or equal to the right-hand side of the Bellman equation. Fix an s_0, and take any allowable sequence $a_t \in \boldsymbol{\mathscr{F}}(s_t)$ and $s_{t+1} = f_t(s_t, a_t)$ for $t \geq 0$.

$$\begin{aligned}\sum_{t=0}^{\infty} \delta^{\,t} r(s_t,a_t) &= r(s_0,a_0) + \delta \sum_{t=1}^{\infty} \delta^{\,t-1} r(s_t,a_t) \\ &\leq r(s_0,a_0) + \delta \sup \Big\{ \textstyle\sum_{t=0}^{\infty} \delta^{\,t-1} r(s'_{t+1},a'_{t+1}) : s'_1 = s_1,\ a'_t \in \boldsymbol{\mathscr{F}}(s'_t), \\ &\qquad\qquad s'_t = f_t(s'_{t-1},a'_{t-1}) \ \text{ for } t \geq 0 \Big\} \\ &= r(s_0,a_0) + \delta\, V(f(s_0,a_0)) \\ &\leq \sup \{\, h(s_0,a'_0) : a'_0 \in \boldsymbol{\mathscr{F}}(s_0) \,\} = h^*(s_0).\end{aligned}$$

Here we define h^* using a supremum like we did earlier for the maximum. Since the total reward for any allowable sequence is less than or equal to $h^*(s_0^0)$, taking the supremum over all allowable choices yields

$$V(s_0) = \sup \left\{ \sum_{t=0}^{\infty} \delta^{\,t}\, r(s_t,a_t) : a_t \in \boldsymbol{\mathscr{F}}(s_t),\ s_{t+1} = f(s_t,a_t) \text{ for } t \geq 0 \right\} \leq h^*(s_0).$$

We will be done if we can show that $V(s_0) \geq h^*(s_0)$.

First, assume that $h^*(s_0) < \infty$. We show that $V(s_0^0) \geq h^*(s_0^0) - \epsilon$ for any $\epsilon > 0$. Fix $\epsilon > 0$. Since $h^*(s_0) - \frac{\epsilon}{2}$ is not an upper bound, there exists an $a'_0 \in \boldsymbol{\mathscr{F}}(s_0)$ such that

$$r(s_0,a'_0) + \delta\, V\left(f(s_0,a'_0)\right) = h(s_0,a'_0) \geq h^*(s_0) - \frac{\epsilon}{2}.$$

Then starting at $s'_1 = f(s_0,a'_0)$, there exist $a'_t \in \boldsymbol{\mathscr{F}}(s'_t)$ and $s'_{t+1} = f(s'_t,a'_t)$ for $t \geq 1$ such that

$$\textstyle\sum_{t=1}^{\infty} \delta^{\,t-1}\, r(s'_t,a'_t) \geq V(s'_1) - \frac{\epsilon}{2}.$$

Combining,

$$\begin{aligned}W(s_0,\{a'_t\}) = r(s_0,a'_0) + \textstyle\sum_{t=1}^{\infty} \delta^{\,t} r(s'_t,a'_t) &= r(s_0,a'_0) + \delta \textstyle\sum_{t=1}^{\infty} \delta^{\,t-1}\, r(s'_t,a'_t) \\ &\geq r_\tau(s_0,a'_0) + \delta\, V(s'_1) - \frac{\epsilon}{2} \\ &\geq h^*_\tau(s_\tau) - \frac{\epsilon}{2} - \frac{\epsilon}{2}.\end{aligned}$$

The supremum over all allowable sequences starting with s_0 is at least as large as the payoff using the above sequence of choices a'_t, so

$$V(s_0) \geq h^*(s_0) - \epsilon.$$

Since $\epsilon > 0$ is arbitrary, $V(s_0) \geq h^*(s_0)$. Combining the two directions, $V(s_0) = h^*(s_0)$ and the Bellman equation is satisfied.

Next, assume that $h^*(s_0) = \infty$. For any $K > 0$, there exists an $a'_0 \in \boldsymbol{\mathscr{F}}(s_0)$ such that

$$r(s_0,a'_0) + \delta\, V\left(f(s_0,a'_0)\right) \geq K.$$

Then starting at $s'_1 = f(s_0,a'_0)$, there exist $a'_t \in \boldsymbol{\mathscr{F}}(s'_t)$ and $s'_{t+1} = f(s'_t,a'_t)$ for $t \geq 1$ such that

$$r(s_0,a'_0) + \delta \textstyle\sum_{t=1}^{\infty} \delta^{\,t-1}\, r(s'_t,a'_t) \geq K - 1.$$

The supremum over all allowable sequences starting with s_0 is at least as large as the payoff using the above sequence of choices a'_t, so

$$V(s_0) \geq \textstyle\sum_{t=0}^{\infty} \delta^{\,t}\, r(s'_t,a'_t) \geq K - 1.$$

Since $K > 0$ is arbitrary, $V(s_0) = \infty$, and $V(s_0) = h^*(s_0^0)$. □

Properties of Value Function and Existence of Optimal Stationary Strategy

We prove the following results under one of two sets of assumptions: (i) the reward function is bounded (SDB), or (ii) SDP is an Optimal Growth Dynamical Program for a one-sector economy satisfying assumptions E1 – E3 that are given when the general model is discussed.

Finite value function. For these two contexts, Theorems 4.23 and 4.26(a) show that $V(s) < \infty$ for each $s \in \mathbf{S}$, so $V(s)$ is a well defined function.

Continuity. Theorems 4.24 and 4.26(b) prove that $V(s)$ is continuous and the unique bounded function satisfying Bellman equation. The proof uses an iterative process to construct a sequence of continuous $V_j(s)$ that converge uniformly to $V(s)$ on compact intervals $[0, \bar{s}]$, so $V(x)$ is continuous.

Optimal Strategy. Once we know that $V(s)$ is continuous, then the right hand side of the Bellman equation is continuous,

$$h(s, a) = r(s, a) + \delta\, V \circ f(s, a).$$

Theorems 4.25 and 4.27(b) prove that any choice function

$$\sigma^*(s) \in \boldsymbol{\mathscr{F}}^*(s) = \{\, a \in \boldsymbol{\mathscr{F}}(s) : h(s, a) = h^*(s) \,\}$$

is an optimal strategy, so an optimal stationary strategy σ^* exists with $W(s, \sigma^*) = V(s)$.

4.3.1. Examples

We delay the precise theorems and proofs until after giving examples using two methods of using the Bellman equation to determine the value function and an optimal strategy. The first method involves construction a sequence of functions by iteratively maximizing the right-hand of the Bellman equation. The proof of Theorem 4.24 the sequence of function converge to the true value function. The second method involves guessing the form of the value function in a form that involves unknown parameters. Then the Bellman equation is used to determine the value of these parameters and so of the true value function.

Example 4.21. This example is a special case of the optimal growth of a one-sector economy considered later. Let

$$\begin{aligned} \mathbf{S} &= \mathbb{R}_+, \\ \mathbf{A} &= \mathbb{R}_+, \\ \boldsymbol{\mathscr{F}}(s) &= [0, s], \\ f(s, a) &= k(s - a), \qquad \text{with } k \geq 1 \\ r(s, a) &= u(a) = a^{\frac{1}{2}}, \qquad 1.0 in \text{and} \\ 0 < \delta < 1, &\qquad \text{with } k\,\delta^2 < 1. \end{aligned}$$

The reward function is not bounded on $\mathbb{R}_+$, but given an initial state s_0, $s_1 \leq k s_0$, $s_2 \leq k\, s_1 \leq k^2\, s_0$, $s_t \leq k^t\, s_0$. If $k\,\delta^2 < 1$, the total reward is bounded as follows:

$$\delta^t\, r(s_t, a_t) \leq \delta^t\, u(s_t) \leq \delta^t\, u(k^t\, s_0) = \left(\delta\, k^{\frac{1}{2}}\right)^t\, s_0^{\frac{1}{2}},$$

$$\sum_{t=0}^{\infty} \delta^t\, r(s_t, a_t) \leq \sum_{t=0}^{\infty} \delta^t k^{\frac{t}{2}}\, s_0^{\frac{1}{2}} \leq \frac{s_0^{\frac{1}{2}}}{1 - \delta k^{\frac{1}{2}}}.$$

Thus, the value function is finite for each s_0.

Solution Method 1: We construct a sequence of functions $V_j(s)$, that converge to the value function $V(s)$. We prove that the $V_j(s)$ are continuous by induction. Start with the zero function $V_0(s) = 0$ for all s. Assume that $V_j(s)$ in our sequence is continuous, and let

$$h_{j+1}(s,a) = r(s,a) + \delta\, V_j\left(f(s,a)\right) = a^{\frac{1}{2}} + \delta\, V_j\left(k(s-a)\right)$$

be the function that is used in the supremum on the right hand-side of the Bellman equation using the value function V_j. The functions $r(s,a) = u(a) = a^{\frac{1}{2}}$, $V_j(s)$, and $f(s,a) = k\,(s-a)$ are all continuous, so $h_{j+1}(s,a)$ is a continuous function. The feasible action correspondence $\mathscr{F}(s) = [0,s]$ is a continuous compact-valued correspondence. So, we can apply the Parametric Maximization Theorem to get a continuous maximal value as a function of s,

$$V_{j+1}(s) = h_1^*(s).$$

None of these functions are the value function, because we will see that $V_{j+1}(s) \neq V_j(s)$. However in this example, we note that this sequence of functions converges to the true value function. Later in the Proof of Theorem 4.24 proving the continuity of the value function, we indicate a little more of why this works.

Since we start with $V_0(s) = 0$ for all s,

$$h_1(s,a) = a^{\frac{1}{2}} + \delta\, V_0\left(k(s-k)\right) = a^{\frac{1}{2}}.$$

This function is increasing on the intervals $\mathscr{F}(s) = [0,s]$, so $h_1(s,a)$ has a maximum at the right end point, $\bar{a} = s$. Then,

$$V_1(s) = h_1^*(x) = h_1(s,\bar{a}) = s^{\frac{1}{2}}.$$

Note that $V_1(s)$ is the maximum over the single period $t = 0$.

For the next step, $h_2(s,a) = a^{\frac{1}{2}} + \delta\, V_1\left(k(s-a)\right) = a^{\frac{1}{2}} + \delta\, k^{\frac{1}{2}}(s-a)^{\frac{1}{2}}$. The critical point satisfies

$$\begin{aligned}
0 = \frac{\partial h_2}{\partial a} &= \tfrac{1}{2}a^{-\frac{1}{2}} - \tfrac{1}{2}\delta\, k^{\frac{1}{2}}(s-a)^{-\frac{1}{2}},\\
a^{-\frac{1}{2}} &= k^{\frac{1}{2}}\,\delta\,(s-a)^{-\frac{1}{2}},\\
s - a &= k\,\delta^{\,2}\,a,\\
s &= (1 + k\,\delta^{\,2})\,a,\\
\bar{a} &= \frac{s}{1 + k\,\delta^{\,2}}.
\end{aligned}$$

Since $\dfrac{\partial^2 h_2}{\partial a^2} < 0$ for all $a \in [0,s]$, $\bar{a}$ is a maximizer.

$$\begin{aligned}
V_2(s) = h_2^*(s) = h_2(s,\bar{a}) &= \bar{a}^{\frac{1}{2}} + \delta\, k^{\frac{1}{2}}(s-\bar{a})^{\frac{1}{2}}\\
&= \bar{a}^{\frac{1}{2}} + \delta\, k^{\frac{1}{2}}\delta\, k^{\frac{1}{2}}\,\bar{a}^{\frac{1}{2}}\\
&= (1 + k\,\delta^{\,2})\,\bar{a}^{\frac{1}{2}}\\
&= (1 + k\,\delta^{\,2})\,\frac{s^{\frac{1}{2}}}{(1 + k\,\delta^{\,2})^{\frac{1}{2}}}\\
&= (1 + k\,\delta^{\,2})^{\frac{1}{2}}\,s^{\frac{1}{2}}.
\end{aligned}$$

Note that $V_2(s)$ is the maximum over the two periods $t = 0, 1$.

