INTRODUCTION TO LINEAR AND NONLINEAR PROGRAMMING

Stephen E. Wright

Department of Statistics
Miami University, Oxford, OH

# Table of Contents

1. INTRODUCTION 1
   1.1 Terminology 2
   1.2 Some Facts about Calculus, Vectors, and Matrices 5
   1.3 Norms and Neighborhoods 7
   1.4 Convergence, Continuity, Closedness and Boundedness 10

2. UNCONSTRAINED OPTIMIZATION 13
   2.1 First-Order Necessary Conditions for Optimality 13
   2.2 Sufficient Conditions for Optimality — a First Look 15
   2.3 Definiteness and Quadratic Forms 19
   2.4 Back to Optimization — Sufficient Conditions 25

3. EXISTENCE OF GLOBAL OPTIMIZERS 28
   3.1 Extreme Value Theory 28
   3.2 Verifying Coercivity 31

4. CONVEXITY AND OPTIMIZATION 36
   4.1 Convex Sets and Functions 36
   4.2 Practical Tests for Convexity 41
   4.3 Further Properties of Convex Sets and Functions 46

5. ITERATIVE METHODS FOR OPTIMIZATION 49
   5.1 Descent Methods 49
   5.2 Classic Examples of Descent Methods 51
   5.3 Solving Systems of Nonlinear Equations 57

6. LINEAR LEAST SQUARES AND RELATED PROBLEMS 59
   6.1 A Motivating Application — Curve-Fitting 59
   6.2 Linear Least Squares 60
   6.3 A Very Brief Look at Nonlinear Least Squares 63
   6.4 Nearest-Point Projections and Complementary Subspaces 64
   6.5 Optimality Conditions for Linear Least Squares 66
   6.6 Nonnegative Linear Least Squares 69
1. INTRODUCTION

Optimization is the search for “extreme” values: minimum cost, greatest distance, least effort, best fit, and so on. Mathematical Programming is a field within mathematics comprising the theory and methods for solving optimization problems. (Here “programming” means “planning.”) Optimization problems are classified according to the different types of objects or decisions under consideration. Here’s a brief list of sample categories:

- Linear and Nonlinear Programming — optimization involving vectors of real numbers and a finite list of constraints.
- Combinatorial and Integer Programming — optimization involving vectors of integers or the selection of objects from a finite set.
- Stochastic Programming — optimization under uncertainty (“stochastic” = involves probability).
- Dynamic Programming — optimization of sequential processes (“dynamic” means time-dependent).

These four categories actually overlap quite a bit. In particular, the area of Linear and Nonlinear Programming provides many of the fundamental concepts and tools needed for most other areas of optimization. That’s why it’s the focus of this introductory course.

Example of a nonlinear programming problem.

\[
\begin{align*}
\text{maximize} & \quad x_1 x_2 - x_3^2 \\
\text{over all} & \quad (x_1, x_2, x_3) \in \mathbb{R}^3 \\
\text{subject to} & \quad x_1 - x_2 \geq 2, \\
& \quad x_1 + 2x_2 + 4x_3 \geq