Our induction hypothesis is that

$$V_j(s) = \left[1 + k\,\delta^{\,2} + \cdots + k^{j-1}\delta^{\,2(j-1)}\right]^{\frac{1}{2}}\; s^{\frac{1}{2}}.$$

We have verified the formula for $j = 1, 2$. Assume this is true for $j = t$ and show is is true for $j = t + 1$. Let

$$h_{t+1}(s,a) = a^{\frac{1}{2}} + \delta \left[1 + k\,\delta^{\,2} + \cdots + k^{t-1}\delta^{\,2(t-1)}\right]^{\frac{1}{2}} k^{\frac{1}{2}}\,(s-a)^{\frac{1}{2}}.$$

The critical point satisfies

$$\begin{aligned}
0 = \frac{\partial h_{t+1}}{\partial a} &= \frac{1}{2}\,a^{-\frac{1}{2}} - \frac{1}{2}\,k^{\frac{1}{2}}\,\delta\,\left[1 + \cdots + k^{t-1}\,\delta^{\,2(t-1)}\right]^{\frac{1}{2}}\,(s-a)^{-\frac{1}{2}} \\
(s-a)^{\frac{1}{2}} &= k^{\frac{1}{2}}\,\delta\,\left[1 + \cdots + k^{t-1}\,\delta^{\,2(t-1)}\right]^{\frac{1}{2}}\,a^{\frac{1}{2}} \\
s - a &= \left[k\,\delta^{\,2} + \cdots + k^{t}\,\delta^{\,2t}\right]\,a \\
s &= \left[1 + k\,\delta^{\,2} + \cdots + k^{t}\,\delta^{\,2t}\right]\,a \\
\bar{a} &= \frac{s}{1 + k\,\delta^{\,2} + \cdots + k^{t}\,\delta^{\,2t}} = \sigma^{*}_{t+1}(s).
\end{aligned}$$

Again, this is a maximizer because $\dfrac{\partial^2 h_{t+1}}{\partial a^2} < 0$ for all $a \in [0, s]$. Then

$$\begin{aligned}
V_{t+1}(s) = h^{*}_{t+1}(s) = h_{t+1}(s,\bar{a}) &= \bar{a}^{\frac{1}{2}} + \delta\,\left[1 + k\,\delta^{\,2} + \cdots + k^{t-1}\,\delta^{\,2(t-1)}\right]^{\frac{1}{2}}\,k^{\frac{1}{2}}\,(s-\bar{a})^{\frac{1}{2}} \\
&= \bar{a}^{\frac{1}{2}} + \delta\,\left[1 + k\,\delta^{\,2} + \cdots + k^{t-1}\,\delta^{\,2(t-1)}\right]^{\frac{1}{2}}\,k^{\frac{1}{2}}\,\left[k\,\delta^{\,2}\,\bar{a}\right]^{\frac{1}{2}} \\
&= \left[1 + k\,\delta^{\,2} + \cdots + k^{t}\,\delta^{\,2t}\right]^{\frac{1}{2}}\,\bar{a}^{\frac{1}{2}} \\
&= \left[1 + k\,\delta^{\,2} + \cdots + k^{t}\,\delta^{\,2t}\right]^{\frac{1}{2}}\left[1 + k\,\delta^{\,2} + \cdots + k^{t}\,\delta^{\,2t}\right]^{-1} s \\
&= \left[1 + k\,\delta^{\,2} + \cdots +, k^{t}\,\delta^{\,2t}\right]^{\frac{1}{2}}\,s^{\frac{1}{2}}
\end{aligned}$$

This verifies the induction hypothesis for the form of $V_{t+1}(s)$.

Taking the limit as t goes to infinity,

$$\begin{aligned}
V_{\infty}(s) = \lim_{t\to\infty} V_t(s) &= \lim_{t\to\infty}\left[1 + k\,\delta^{\,2} + \cdots + k^{t-1}\,\delta^{\,2(t-1)}\right]^{\frac{1}{2}}\,s^{\frac{1}{2}} \\
&= \left[1 - \delta^{\,2}k\right]^{-\frac{1}{2}}\,s^{\frac{1}{2}}.
\end{aligned}$$

Also, if we take the limit in the inductive equation defining the $V_j(s)$,

$$\begin{aligned}
V_{\infty}(s) = \lim_{t\to\infty} V_{j+1}(s) &= \lim_{t\to\infty} \max\{\, a^{\frac{1}{2}} + \delta\, V_j(k(s-a)) : 0 \le a \le s\,\} \\
&= \max\{\, a^{\frac{1}{2}} + \delta\, V_{\infty}(k(s-a)) : 0 \le a \le s\,\}. \qquad \text{(Bellman equation).}
\end{aligned}$$

Since the value function is the unique locally bounded solution of Bellman equation, $V(s) = V_{\infty}(s)$ and

$$V(s) = V_{\infty}(s) = \left[1 - \delta^{\,2}k\right]^{-\frac{1}{2}}\,s^{\frac{1}{2}}.$$

The optimal strategy is also the limit of the strategies $\sigma^{*}_{t}(s)$,

$$\sigma^{*}(s) = \lim_{t\to\infty}\sigma^{*}_{t}(s) = \lim_{t\to\infty}(1 + k\,\delta^{\,2} + \cdots + k^{t-1}\,\delta^{\,2t-2})^{-1}s = (1 - k\,\delta^{\,2})\,s.$$

If $V_0(s) \equiv 0$, then $V_1(s)$ maximum over the single period $t = 1$; $V_2(s)$ is the maximum over the two periods $t = 0, 1$; by induction, $V_j(s)$ is the maximum over the j periods $t = 0, \ldots, j-1$. Taking the limit, $V_{\infty}(s)$ is the maximum over all the periods $t \ge 0$, which is the true value function.

Steps to Solve a SDP by Iteration, Method 1	
1.	Start with $V_0(s) \equiv 0$ for all s.
2.	By induction define $V_{j+1}(s) = \max\{\, r(s,a) + \delta\, V_j(f(s,a)) : a \in \mathscr{F}(s)\,\}$. $V_1(s)$ is the maximum over one period, $t = 0$. $V_2(s)$ is the maximum over two periods, $t = 0, 1$. $V_j(s)$ is the maximum over j periods, $t = 0, \ldots, j-1$.
3.	$V_j(s)$ converges uniformly to $V(s)$ on compact intervals, so $V(s)$ is continuous. $V(s)$ is maximum for all periods.
4.	The maximizer $\sigma_j(s)$ for each step converges to the optimal strategy, $\sigma^*(s)$.

Solution Method 2: In this method, the way we find the optimal solutions is to *guess* the form of the value function V with unspecified parameters. Next, we use the Bellman equation to determine the unspecified parameters in the guess. In the process, the optimal strategy is determined.

For the present problem, an outline of the solutions method is as follows. (1) Based on the reward function, we guess that it is of the form $V(s) = M s^{\frac{1}{2}}$ where M is to be determined. We could also use the first few $V_j(s)$ calculated by Method 1 to motivate a guess of the form of the true value function. (2) Next, determine the critical point $\bar{a}$ of

$$h(s,a) = r(s,a) + \delta\, V(f(s,a)) = a^{\frac{1}{2}} + \delta\, M\, k^{\frac{1}{2}} (s-a)^{\frac{1}{2}}.$$

This is the sum of the immediate payoff plus the payoff for what is carried forward to the future calculated by the value function. Verify that these are a maximum of $h(s,a)$ for $a \in \mathscr{F}(s) = [0, s]$. This value $\bar{a}$ can depend on the unspecified parameters of V as well as the data for the problem. (3) Calculate $h^*(s) = h(s, \bar{a})$. (4) Use the Bellman equation, $V(s) = h^*(s)$, to solve for the unspecified parameters of the guess of V. Finally, (5) substitute the parameters into $\bar{s}$ to determine the optimal strategy.

(1) Using the guess is that the value function has the form $V(s) = M s^{\frac{1}{2}}$, define

$$h(s,a) = a^{\frac{1}{2}} + \delta\, M\, k^{\frac{1}{2}} (s-a)^{\frac{1}{2}}.$$

(2) The critical point of h as a function of a satisfies

$$\begin{aligned}
0 = \frac{\partial}{\partial a} h(s,a) &= \tfrac{1}{2} a^{-\frac{1}{2}} - k^{\frac{1}{2}}\, \delta\, M\, \tfrac{1}{2}\, (s-a)^{-\frac{1}{2}} \\
a^{-\frac{1}{2}} &= k^{\frac{1}{2}}\, \delta\, M\, (s-a)^{-\frac{1}{2}} \\
(s-a)^{\frac{1}{2}} &= k^{\frac{1}{2}}\, \delta\, M\, a^{\frac{1}{2}}, \\
s - a &= k\, \delta^2\, M^2\, a, \\
s &= (1 + k\, \delta^2\, M^2)\, a, \qquad \text{and} \\
\bar{a} &= \frac{s}{1 + k\, \delta^2 M^2} \leq s.
\end{aligned}$$

Since $\dfrac{\partial^2}{\partial a^2} h(s,a) < 0$, the critical point $\bar{a}$ indeed maximizes h and is an optimal strategy.

(3) The maximal value of h as a function of a, with s as a parameter is as follows:

$$\begin{aligned} h^*(s) &= \bar{a}^{\frac{1}{2}} + k^{\frac{1}{2}}\,\delta\,M\,(s-\bar{a})^{\frac{1}{2}}, \\ &= \bar{a}^{\frac{1}{2}} + k^{\frac{1}{2}}\,\delta\,M\,\delta\,M\,k^{\frac{1}{2}}\bar{a}^{\frac{1}{2}} = \left(1+k\,\delta^{\,2}M^2\right)\bar{a}^{\frac{1}{2}} \\ &= \left(1+k\,\delta^{\,2}M^2\right)\left[\frac{s}{1+k\,\delta^{\,2}M^2}\right]^{\frac{1}{2}} \\ &= \left(1+k\,\delta^{\,2}M^2\right)^{\frac{1}{2}}\,s^{\frac{1}{2}}. \end{aligned}$$

(4) The value function must satisfy the Bellman equation (11), $V(s) = h^*(s)$, so

$$\begin{aligned} M\,s^{\frac{1}{2}} &= \left(1+k\,\delta^{\,2}M^2\right)^{\frac{1}{2}}\,s^{\frac{1}{2}}, \\ M^2 &= 1+k\,\delta^{\,2}M^2, \\ M^2(1-k\,\delta^{\,2}) &= 1, \\ M^2 &= \frac{1}{1-k\,\delta^{\,2}}, \qquad \text{and} \\ \bar{M} &= \left[\frac{1}{1-k\,\delta^{\,2}}\right]^{\frac{1}{2}}. \end{aligned}$$

Because the solution of the Bellman equation is unique and this function using $\bar{M}$ satisfies it, it must be the value function,

$$V(s) = \left[\frac{s}{1-k\,\delta^{\,2}}\right]^{\frac{1}{2}}.$$

(5) The optimal strategy is

$$\begin{aligned} \sigma^*(s) = \bar{a} &= \frac{s}{1+k\,\delta^{\,2}M^2} = \frac{s}{M^2} \\ &= (1-k\,\delta^{\,2})s. \end{aligned}$$

Note that we need $k\,\delta^{\,2} < 1$ for $\sigma^*(s) \geq 0$ and $V(s)$ to be defined. ■

Steps to Solve a SDP using Method 2	
1.	Guess the form of the value function, with unspecified parameters. Base the guess on either the form of $r(s,a)$ or a few steps in the iterative method.
2.	Determine the critical point $\bar{a}$ of $h(s,a) = r(s,a) + \delta\,V(f(s,a)$ using the guess for $V(s)$. The value $\bar{a}$ can depend on the unspecified parameters of V. Verify that $\bar{a}$ is a maximizer on $\mathscr{F}(s)$.
3.	Substitute $\bar{a}$ into $h(s,a)$ to determine the maximal value $h^*(s) = h(s,\bar{a})$.
4.	Use Bellman equation, $V(s) = h^*(s)$, to solve for the unspecified parameters of the guess for V. Substitute the parameters into $V(s)$ to get the value function.
5.	Substitute the parameters into $\bar{a}$ to get the optimal strategy.

Example 4.22. This is an example attributed to Weitzman. On each day, a vintner can split his time between baking bread or squeezing grapes: The amounts of effort are b_t and $1-b_t$ respectively with both b_t and w_{t+1} elements of $[0,1]$. In the next period, the amount of wine available is $w_{t+1} = 1-b_t$. The reward or utility at each period is $u(w_t,b_t) = w_t^{\frac{1}{2}}b_t^{\frac{1}{2}}$. There is a discount factor $0<\delta<1$. The quantity to be maximized is $\sum_{t=0}^{\infty}\delta^t w_t^{\frac{1}{2}}b_t^{\frac{1}{2}}$. We consider

w_t as the state variable and b_t as the action with $w_{t+1} = 1 - b_t$ the transition function. The Bellman equation is

$$V(w) = \max\{\sqrt{wb} + \delta\, V(1-b) : b \in [0,1]\,\}.$$

Method 1: In this example, the function to be maximized is

$$h_j(w,b) = w^{\frac{1}{2}} b^{\frac{1}{2}} + \delta\, V_{j-1}(1-b) \qquad b \in [0,1].$$

Taking $V_0(w) = 0$ for all w, h_1 is an increasing function of b so the maximal value is $V_1(w) = h_1^*(w) = h_1(w,1) = w^{\frac{1}{2}}$. Then, $h_2(w,b) = w^{\frac{1}{2}} b^{\frac{1}{2}} + \delta\,(1-b)^{\frac{1}{2}}$, and the critical point satisfies

$$\begin{aligned} 0 = \frac{\partial h_2}{\partial b} &= \tfrac{1}{2} w^{\frac{1}{2}} b^{-\frac{1}{2}} - \tfrac{1}{2}\delta\,(1-b)^{-\frac{1}{2}} \\ w\, b^{-1} &= \delta^{\,2}(1-b)^{-1} \\ w(1-b) &= \delta^{\,2} b \\ w &= (w + \delta^{\,2})\, b \\ \bar{b} &= \frac{w}{w+\delta^{\,2}}. \end{aligned}$$

The second derivative $\dfrac{\partial^2 h_2}{\partial a^2} < 0$, so $\bar{b}$ is a maximizer. Then, $V_2(w) = h_2^*(w) = h_2(w,\bar{b})$, so

$$\begin{aligned} V_2(w) &= w^{\frac{1}{2}} \left(\frac{w}{w+\delta^{\,2}}\right)^{\frac{1}{2}} + \delta \left(\frac{\delta^{\,2}}{w+\delta^{\,2}}\right)^{\frac{1}{2}} \\ &= \frac{w+\delta^{\,2}}{(w+\delta^{\,2})^{\frac{1}{2}}} = (w+\delta^{\,2})^{\frac{1}{2}}. \end{aligned}$$

Rather than calculate more terms, we turn to Method 2. In this treatment, we will derive formulas that allows to determine the rest of the sequence of functions for Method 1.

Method 2: (1) We look for a solution of the form $V(w) = A(w+C)^{\frac{1}{2}}$, where A and C are parameters to be determined. (2) We introduce the function to be maximized: $h(w,b) = b^{\frac{1}{2}} w^{\frac{1}{2}} + \delta A(1-b+C)^{\frac{1}{2}}$. (3) The critical point satisfies

$$\begin{aligned} 0 = \frac{\partial h}{\partial b} &= \tfrac{1}{2} b^{-\frac{1}{2}} w^{\frac{1}{2}} - \tfrac{1}{2}\delta A (C+1-b)^{-\frac{1}{2}} \\ w(C+1-b) &= \delta^{\,2} A^2 b \\ w(C+1) &= b\left[w + \delta^{\,2} A^2\right] \\ \bar{b} &= \frac{w(C+1)}{w+\delta^{\,2}A^2}. \end{aligned}$$

The second derivative $\dfrac{\partial^2 h}{\partial b^2} < 0$, so the critical point is a maximum.

(4) As a preliminary step to calculate the maximal value,

$$\begin{aligned}(C+1-\bar{b}) &= \left(\frac{\bar{b}}{w}\right)\delta^2 A^2,\\ \delta\, A(C+1-\bar{b})^{\frac{1}{2}} &= \left(\frac{\bar{b}}{w}\right)^{\frac{1}{2}}\delta^2 A^2\\ &= \frac{(C+1)^{\frac{1}{2}}\delta^2 A^2}{\left[w+\delta^2 A^2\right]^{\frac{1}{2}}}.\end{aligned}$$

The maximal value $h(w,\bar{b})$ can be given as

$$\begin{aligned}h^*(w) &= \frac{(C+1)^{\frac{1}{2}}w}{\left[w+\delta^2 A^2\right]^{\frac{1}{2}}} + \frac{(C+1)^{\frac{1}{2}}\delta^2 A^2}{\left[w+\delta^2 A^2\right]^{\frac{1}{2}}}\\ &= \frac{(C+1)^{\frac{1}{2}}\left[w+\delta^2 A^2\right]}{\left[w+\delta^2 A^2\right]^{\frac{1}{2}}}\\ &= (C+1)^{\frac{1}{2}}\left[w+\delta^2 A^2\right]^{\frac{1}{2}}.\end{aligned}$$

(5) The Bellman equation becomes

$$A(w+C)^{\frac{1}{2}} = (C+1)^{\frac{1}{2}}\left[w+\delta^2 A^2\right]^{\frac{1}{2}}.$$

Equating similar coefficients, we get $A=(C+1)^{\frac{1}{2}}$ and $C=\delta^2 A^2$, so

$$\begin{aligned}A^2 &= \delta^2 A^2+1,\\ (1-\delta^2)A^2 &= 1,\\ A^2 &= \frac{1}{1-\delta^2}, \qquad \text{and}\\ C &= \frac{\delta^2}{1-\delta^2}.\end{aligned}$$

(6) Therefore, the value function is

$$V(w) = \left(\frac{1}{1-\delta^2}\right)^{\frac{1}{2}}\left[w+\frac{\delta^2}{1-\delta^2}\right]^{\frac{1}{2}} = \frac{\left[w(1-\delta^2)+\delta^2\right]^{\frac{1}{2}}}{1-\delta^2}.$$

Using $A^2\dfrac{1}{1-\delta^2}$ and $C+1=\dfrac{\delta^2}{1-\delta^2}+1=\dfrac{\delta^2+1-\delta^2}{1-\delta^2}=\dfrac{1}{1-\delta^2}$, we get the optimal strategy

$$\bar{b}=\sigma^*(w)=\frac{w(C+1)}{w+\delta^2 A^2}=\frac{w\left[\frac{1}{1-\delta^2}\right]}{w+\frac{\delta^2}{1-\delta^2}}=\frac{w}{w(1-\delta^2)+\delta^2}.$$

Return to Method 1: In the consideration of Method 2, we saw that if $V_j(w)=A(w+C)^{\frac{1}{2}}$, then $V_{j+1}(w)=(1+C)^{\frac{1}{2}}\left[w+\delta^2 A^2\right]^{\frac{1}{2}}$ and $\sigma_{j+1}(w)=\dfrac{w(1+C)}{w+\delta^2 A^2}$. Using that $V_2(w)=$

$(w+\delta^2)^{\frac{1}{2}}$,

$$\begin{aligned}
V_3(w) &= \left(1+\delta^2\right)^{\frac{1}{2}}\left[w+\delta^2\right]^{\frac{1}{2}},\\
V_4(w) &= \left(1+\delta^2\right)^{\frac{1}{2}}\left[w+\delta^2+\delta^4\right]^{\frac{1}{2}},\\
V_{2j}(w) &= \left(1+\cdots+\delta^{2j-2}\right)^{\frac{1}{2}}\left[w+\delta^2+\cdots+\delta^{2j}\right]^{\frac{1}{2}},\\
V_{2j+1}(w) &= \left(1+\cdots+\delta^{2j}\right)^{\frac{1}{2}}\left[w+\delta^2+\cdots+\delta^{2j}\right]^{\frac{1}{2}}.
\end{aligned}$$

This sequence of functions converges to the value function found using Method 2,

$$V(w) = \left(\frac{1}{1-\delta^2}\right)^{\frac{1}{2}}\left[w+\frac{\delta^2}{1-\delta^2}\right]^{\frac{1}{2}} = \frac{[w(1-\delta^2)+\delta^2]^{\frac{1}{2}}}{1-\delta^2}.$$

The optimal strategy at each step in the process is

$$\begin{aligned}
\sigma_3(w) &= \frac{w(1+\delta^2)}{w+\delta^2},\\
\sigma_4(w) &= \frac{w(1+\delta^2)}{w+\delta^2+\delta^4},\\
\sigma_{2j+1}(w) &= \frac{w(1+\delta^2+\cdots+\delta^{2j})}{w+\delta^2+\cdots+\delta^{2j}},\\
\sigma_{2j+2}(w) &= \frac{w(1+\delta^2+\cdots+\delta^{2j})}{w+\delta^2+\cdots+\delta^{2j+2}}.
\end{aligned}$$

This sequence of strategies converges to the optimal strategy found using Method 2,

$$\sigma^*(w) = \frac{w\left[\frac{1}{1-\delta^2}\right]}{w+\frac{\delta^2}{1-\delta^2}} = \frac{w}{w(1-\delta^2)+\delta^2},$$

■

4.3.2. Theorems for Bounded Reward Function

In this section, we consider the case when the reward function $r(\mathbf{s},\mathbf{a})$ is bounded on $\mathbf{S}\times\mathbf{A}$.

SDB. The reward function r is continuous and bounded on $\mathbf{S}\times\mathbf{A}$, i.e., there exists $K>0$ such that $|r(\mathbf{s},\mathbf{a})|\le K$ for all $(\mathbf{s},\mathbf{a})\in\mathbf{S}\times\mathbf{A}$.

We show that if a SDP satisfies SDB, then the value function is finite and continuous and an optimal strategy exists. SDB is not satisfied by any of our examples, but we use this consideration to later show that the results hold for the types of examples that we have considered.

Theorem 4.23 (SDP Finite Value Function). *If a SDP satisfies SDB in addition to SD1 – SD6, then the value function $V(\mathbf{s})$ is a bounded function, so $V(\mathbf{s})<\infty$ for each $\mathbf{s}\in\mathbf{S}$.*

Proof. The total reward for any choice of actions is bounded:

$$\left|\sum_{t=0}^{\infty}\delta^t\, r(\mathbf{s}_t,\mathbf{a}_t)\right| \le \sum_{t=0}^{\infty}\delta^t\,|r(\mathbf{s}_t,\mathbf{a}_t)| \le \sum_{t=0}^{\infty}\delta^t\,K = \frac{K}{1-\delta} = K'.$$

The supremum over all allowable choices of all the $\mathbf{a}_t$ is bounded by this same constant, so the value function $V(\mathbf{s}_0)$ is bounded and takes on finite values. □

Theorem 4.24 (SDP Continuity of Value Function). *Assume that a SDP satisfies SDB and has a bounded value function V. Then the following hold.*

- **a.** *There exists a unique bounded solution of Bellman equation, and unique solution is continuous.*
- **b.** *The value function $V(\mathbf{s})$ is the unique continuous function that satisfies the Bellman equation.*

Proof. The continuity of the value function $V : \mathbf{S} \to \mathbb{R}$ cannot be proved directly from the Bellman equation because we do not know *a priori* that the right hand side is continuous.

Instead the continuity is proved by means of a process that takes a bounded function and returns another bounded function by the solution process of Method 1. In determining the value function for examples, Method 1 finds it as the limit of a sequence of functions calculated using the right-hand side of the Bellman equation. The theory behind these calculations involves (i) putting a distance between two functions (a metric on a space of functions), (ii) showing the space of functions has a property called complete (a sequence of functions that gets closer together has to converge to a function in the space of functions), and finally, showing that the construction of the sequence of function is really a contraction mapping of the space of functions. We showed earlier how this process works for two examples.

Assume $G : \mathbf{S} \to \mathbb{R}$ is any bounded function. Let

$$h_G(\mathbf{s}, \mathbf{a}) = r(\mathbf{s}, \mathbf{a}) + \delta\, G(f(\mathbf{s}, \mathbf{a})), \qquad \text{and}$$
$$\mathscr{T}(G)(\mathbf{s}) = \sup\{\, h_G(\mathbf{s}, \mathbf{a}) : \mathbf{a} \in \mathscr{F}(\mathbf{s}) \,\}.$$

It is shown that $\mathscr{T}(G) : \mathbf{S} \to \mathbb{R}$ is a new bounded function, and if G_1 and G_2 are two such functions then $\mathscr{T}(G_1)$ and $\mathscr{T}(G_2)$ are closer together than G_1 and G_2, i.e., $\mathscr{T}$ is a contraction mapping on the space of bounded functions. The set of bounded functions needs to be shown to be complete, i.e., a sequence of functions getting closer together (is a Cauchy sequence) must converge to a bounded function. Then it follows that there is a unique bounded function that is taken to itself. Since the value function is one such function, it must be the unique function satisfying the Bellman equation.

If $V_0 : \mathbf{S} \to \mathbb{R}$ is any bounded function and inductively $V_{j+1} = \mathscr{T}(V_j)$, then the sequence $V_j(s)$ converges to the unique function fixed by $\mathscr{T}$. If we start with a continuous function V_0, then all the functions $V_j(\mathbf{s})$ in the sequence are continuous by the Parametric Maximization Theorem. Because the distance between the functions $V_j(\mathbf{s})$ and $V(\mathbf{s})$ goes to zero in terms of the distance on the function space, the sequence converges uniformly to $V(S)$. But the uniform limit of continuous functions is continuous, so the value function $V(S)$ must be continuous.

See Section 4.3.4 for more details. □

Remark. Assume we start with $V_0(\mathbf{s}) \equiv 0$ and inductively let $V_{j+1} = \mathscr{T}(V_j)$. Then, $V_1(\mathbf{s})$ is the maximum over the one period $t = 0$; $V_2(\mathbf{s})$ is the maximum over the two periods $t = 0,\ 1$; $V_j(\mathbf{s})$ is the maximum over the j periods $t = 0, \ldots, j-1$. The proof of the theorem shows that $V_j(s)$ converges to the value function $V(\mathbf{s})$ that is the maximum over all periods $t \geq 0$.

Theorem 4.25 (SDP Optimal Strategy). *Assume that a SDP has a continuous value function V such that $\delta^t V(\mathbf{s}_t)$ goes to zero as t goes to infinity for any allowable sequence of $\{\mathbf{a}_t\}$ with $\mathbf{s}_{t+1} = f(\mathbf{s}_t, \mathbf{a}_t)$.*

Then optimal stationary strategy σ^* exists with $W(\mathbf{s}, \sigma^*) = V(\mathbf{s})$.
In fact, for $h(\mathbf{s}, \mathbf{a}) = r(\mathbf{s}, \mathbf{a}) + \delta\, V \circ f(\mathbf{s}, \mathbf{a})$, an optimal strategy is any choice function

$$\sigma^*(\mathbf{s}) \in \mathscr{F}^*(\mathbf{s}) = \{\, \mathbf{a} \in \mathscr{F}(\mathbf{s}) : h(\mathbf{s}, \mathbf{a}) = h^*(\mathbf{s}) \,\}.$$

Remark. Note that the Theorem is valid if $r(\mathbf{s}, a)$ is bounded (SDB) so $V(\mathbf{s})$ is bounded.

Proof. Since r, f, and V are continuous, $h(\mathbf{s},\mathbf{a}) = r(\mathbf{s},\mathbf{a}) + \delta\, V \circ f(\mathbf{s},\mathbf{a})$ is continuous. By the Parametric Maximization Theorem, the maximal value $h^*(\mathbf{s})$ is realized on $\boldsymbol{\mathscr{F}}^*(\mathbf{s}) = \{\, \mathbf{a} \in \boldsymbol{\mathscr{F}}(\mathbf{s}) : h(\mathbf{s},\mathbf{a}) = h^*(\mathbf{s}) \,\}$, which is set is nonempty, compact-valued, upper-hemicontinuous correspondence.

$$\begin{aligned} h^*(\mathbf{s}_0) &= \max\{\, r(\mathbf{s}_0,\mathbf{a}_0) + \delta\, V(f(\mathbf{s}_0,\mathbf{a}_0)) : \mathbf{a}_0 \in \boldsymbol{\mathscr{F}}^*(\mathbf{s}_0)\,\} \\ &= \max\left\{ r(\mathbf{s}_0,\mathbf{a}_0) + \delta \max_{\{\mathbf{a}_t\}} \sum_{t\ge 1} \delta^{t-1} r(\mathbf{s}_t,\mathbf{a}_t) : \mathbf{a}_0 \in \boldsymbol{\mathscr{F}}^*(\mathbf{s}) \right\} \\ &= \max_{\{\mathbf{a}_t\}} \sum_{t\ge 0} \delta^t r(\mathbf{s}_t,\mathbf{a}_t) \\ &= V(\mathbf{s}_0). \end{aligned}$$

Let $\sigma^*(\mathbf{s}) \in \boldsymbol{\mathscr{F}}^*(\mathbf{s})$ be any choice function. We next show that $V(\mathbf{s}) = W(\mathbf{s},\sigma^*)$. For $\mathbf{s}_0 \in \mathbf{S}$, let $\mathbf{a}_t^* = \mathbf{a}_t(\sigma^*,\mathbf{s}_0) = \sigma^*(\mathbf{s}_t)$ and $\mathbf{s}_{t+1}^* = \mathbf{s}_{t+1}(\sigma^*,\mathbf{s}_0) = f(\mathbf{s}_t, a_t^*)$. Also, $\mathbf{s}_0^* = \mathbf{s}_0$. By the equation for V and σ^* above and the definitions of $\mathbf{a}_t^*$ and $\mathbf{s}_{t+1}^*$,

$$\begin{aligned} V(\mathbf{s}_t^*) &= h^*(\mathbf{s}_t^*) = h(\mathbf{s}_t^*, \sigma^*(\mathbf{s}_t^*)) \\ &= r(\mathbf{s}_t^*, \sigma^*(\mathbf{s}_t^*)) + \delta\, V \circ f(\mathbf{s}_t^*, \sigma^*(\mathbf{s}_t^*)) \\ &= r(\mathbf{s}_t^*, \mathbf{a}_t^*) + \delta\, V(\mathbf{s}_{t+1}^*). \end{aligned}$$

By repeated uses of this formula,

$$\begin{aligned} V(\mathbf{s}_0) &= r(\mathbf{s}_0,\mathbf{a}_0^*) + \delta\, V(\mathbf{s}_1^*) \\ &= r(\mathbf{s}_0,\mathbf{a}_0^*) + \delta\, r(\mathbf{s}_1^*,\mathbf{a}_1^*) + \delta^2 V(\mathbf{s}_2^*) \\ &= r(\mathbf{s}_0,\mathbf{a}_0^*) + \delta\, r(\mathbf{s}_1^*,\mathbf{a}_1^*) + \delta^2 r(\mathbf{s}_1^*,\mathbf{a}_1^*) + \delta^3 V(\mathbf{s}_3^*) \\ &\;\;\vdots \\ &= \left(\sum\nolimits_{t=0}^{T-1} \delta^t r(\mathbf{s}_t^*,\mathbf{a}_t^*) \right) + \delta^T V(\mathbf{s}_T^*). \end{aligned}$$

If we let T go to infinity, then $\delta^T\, V(\mathbf{s}_T^*)$ goes to zero by the hypothesis. Therefore, the right hand side converges to

$$\sum\nolimits_{t=0}^{\infty} \delta^t r(\mathbf{s}_t^*,\mathbf{a}_t^*) = W(\mathbf{s}_0,\sigma^*),$$

and $V(\mathbf{s}_0) = W(\mathbf{s}_0,\sigma^*)$. Thus, σ^* is an optimal strategy. □

4.3.3. Theorems for One-Sector Economy

Example 4.21 considered a one-sector economy with specific feasible correspondence and reward and transition functions. In this section, we show that some of the properties of the value function and optimal strategy for an optimal growth of a one-sector economy can be proved without specifying a specific utility function (reward function) or specific production function(transition function) but merely giving assumptions on their properties. Note that the reward function is not assumed to be bounded so we cannot apply Theorems 4.23, 4.24, 4.25.

One-Sector Economy, 1-SecE: In this model, the state $s_t \ge 0$ in period t is the supply of consumption goods available. The choice $c_t \in [0, s_t] = \boldsymbol{\mathscr{F}}(s_t)$ is the consumption (action) in period t, and $r(s,c) = u(c)$ is the utility of the consumption. Given the state and consumption in period t, then $s_{t+1} = f(s_t - c_t)$ the supply at the next period, so $f(x)$ is the production function. The discount factor is $0 < \delta < 1$. The utility function u and production function f are assumed to satisfy the following.

E1. $u : \mathbb{R}_+ \to \mathbb{R}$ is continuous and strictly increasing with $u(0) = 0$ for convenience.

E2. $f : \mathbb{R}_+ \to \mathbb{R}$ is continuous and nondecreasing on $\mathbb{R}_+$ with $f(0) = 0$ (there is no free production).

E3. At least one of the following two conditions holds.

a. There is $\bar{x} > 0$ such that $f(x) \le x$ for $x \ge \bar{x}$.

b. There is a $0 < \lambda < 1$, such that

$$\delta\, u(f(x)) \le \lambda\, u(x) \qquad \text{for all } x \ge 0.$$

In Example 4.21, $u(a) = a^{\frac{1}{2}}$ and $f(x) = k\,x$. This utility function is unbounded but u and f satisfy E1, E2, and E3.b using $\lambda = \delta\, k^{\frac{1}{2}} < 1$:

$$\delta\, u(f(x)) = \delta\, k^{\frac{1}{2}} x^{\frac{1}{2}} = \lambda\, u(x) \qquad \text{for } x \ge 0.$$

Theorem 4.26. *If a* 1-SecE *satisfies E1 – E3, then the following are true.*

a. *The value function $V(s) < \infty$ for each $s \in \mathbb{R}_+$.*

b. *The value function $V(s)$ is the unique bounded solution of Bellman equation and is continuous.*

Proof using E3.a. Take any $\bar{s} \ge \bar{x}$. We restrict to $\mathbf{S} = [0, \bar{s}]$. The reward function $r(s, c)$ is bounded on $[0, \bar{s}]$. Take $s_0 \in [0, \bar{s}]$. The following inequalities proves by induction that $s_t \in [0, \bar{s}]$ for all $t \ge 0$: If $s_t \in [0, \bar{s}]$, then

$$0 = f(0) \le f(s_t - c_t) = s_{t+1} \le f(s_t) \le f(\bar{s}) \le \bar{s}.$$

(**a**) The one stage reward is bounded on compact interval $[0, \bar{s}]$ and series defining the value function converges on $[0, \bar{s}]$ as in the proof of Theorem 4.23. Since $\bar{s} \ge \bar{x}$ is arbitrary, $V(s) < \infty$ for any $s \ge 0$.

(**b**) Since $r(s, a)$ is bounded for $s_t \in [0, \bar{s}]$, so the proof of Theorem 4.24 shows $V(s)$ is continuous. □

Proof using E3.b. (**a**) Take $s_0 \ge 0$. For any allowable sequence $\{c_t\}$,

$$\delta\, u(c_t) \le \delta\, u(s_t) = \delta\, u(f(s_{t-1} - c_{t-1})) \le \delta\, u(f(s_{t-1})) \le \lambda\, u(s_{t-1}).$$

By induction,

$$\delta^t u(c_t) \le \delta^t u(s_t) \le \lambda^t u(s_0) \qquad \text{and}$$

$$V(s_0) = \sup \sum_t \delta^t u(c_t) \le \sum_t \lambda^t u(s_0) = \frac{u(s_0)}{1 - \lambda} < \infty.$$

(**b**) The proof follows the ideas in [**12**]. For an alternate proof see Section 4.3.4. Let $V^*(s) = A\, u(s)$ for $1 + \lambda\, A = A$ or $A = \dfrac{1}{1 - \lambda}$. Then,

$$\begin{aligned} u(c) + \delta\, A\, u(f(s - a)) &\le u(a) + \delta\, A\, u(f(s)) \\ &\le u(s) + A\, \lambda\, u(s) \\ &= A\, u(s), \qquad \text{so} \\ \mathscr{T}(V^*)(s) &= \sup_{0 \le c \le s} u(c) + \delta\, A\, u(f(s, c)) \\ &\le A\, u(s) = V^*(s). \end{aligned}$$

Let $V_0^*(s) = V^*(s)$ and $V_{j+1}^* = \mathscr{T}(V_j^*)$ for $j \ge 0$. Since $V_1^*(s) = \mathscr{T}(V_0^*)(s) \le V_0^*(s)$ for all s, $V_{j+1}^*(s) \le V_j^*(s)$ for all s by induction. Thus for each $s \ge 0$, $V_j^*(s) \ge 0$ is a decreasing sequence, and $\lim_{j \to \infty} V_j^*(s)$ converges to $V_\infty^*(s)$ that satisfies the Bellman equation and so is the value function. □

Theorem 4.27. *If a* 1-SecE *satisfies E1 – E3, then the following hold.*

a. *The value function* $V : \mathbb{R}_+ \to \mathbb{R}$ *is increasing.*

b. *There is an optimal strategy* $\sigma^*(s)$*, satisfying* $V(s) = u(\sigma^*(s)) + \delta\, V \circ f(s - \sigma^*(s))$.

Proof. (a) Assume $s_0 < s_0'$. Let s_t^* and c_t^* be the optimal sequences starting at $s_0^* = s_0$ with $c_t^* = \sigma^*(s_t^*)$ and $s_{t+1}^* = f(s_t - c_t)^*$. Set $c_0' = c_0^* + s_0' - s_0 > c_0$. Then $c_0' \leq s_0'$ and $s_0' - c_0' = s_0 - c_0^*$ so $s_1' = f(s_0' - c_0') = s_1^*$. Therefore, $s_t' = s_t^*$ and $c_t' = c_t^*$ are allowable for $t \geq 1$. The sequence c_t' for $t \geq 1$ is allowable starting at s_0', but not necessarily optimal. Then

$$V(s_0') \geq \sum_{t=0}^{\infty} u(c_t') = V(s_0) - u(c_0^*) + u(c_0') > V(s_0)$$

because u is strictly increasing. This proves that V is increasing.

(**b**) By Theorem 4.26, $V(s)$ is continuous so $h(s,a) = r(s,a) + \delta\, V \circ f(s,a)$ is continuous.

If E3.a is satisfied, then $V(s)$ is bounded on each $[0, \bar{s}]$ as in the proof of Theorem 4.26. Therefore, Theorem 4.25 implies that an optimal strategy exists.

If E3.b is satisfied, then by the proof of Theorem 4.26,

$$\delta^T V(s_T) = \sum_{t=T}^{\infty} \delta^t u(c_t) \leq \sum_{t=T}^{\infty} \lambda^t u(s_0) \to 0 \text{ as } T \to \infty$$

since series $\sum_{t=1}^{\infty} \lambda^t$ converges. Again, Theorem 4.25 implies that an optimal strategy exists on all $\mathbb{R}_+$. □

Theorems 4.26 and 4.27 show why the value function and optimal strategy exist for Example 4.21. We proceed to show that more of the properties of the value function and optimal strategy that Example 4.21 possessed are true for more general models of a one-sector economy with additional assumptions. The reference to this material is Section 12.6 of [**14**] by Sundaram. To prove these results, we need further assumptions on the production function f and the utility function u.

E4. The utility function u is strictly concave on $\mathbb{R}_+$.

E5. The production function f is concave on $\mathbb{R}_+$.

E6. The utility function u is C^1 on $\mathbb{R}_{++}$ with $\lim_{c \to 0+} u'(c) = \infty$.

E7. The production function f is C^1 on $\mathbb{R}_{++}$ with $f(0+) = \lim_{x \to 0+} f'(x) > 0$.

Remark. Assumptions E1 – E7 are satisfied for $u(c) = c^{\frac{1}{2}}$ and $f(x) = k\,x$ of the earlier example.

Theorem 4.28. *If a* 1-SecE *satisfies E1 – E4, then the savings function* $\xi^*(s) = s - \sigma^*(s)$ *is nondecreasing on* $\mathbb{R}_+$.

Proof. Assume not. Then there exist $s, s' \in \mathbf{S}$ with $s < s'$ and $\xi^*(s) > \xi(s')$. Let $x = \xi^*(s)$ and $x' = \xi(s')$. Since $x \leq s < s'$, x is a feasible savings level for s'. Also, $x' < x \leq s$, so x' is a feasible level of savings for s. Since x and x' are the optimal savings levels for s and s', we have

$$\begin{aligned} V(s) &= u(s - x) + \delta V(f(x)) \\ &\geq u(s - x') + \delta V(f(x')), \\ V(s') &= u(s' - x') + \delta V(f(x')) \\ &\geq u(s' - x) + \delta V(f(x)). \end{aligned}$$

Therefore,

$$u(s - x') - u(s - x) \leq \delta\,[V(f(x)) - V(f(x'))] \leq u(s' - x') - u(s' - x).$$

The function u is strictly concave and increasing and $(s-x')-(s-x) = (s'-x')-(s'-x)$, and the points on the right are larger, so

$$u(s-x') - u(s-x) > u(s'-x') - u(s'-x).$$

This contradiction proves the theorem. □

Theorem 4.29. *If a* 1-SecE *satisfies E1 – E5, then the value function* V *is concave on* $\mathbb{R}_+$.

Proof. Let $s, s' \in \mathbf{S}$ with $s < s'$. Set $s^\tau = (1-\tau)s + \tau s'$ for $0 \le \tau \le 1$. Let s_t and s'_t be the optimal sequences of states and c_t and c'_t be the sequences of consumptions starting at s and s' respectively. Note that $c_t \le s_t$ and $c'_t \le s'_t$. For each t, let $c^\tau_t = (1-\tau)c_t + \tau c'_t$. We will show that c^s_t is an allowable consumption for s^τ_t. Let x^*_t denote the sequence of investment levels if we use the sequence c^τ_t starting at s^τ.

First,

$$s^\tau = (1-\tau)s + \tau s' \ge (1-\tau)c_0 + \tau a'_0 = c^\tau_0.$$

Then, $x^*_0 = s^\tau - c^\tau_0$. Using the concavity of f,

$$\begin{aligned} f(x^*_0) &\ge (1-\tau)\, f(s-c_0) + s\, f(s'-c'_0) \\ &= (1-\tau)\, s_1 + \tau\, s'_1 \ge (1-\tau)\, c_1 + \tau\, c'_1 \\ &= c^\tau_1. \end{aligned}$$

Continuing by induction, we get that $f(x^*_t) \ge a^\tau_{t+1}$. Therefore, the sequence c^s_t is an allowable sequence for s^τ.

The sequence c^τ_t is feasible, but not necessarily optimal. The utility function is concave, so

$$\begin{aligned} V(s^\tau) &\ge \sum\nolimits_{t=0}^{\infty} \delta^t\, u(c^\tau_t) \\ &= \sum\nolimits_{t=0}^{\infty} \delta^t u\left((1-\tau)c_t + \tau a'_t\right) \\ &\ge (1-\tau) \sum\nolimits_{t=0}^{\infty} \delta^t\, u(c_t) + \tau \sum\nolimits_{t=0}^{\infty} \delta^t\, u(a'_t) \\ &= (1-\tau)V(s) + \tau\, V(s'). \end{aligned}$$

This shows that V is concave. □

Theorem 4.30. *If a* 1-SecE *satisfies E1 – E5, then the correspondence* $\boldsymbol{\mathcal{F}}^*$ *that gives the maximizers of the Bellman equation is single-valued. Therefore, the optimal strategy* σ^* *is uniquely determined and is a continuous function on* $\mathbb{R}_+$.

Proof. We are assuming that u is strictly concave and that f is concave. By Theorem 4.29, V is concave. Combining, $u(a) + \delta\, V(f(s-a))$ is a strictly concave function of a. It follows that there can be a single point that maximizes this function, so $\boldsymbol{\mathcal{F}}^*(s)$ is a single point. It follows that the optimal strategy is unique. Since an upper-hemicontinuous single valued correspondence is continuous, $\boldsymbol{\mathcal{F}}^*$ or σ^* is continuous. □

Theorem 4.31. *If a* 1-SecE *satisfies E1 – E7, then for all* $s > 0$*, the optimal strategy is an interior point of* $\boldsymbol{\mathcal{F}}(s) = [0, s]$, $0 < \sigma^*(s) < s$.

Proof. Let $s_0 > 0$, $\bar{s}_t$, $\bar{c}_t = \sigma^*(\bar{s}_t)$, and $\bar{x}_t = \bar{s}_t - \bar{c}_t$ be optimal sequence of states, consumption, and savings. Assume that not all the $\bar{x}_t$ are interior. Let $\bar{x}_\tau$ be the first $\bar{x}_t$ that is not interior. We get a contradiction to the fact that these are optimal. Since $\bar{x}_{\tau\text{-}1}$ is interior, $\bar{s}_\tau = f(\bar{x}_{\tau\text{-}1}) > 0$.

First we show that it cannot happen that $\bar{x}_\tau = 0$. If this were true, then $\bar{s}_{\tau+1} = f(\bar{x}_\tau) = 0$ and $\bar{s}_t = 0$ for all $t \ge \tau + 1$. The value function starting at $\bar{s}_\tau$ is $V(\bar{s}_\tau) = u(\bar{s}_\tau - \bar{x}_\tau) +$

$\delta V(\bar{s}_{\tau+1}) = u(\bar{s}_\tau)$. This is greater than the payoff choosing saving $z > 0$ for $t = \tau$ and savings $0 = f(z) - c_{\tau+1}$ for $t = \tau + 1$,

$$u(\bar{s}_\tau) \geq u(\bar{s}_\tau - z) + \delta\, u(f(z)) + \delta^2 u(0).$$

But

$$\begin{aligned}\frac{d}{dz}\left[u(\bar{s}_\tau - z) + \delta\, u(f(z))\right]\Big|_{z=0+} &= -u'(\bar{s}_\tau) + \delta\, u'(0+)\, f'(0+) = \infty \\ &= -u'(\bar{s}_\tau) + \delta\,(\infty)\, f'(0+) = \infty.\end{aligned}$$

Since this derivative is positive, $z = 0$ cannot be a maximum.

Second, we show that if $\bar{x}_\tau = \bar{s}_\tau$, then $\bar{x}_{\tau+1} = \bar{s}_{\tau+1}$.

$$\begin{aligned}V(\bar{s}_\tau) &= u(\bar{s}_\tau - \bar{x}_\tau) + \delta\, u\left(f(\bar{s}_\tau) - \bar{x}_{\tau+1}\right) + \delta^2 V(f(\bar{x}_{\tau+1}) \\ &= u(0) + \delta\, u\left(f(\bar{s}_\tau) - \bar{x}_{\tau+1}\right) + \delta^2 V\left(f(\bar{x}_{\tau+1})\right).\end{aligned}$$

If also $\bar{x}_{\tau+1} < \bar{s}_{\tau+1}$, then savings z at period τ can be decreased from $z = \bar{x}_\tau = \bar{s}_\tau$ while keeping $\bar{x}_{\tau+1}$ fixed,

$$\begin{aligned}u(0) + \delta\, u\left(\bar{s}_{\tau+1} - \bar{x}_{\tau+1}\right) + \delta^2 V\left(f(\bar{x}_{\tau+1})\right) \\ \geq u(\bar{s}_\tau - z) + \delta\, u\left(f(z) - \bar{x}_{\tau+1}\right) + \delta^2 V\left(f(\bar{x}_{\tau+1})\right)\end{aligned}$$

or

$$u(0) + \delta\, u\left(\bar{s}_{\tau+1}\right) - \bar{x}_{\tau+1}) \geq u(\bar{s}_\tau - z) + \delta\, u\left(f(z) - \bar{x}_{\tau+1}\right).$$

Since $\bar{x}_{\tau+1}$ is fixed, we can keep all the choices fixed for $t \geq \tau + 1$. Since this must be an optimal choice,

$$\begin{aligned}0 \leq \frac{d}{dz}\left[u(\bar{s}_\tau - z) + \delta\, u\left(f(z) - \bar{x}_{\tau+1}\right)\right]\Big|_{z=\bar{s}_\tau} \\ = -u'(0) + \delta\, u'(\bar{s}_{\tau+1} - \bar{x}_{\tau+1})\, f'(\bar{s}_\tau) \\ = -\infty + \delta\, u'(\bar{s}_{\tau+1} - \bar{x}_{\tau+1})\, f'(\bar{s}_\tau) = -\infty.\end{aligned}$$

Since this derivative is negative, it cannot be a maximum. Therefore, we would need $\bar{x}_{\tau+1} = \bar{s}_{\tau+1}$.

We have shown that if $\bar{x}_\tau = \bar{s}_\tau$, then $\bar{x}_{\tau+1} = \bar{s}_{\tau+1}$. Continuing by induction, we would need $\bar{x}_t = \bar{s}_t$ and $\bar{c}_t = 0$ for all $t \geq \tau$. Therefore, the value function starting at τ would be zero. Keeping all the $c_t = 0$ for $t \geq \tau + 1$ and increasing c_τ, we can increase the payoff. Therefore, this would not be an optimal sequence.

Thus, we have ruled out the possibility of $\bar{x}_\tau$ being on either end point, and so it and $\bar{c}_\tau$ must be interior. □

Theorem 4.32. *If a* 1-SecE *satisfies E1 – E7, then for* $s_0 > 0$*, the optimal strategy* σ^* *satisfies the Ramsey-Euler equation*

$$u'(\sigma^*(s_t)) = \delta\, u'(\sigma^*(s_{t+1}))\, f'(s_t - \sigma^*(s_t))$$

where $s_{t+1} = f(s_t - \sigma^*(s_t))$.

Proof. In the proof of the last theorem, we showed that $c = c_t = \sigma^*(s_t)$ is an interior maximum of

$$u(c) + \delta\, u\left(f(s_t - c) - x_{t+1}\right).$$

The Ramsey-Euler equation is the first order condition for an interior maximum of this function. □

Theorem 4.33. *If a* 1-SecE *satisfies E1 – E7, then the optimal strategy* σ^* *is increasing on* $\mathbb{R}_+$.

Proof. Suppose the theorem is false and there exist $s < \hat{s}$ with $c = \sigma^*(s) \geq \sigma^*(\hat{s}) = \hat{c}$. Since $s - c < \hat{s} - \hat{c}$, $s_1 = f(s-c) \leq f(\hat{s} - \hat{c}) = \hat{s}_1$. Let $c_1 = \sigma^*(s_1)$ and $\hat{c}_1 = \sigma^*(\hat{s}_1)$. By the Ramsey-Euler equation, $u'(c) = \delta\, u'(c_1)\, f'(s-c)$ and $u'(\hat{c}) = \delta\, u'(\hat{c}_1)\, f'(\hat{s} - \hat{c})$, so

$$\frac{u'(a)}{u'(\hat{c})} = \frac{u'(c_1)\, f'(s-c)}{u'(\hat{c}_1)\, f'(\hat{s} - \hat{c})} \qquad \text{or}$$
$$\frac{u'(c_1)}{u'(\hat{c}_1)} = \frac{u'(c)\, f'(\hat{s} - \hat{c})}{u'(\hat{c})\, f'(s-c)}.$$

Since u is strictly concave $u'(c) \leq u'(\hat{c})$. Also, $s - c < \hat{s} - \hat{c}$, so $f'(s-c) \geq f'(\hat{s} - \hat{c})$. Combining, we must have $u'(c_1) \leq u'(\hat{c}_1)$ and $c_1 \geq \hat{c}_1$. Thus, the situation is repeated at the next period with $s_1 < \hat{s}_1$ and $c_1 \geq \hat{c}_1$. Continuing by induction, we get that $s_t < \hat{s}_t$ and $c_t \geq \hat{c}_t$ for all t, so that $V(s) \geq V(\hat{s})$. This contradicts the fact that V is increasing and proves the theorem. □

The last theorem shows that under one additional assumption, there is a positive steady state of consumption and savings. For this theorem , we need the following assumption.

E8. The production function f is strictly concave on $\mathbb{R}_+$ and $\delta\, f(0+) > 1$.

Note that for $f(x) = kx$, then it requires $k\delta > 1$, so $k > 1$ and E3.a is not valid. However, if $k\delta^2 < 1$, then E3.b is satisfied and we still get the results.

Theorem 4.34. *If a* 1-SecE *satisfies E1 – E8, then there is a unique* x^* *such that* $\delta f'(x^*) = 1$. *Let* $s^* = f(x^*)$ *and* $c^* = s^* - x^*$ *be the associated state and consumption. For* $s_0 > 0$, *define* c_t *and* s_t *inductively by* $c_t = \sigma^*(s_t)$ *and* $s_{t+1} = f(s_t - c_t)$. *Then* $\lim_{t\to\infty} s_t = s^*$ *and* $\lim_{t\to\infty} c_t = c^*$.

Proof. Since $\delta\, f'(0+) > 1$, $\delta\, f'(\bar{x}) < 1$, and f is strictly concave, there is a unique x^* such that $\delta\, f'(x^*) = 1$. Since the savings function ξ and production function f are nondecreasing, the sequence s_t is nondecreasing function of t and has a limit s_∞. Since σ^* is nondecreasing, c_t is nondecreasing and has a limit c_∞ with $c_\infty = \sigma^*(s_\infty)$. Since $s_{t+1} = f(s_t - c_t)$,

$$s_\infty = f(s_\infty - c_\infty).$$

Also, since $u'(c_t) = \delta\, u'(c_{t+1})\, f'(s_t - c_t)$, we get that

$$u'(c_\infty) = \delta\, u'(c_\infty)\, f'(s_\infty - c_\infty) \qquad \text{so}$$
$$1 = \delta\, f'(s_\infty - c_\infty)$$

Thus, $x_\infty = s_\infty - c_\infty$ satisfies the equation for x^* and so $x_\infty = x^*$, $s_\infty = f(x_\infty) = f(x^*) = s^*$, and $c_\infty = s_\infty - x_\infty = s^* - x^* = c^*$. Therefore, these quantities are x^*, s^*, and c^*. □

For our specialized optimal growth Example 4.21, the hypotheses of Theorem 4.34 are not valid: the production function $f(x) = k\,x$ is not strictly concave and we cannot satisfy both E3.b and $f'(0) > 1/\delta$: We need $k\,\delta^2 \leq 1$ so we can satisfy E3.b, and $f'(0) = k \geq 1/\delta$ for E8. For that example, the conclusion of Theorem 4.34 are not valid since s_t and c_t both go to zero and not a nonzero limit.

4.3.4. Continuity of Value Function

In this section, we present the essential steps to prove the continuity of the value function. See Theorem 12.13 and Step 1 in Section 12.5.3 in Sundaram [**14**] for details. In order to define a convergent sequence, it is necessary to already know the point to which it converges. It is convenient to have a criterion for a sequence to converge that uses only the points in the sequence and not the limit point.

Definition. Given a norm $\| \ \|_*$ on a space $\mathbf{S}$, a sequence $\{\mathbf{x}_k\}$ in $\mathbf{S}$ is called *Cauchy* in terms of the norm $\| \ \|_*$, or is a *Cauchy sequence*, provided that for all $\epsilon > 0$ there exists a $K(\epsilon)$ such that

$$\|\mathbf{x}_j - \mathbf{x}_k\|_* < \epsilon \qquad \text{whenever } j, k \geq K(\epsilon).$$

If we do not specify the norm for $\mathbf{S} = \mathbb{R}^n$, then we mean the Euclidean norm.

Definition. A space $\mathbf{S}$ with a norm $\| \ \|_*$ is called *complete* provided that for any Cauchy sequence $\{\mathbf{x}_k\}$ in $\mathbf{S}$, there exists a point $\mathbf{a} \in \mathbf{S}$ such that the the sequence $\{\mathbf{x}_k\}$ converges to $\mathbf{a}$.

Theorem 4.35 (S1.11). *Euclidean spaces $\mathbb{R}^n$ with the Euclidean norm are complete.*

The proof must be given in $\mathbb{R}$ first. Then the fact that a sequence in $\mathbb{R}^n$ converges if and only if each of its components converge, can be used prove it in any dimension.

Since we are going to seek the solution for the Bellman equation both in the space of all bounded functions and the space of all continuous functions, we introduce a notation for these function spaces. For $\mathbf{S} \subset \mathbb{R}^n$, defined the function space of *bounded* and *bounded continuous* functions as follows:

$$\mathbf{B}(\mathbf{S}) = \{\, W : \mathbf{S} \to \mathbb{R} : W \text{ is bounded } \},$$
$$\mathbf{C}(\mathbf{S}) = \{\, W \in \mathbf{B}(\mathbf{S}) : W \text{ is continuous } \}.$$

The sup norm on these spaces is

$$\|W - V\|_0 = \sup\{\, |W(\mathbf{x}) - V(\mathbf{x})| : \mathbf{x} \in \mathbf{S} \,\}.$$

Proposition 4.36 (Uniform convergence of functions). *The metric spaces $\mathbf{B}(\mathbf{S})$ and $\mathbf{C}(\mathbf{S})$ are complete with the metric $\| \cdot \|_0$, i.e., every Cauchy sequence in one of these spaces converges to a function in the same space.*

A sketch of the proof is as follows. Let $\mathscr{X}$ equal to $\mathbf{B}(\mathbf{S})$ or $\mathbf{C}(\mathbf{S})$. If W_j is a sequence in $\mathscr{X}$ that is Cauchy, then for each $\mathbf{s}$, the sequence $W_j(\mathbf{s})$ is Cauchy in $\mathbb{R}$ and converges to a limit value $W_\infty(\mathbf{s})$ since $\mathbb{R}$ is complete. This defines the limit function. It must then be shown that W_j converges to W_∞ in terms of the norm $\| \ \|_0$ and that $W_\infty \in \mathscr{X}$. We omit the details.

Theorem 4.37. *Assume that a SDP satisfies SDB. For $W \in \mathbf{B}(\mathbf{S})$ or $\mathbf{C}(\mathbf{S})$ define*

$$\mathscr{T}(W)(\mathbf{s}) = \sup\{\, r(\mathbf{s}, \mathbf{a}) + \delta\, W \circ f(\mathbf{s}, \mathbf{a}) : \ \mathbf{a} \in \mathscr{F}(\mathbf{s}) \,\}.$$

Then, $\mathscr{T} : \mathbf{B}(\mathbf{S}) \to \mathbf{B}(\mathbf{S})$, $\mathscr{T} : \mathbf{C}(\mathbf{S}) \to \mathbf{C}(\mathbf{S})$, and $\mathscr{T}$ has a unique fixed function in $\mathbf{B}(\mathbf{S})$ and $\mathbf{C}(\mathbf{S})$. In fact, for any $W_0 \in \mathbf{B}(\mathbf{S})$ (or $\mathbf{C}(\mathbf{S})$), the sequence of functions defined inductively by $W_{j+1} = \mathscr{T}(W_j)$ converges to the fixed function.

The value function V is this fixed function, so is continuous.

Proof. If $W \in \mathbf{B}(\mathbf{S})$, then $h_W(\mathbf{s}, \mathbf{a}) = r(\mathbf{s}, \mathbf{a}) + \delta\, W \circ f(\mathbf{s}, \mathbf{a})$ is bounded, so $\mathscr{T}(W)$, which is the supremum over a, is bounded and is in $\mathbf{B}(\mathbf{S})$. If $W \in \mathbf{C}(\mathbf{S})$, then $h_W(\mathbf{s}, \mathbf{a})$ is continuous and $\mathscr{F}$ is a continuous correspondence. By the Maximum Theorem, $\mathscr{T}(W)$ is continuous on $\mathbf{S}$, and so $\mathscr{T}(W) \in \mathbf{C}(\mathbf{S})$. Thus, $\mathscr{T}$ takes the spaces into themselves.

Let $\mathscr{X}$ be either $\mathbf{B}(\mathbf{S})$ or $\mathbf{C}(\mathbf{S})$. It can be directly checked that the operator $\mathscr{T}$ has the following two properties for $W_1, W_2 \in \mathscr{X}$ and any constant c:

(i) If $W_2(\mathbf{s}) \geq W_1(\mathbf{s})$ for all $\mathbf{s} \in \mathbf{S}$, then $\mathscr{T}(W_2)(\mathbf{s}) \geq \mathscr{T}(W_1)(\mathbf{s})$ for all $\mathbf{s} \in \mathbf{S}$.

(ii) $\mathscr{T}(W_1(\cdot) + c) = \mathscr{T}(W_1(\cdot)) + \delta\, c$, where $0 < \delta < 1$ is the discount factor.

The following lemmas complete the proof that there is a unique fixed function by $\mathscr{T}$.

The value function satisfies the Bellman equation by Theorem 4.20 and so is the unique fixed function by $\mathscr{T}$ in $\mathbf{B}(\mathbf{S})$. Since the unique fixed function is also in $\mathbf{C}(\mathbf{S})$, $V(\mathbf{s})$ must be continuous. □

Lemma 4.38. *If a SDP satisfies SDB, then $\mathscr{T}$ is a contraction, $\|\mathscr{T}(W_2) - \mathscr{T}(W_1)\|_0 \leq \delta \|W_2 - W_1\|_0$ for all $W_2, W_1 \in \mathscr{X}$, where $\mathscr{X}$ equals $\mathbf{B}(\mathbf{S})$ or $\mathbf{C}(\mathbf{S})$.*

Proof. For any two $W_1, W_2 \in \mathscr{X}$, $W_2(\mathbf{s}) \leq W_1(\mathbf{s}) + \|W_2 - W_1\|_0$. By properties (i) and (ii),

$$\mathscr{T}(W_2)(\mathbf{s}) \leq \mathscr{T}(W_1 + \|W_2 - W_1\|)(\mathbf{s}) = \mathscr{T}(W_1)(\mathbf{s}) + \delta \|W_2 - W_1\|_0 \qquad \text{so}$$
$$\mathscr{T}(W_2)(\mathbf{s}) - \mathscr{T}(W_1)(\mathbf{s}) \leq \delta \|W_2 - W_1\|_0.$$

Reversing the roles of W_1 and W_2, we get $\mathscr{T}(W_1)(\mathbf{s}) - \mathscr{T}(W_2)(\mathbf{s}) \leq \delta \|W_2 - W_1\|_0$. Thus,
$|\mathscr{T}(W_2)(\mathbf{s}) - \mathscr{T}(W_1)(\mathbf{s})| \leq \delta \|W_2 - W_1\|_0$ for all $\mathbf{s} \in \mathbf{S}$. Taking the supremum, we get $\|\mathscr{T}(W_2) - \mathscr{T}(W_1)\|_0 \leq \delta \|W_2 - W_1\|_0$. □

Lemma 4.39. *A contraction $\mathscr{T}$ on a complete metric space $\mathscr{X}$ has a unique fixed point.*

Proof. Take any $W_0 \in \mathscr{X}$. By induction, define $W_{j+1} = \mathscr{T}(W_j)$. We estimate the norms of the differences:

$$\|W_{j+1} - W_j\|_0 \leq \delta \|W_j - W_{j\text{-}1}\|_0 \leq \delta^2 \|W_{j\text{-}1} - W_{j\text{-}2}\|_0 \leq \cdots \leq \delta^j \|W_1 - W_0\|_0,$$

so

$$\begin{aligned}\|W_{j+k+1} - W_j\|_0 &\leq \|W_{j+k+1} - W_{j+k}\|_0 + \|W_{j+k} - W_{j+k\text{-}1}\|_0 + \cdots + \|W_{j+1} - W_j\|_0 \\ &\leq (\delta^{j+k} + \delta^{j+k-1} + \cdots + \delta^j) \|W_1 - W_0\|_0 \\ &\leq \delta^j (1 + \delta + \delta^2 + \cdots) \|W_1 - W_0\|_0 \\ &\leq \left(\frac{\delta^j}{1-\delta}\right) \|W_1 - W_0\|_0.\end{aligned}$$

Therefore, the sequence $\{W_j\}$ is Cauchy, and so it converges to some limit W^* in $\mathscr{X}$.

Since a contraction is continuous,

$$W^* = \lim_{j\to\infty} W_j = \lim_{j\to\infty} W_{j+1} = \lim_{j\to\infty} \mathscr{T}(W_j) = \mathscr{T}(W^*),$$

so W^* is a fixed point of $\mathscr{T}$.

The fixed point can be shown unique. Assume both W^* and $\bar{W}$ are fixed. Then

$$\|W^* - \bar{W}\|_0 = \|\mathscr{T}(W^*) - \mathscr{T}(\bar{W})\|_0 \leq \delta \|W^* - \bar{W}\|_0$$
$$(1-\delta)\|W^* - \bar{W}\|_0 \leq 0.$$

Therefore, $\|W^* - \bar{W}\|_0$ must be zero and $W^* = \bar{W}$. Thus, the fixed point is unique. □

Alternative Proof of Theorem 4.26 with E3.b.

For this alternative proof when the reward function is unbounded but satisfies E3.b, we define the following new norm on functions,

$$\|W\|_* = \sup\left\{\left|\frac{W(s)}{u(s)}\right| : s > 0\right\}. \qquad (*)$$

The spaces of functions for which this norm is finite,

$$\mathbf{B}_*(\mathbb{R}_+) = \{\, W : \mathbb{R}_+ \to \mathbb{R}_+ : W(0) = 0,\ \|W\|_* < \infty \,\} \quad \text{and}$$
$$\mathbf{C}_*(\mathbb{R}_+) = \{\, W \in \mathbf{B}_*(\mathbb{R}_+) : W \text{ is continuous} \,\}.$$

are complete as before.

Lemma 4.40. *Assume a* 1-SecE *satisfies E1, E2, and E3.b.*
a. $\mathscr{T}$ *preserves both* $\mathbf{B}_*(\mathbb{R}_+)$ *and* $\mathbf{C}_*(\mathbb{R}_+)$.
b. $\mathscr{T}$ *is a contraction on* $\mathbf{B}_*(\mathbb{R}_+)$, *i.e.,* $\|\mathscr{T}(W_2) - \mathscr{T}(W_1)\|_* \leq \delta \|W_2 - W_1\|_*$ *for all* $W_2, W_1 \in \mathbf{B}_*(\mathbb{R}_+)$.

Proof. (a) All $W(s) \geq 0$, so

$$\begin{aligned}\mathscr{T}(W)(s) &= \sup\{\, u(a) + \delta\, W(f(s-a)) : a \in [0,s]\,\}\\ &\leq u(s) + \delta\, W(f(s))\\ &\leq u(s) + \delta\, \|W\|_*\, u(f(s))\\ &\leq u(s) + \lambda\, \|W\|_*\, u(s)\\ &= (1 + \lambda\, \|W\|_*)\, u(s).\end{aligned}$$

So,

$$\|\mathscr{T}(W)\|_* \leq (1 + \lambda\, \|W\|_*) < \infty.$$

Convergence in the $\|\cdot\|_*$ implies uniform convergence on compact intervals, so a Cauchy sequence in $\mathbf{C}_*(\mathbb{R}_+)$ converges to a continuous function in $\mathbf{C}_*(\mathbb{R}_+)$.

(b) For two functions $W_1, W_2 \in \mathbf{B}_*(\mathbb{R}_+)$,

$$\begin{aligned}\mathscr{T}(W_2)(s) &= \sup\{u(a) + \delta\, W_2(f(s-a)) : a \in [0,s]\,\}\\ &\leq \sup\{u(a) + \delta\, W_1(f(s-a)) + \delta\, \|W_2 - W_1\|_* u(f(s-a)) : a \in [0,s]\,\}\\ &\leq \sup\{u(a) + \delta\, W_1(f(s-a)) : a \in [0,s]\,\} + \delta\, \|W_2 - W_1\|_* u(f(s))\\ &\leq \mathscr{T}(W_1)(s) + \lambda\, \|W_2 - W_1\|_* u(s),\\ \mathscr{T}(W_2)(s) - \mathscr{T}(W_1)(s) &\leq \lambda\, \|W_2 - W_1\|_* u(s).\end{aligned}$$

Reversing the roles of W_1 and W_2, we get the other inequality, so

$$\begin{aligned}\left|\frac{\mathscr{T}(W_2)(s) - \mathscr{T}(W_1)(s)}{u(s)}\right| &\leq \lambda\, \|W_2 - W_1\|_*,\\ \|\mathscr{T}(W_2) - \mathscr{T}(W_1)\|_* &\leq \lambda\, \|W_2 - W_1\|_*.\end{aligned}$$

□

The rest of the proof goes as before since convergence in the $\|\cdot\|_*$ norm implies uniform convergence on compact intervals. □

4.3. Exercises

4.3.1. Consider the SDP problem with reward function $r(s,a) = u(a) = a^{2/3}$, transition function $f(s,a) = k(s-a)$ with $k \geq 1$, $\mathscr{F}(s) = [0,s]$, and $0 < \delta < 1$.

- **a.** Using the guess that $V(s) = M\, s^{2/3}$, find the action $a = \sigma(s)$ in terms of M that maximizes the right hand side of the Bellman equation.
- **b.** Substitute the solution of part (a) in the Bellman equation to determine the constant M and $V(s)$.
- **c.** What is the optimal strategy?

4.3.2. (Brock-Mirman growth model.) Consider the SDP problem with s_t the amount of capital at period-t, $\mathscr{F}(s) = (0,s]$ the allowable consumption, a_t consumption at period-t, reward function $r(s,a) = u(a) = \ln(a)$, discount $0 < \delta < 1$, and transition (production) function $f(s,a) = (s-a)^\beta$ with $0 < \beta \leq 1$ that determines the capital at the next period. (Note that $u(a) = \ln(a)$ is unbounded below at 0, and the choices are open at 0, but it turns out that the Bellman equation does have a solution.)

a. Using the guess that $V(s) = A + B\ln(s)$, find the action $a = \sigma(s)$ in terms of A and B that maximizes the right hand side of the Bellman equation.
b. Substitute the solution of part (a) in the Bellman equation to determine the constants A and B. *Hint*: The coefficients of $\ln(y)$ on the two sides of the Bellman equation must be equal and the constants must be equal. Solve for B first and then A.
c. What are the value function and optimal strategy?

4.3.3. Consider the SDP problem with $\mathbf{S} = [0,\infty)$, $\mathbf{A}$, $\mathscr{F}(s) = [0,s]$, $f(s,a) = 2s - 2a$, $r(s,a) = 2 - 2e^{-a}$, and $\delta = 1/2$. Start with the guess that this SDP has a value function of the form $V(s) = A - B\,e^{-bs}$.
a. Find the action $\bar{a} = \sigma(s)$ that maximizes the right hand side of the Bellman equation. (The answer can contain the unspecified parameters A, B, and b.)
b. What equations must A, B, and b satisfy to be a solution of the Bellman equation? Solve these equations for A, B, and b.
c. What are the value function and the optimal strategy?

4.3.4. Consider the SDP problem with discount $0 < \delta < 1$, $\mathbf{S} = [0,\infty)$, $\mathbf{A}$, $\mathscr{F}(s) = [0,s]$, bounded reward function $r(s,a) = \dfrac{a}{1+a} = 1 - \dfrac{1}{1+a}$, and transition function $f(s,a) = k(s-a)$ with $k > 1$ and $k\delta = 1$. Start with the guess that this SDP has a value function of the form $V(s) = \dfrac{s}{1+Bs} = \dfrac{1}{B}\left[1 - \dfrac{1}{1+Bs}\right]$.
a. What is the Bellman equation for this problem?
b. Find the action $a = \sigma(s)$ that maximizes the right hand side of the Bellman equation.
c. Substitute the solution of part (b) in the Bellman equation to determine the constant B.
d. What is the optimal strategy?

4. Exercises for Chapter 4

4.1. *Indicate* which of the following statements are always *true* and which are *false*. Also give a short *reason* for you answer.
a. For a finite horizon dynamic programming problem with C^1 reward functions $r_t(s,a)$ and C^1 transition functions $f_t(s,a)$ and continuous feasible action correspondences $\mathscr{F}_t(s)$, the optimal strategy *must be* a continuous function.
b. The feasible set $\mathscr{F}$ for a linear program is always a convex set.
c. If $f : [0,1] \times [0,1] \to \mathbb{R}$ is continuous, then the point $c^* = \sigma^*(s)$ that maximizes $f(s,c)$ for $c \in [0,s]$ is a continuous function of s.

Appendix A

Mathematical Language

We summarize some of the mathematical language with which the reader should be familiar.

We denote the set of real numbers by $\mathbb{R}$ and n-dimensional Euclidean space by $\mathbb{R}^n$.

We assume that the reader can understand common notations related to sets. To express that $\mathbf{p}$ is an element of a set $\mathbf{S}$, we write $\mathbf{p} \in \mathbf{S}$, and $\mathbf{p} \notin \mathbf{S}$ means that $\mathbf{p}$ is not an element of $\mathbf{S}$. We often define sets using notation like $\{\, x \in \mathbb{R} : x^2 \leq 4 \,\}$ for the interval $[-2, 2]$. The set $\{\, x \in \mathbb{R} : x^2 < 0 \,\}$ is the empty set, denoted by $\emptyset$. If every element of the set $\mathbf{S}_1$ is an element of $\mathbf{S}_2$, then we call $\mathbf{S}_1$ a subset of $\mathbf{S}_2$, and write $\mathbf{S}_1 \subset \mathbf{S}_2$. When $\mathbf{S}_1 \subset \mathbf{S}_2$, we allow the two sets to be equal. Given two sets $\mathbf{S}_1$ and $\mathbf{S}_2$, $\mathbf{S}_1 = \mathbf{S}_2$ if and only if $\mathbf{S}_1 \subset \mathbf{S}_2$ and $\mathbf{S}_2 \subset \mathbf{S}_1$. We often express the domain and target space of a function by a notation $f : \mathbf{S} \subset \mathbb{R}^n \to \mathbb{R}$: This means that $\mathbf{S}$ is the domain and a subset of $\mathbb{R}^n$, and f takes values in the real numbers.

Quantifiers are very important. The phrases "for all", "for every", "for any", and "for each" are all expressions for the universal quantifier and mean essentially the same thing, but they do have different connotations. When we say "for all $x \in \mathbf{S}$, $x^2 \geq x$", we are thinking collectively of the whole set of x. Consider the statement "For a function $g : \mathbb{R} \to \mathbb{R}$ and for any $b \in \mathbb{R}$, we can form the set $\{\, x \in \mathbb{R} : g(x) \leq b \,\}$." The set that follows depends on the particular value of b, so we are taking the b one at a time so we use "for any" or "for each" and not "for all".

The phases "for some" and "there exists" are expressions for the existential quantifier. The statement "for some $x \in \mathbb{R}$, $x^2 \geq 4 + x$" is true. But "for all $x \in \mathbb{R}$, $x^2 \geq 4 + x$" is false. This latter is false since there is some x, for example $x = 1$, for which the statement is not true. For a statement containing a universal quantifier to be true, it must be true for every object satisfying the quantifier. Thus, we only need to find one counter-example. Also, quantities can generally depend on things earlier in the sentence: "For any $\epsilon > 0$, there exists a $\delta > 0$ such that", means that the δ can depend on ϵ. Therefore, the order of quantifiers is important. The sentence "There exists a $\delta > 0$ such that for all $\epsilon > 0$" would mean that the same δ works for all of the ϵ.

The use of "or" is inclusive, with one or both being true. The statement "I am taking a class in Mathematics or Economics" is true even if I am taking both classes.

"If A is true, then B is true" and "A implies B" are two ways of saying the same thing that we sometimes denote by the notation "A $\Rightarrow$ B". 'A' is the hypothesis and 'B' is the conclusion. Either of these statements is equivalent to its contrapositive statement: "If B is not true, then A is not true." We often use this type of logic in our proof without explicitly commenting on it. The converse is a very different statement: "If B is true, then A is true" or in symbolic notation

"A $\Leftarrow$ B." We use "iff" to mean "if and only if", so each implies the other, "A $\Leftrightarrow$ B." Notice that the statement "If the moon is made of blue cheese then pigs can fly" is true because the hypothesis is false.

At various places we talk about necessary conditions and sufficient conditions. In the statement "If A, then B", 'A' is a sufficient condition for "B" and "B" is a necessary condition for "A". Theorem 3.12 states that "if the feasible set satisfies the constraint qualification and a differentiable function f attains a maximum for the feasible set at a point $\mathbf{x}^*$, then conditions KKT-1,2,3 hold." Thus, KKT-1,2,3 are necessary conditions for a maximum. Theorem 3.19 states that "if the feasible set is convex and a differentiable, concave function f satisfies KKT-1,2,3 at a point $\mathbf{x}^*$, then f attains a maximum at $\mathbf{x}^*$." Thus, when f is concave, conditions KKT-1,2,3 are sufficient for a maximum. The Extreme Value Theorem 2.14 states that continuity of a function on a closed and bounded set are sufficient conditions for the existence of a maximum.

The reader should be familiar with how quantifiers change when a statement is negated. The negation of "all mathematics classes are difficult" is "at least one (or some) mathematics class is not difficult." The negation of "some mathematics classes are difficult" is "all mathematics classes are not difficult."

In definitions, we use "provided that" to define a term. What follows "provided that" is the defining conditions for the word or phrase being defined. Thus, its meaning is "if and only if." Many authors use "if", but this has a different meaning than the use of "if" in theorems, so we prefer "provided that" which is not often used in other contexts.

In Chapter 4, we prove some equalities by induction, although it is only for a finite range of integers.

This quick summary of some mathematical language and logic is not complete by any means. A more complete introduction is given in a book on the foundation of higher mathematics such as Bond and Keane [**3**]. The main point is to be able to read definitions and theorems that are stated using formal mathematical language. Such statements must be read carefully and digested in order to understand their meaning. In theorems, be sure to understand what are the assumptions or hypotheses and what is the conclusion. In a definition, be sure to understand the key aspects of the concept presented.

Bibliography

1. K. Arrow and F.H. Han, *General Competitive Analysis*, Holden-Day, 1971.
2. M. Bazaraa, H. Sherali, and C. Shetty, *Nonlinear Programming: Theory and Algorithms*, Wiley Inter-Science, Hoboken NJ, 2006.
3. R. Bond and W. Keane, *An Introduction to Abstract Mathematics*, Waveland Press, 1999.
4. D. Besanko and R. Braeutigam, *Microeconomics, 4th edition*. John Wiley & Sons, New York, 2010.
5. A. Chiang, *Fundamental Methods of Mathematical Economics*, McGraw-Hill Inc., New York, 1984.
6. S. Colley, *Vector Calculus, 4th edition*, Pearson Prentice Hall, 2012.
7. C. Hassell and E. Rees, "The index of a constrained critical point", The Mathematical Monthly, October 1993, pp. 772–778.
8. H. Jongen, K. Meer, and E. Triesch, *Optimization Theory*, Kluwer Academic Publishers, Norwell MA, 2004.
9. D. Lay, *Linear Algebra and its Applications, fourth edition*, Addison-Wesley, Boston, 2012.
10. W. Rudin, *Principles of Mathematical Analysis, third edition*, McGraw-Hill, New York, 1976.
11. C. Simon and L. Blume, *Mathematics for Economists*, W. W. Norton & Company, New York, 1994.
12. N. Stokey and R. Lucas Jr, *Recursive Methods in Economic Dynamics*, Harvard University Press, Cambridge MA, 1989.
13. G. Strang, *Linear Algebra and its Applications*, Harcourt Brace Jovanovich, Publ., San Diego, 1976.
14. R. Sundaram, *A First Course in Optimization Theory*, Cambridge University Pres, New York & Cambridge UK, 1996.
15. W. Wade, *Introduction to Analysis, Fourth Edition*, Pearson Prentice Hall, Englewood Cliffs, NJ, 2010.
16. R. Walker, *Introduction to Mathematical Programming*, Pearson Learning Solutions, Boston MA, fourth edition, 2013.

Index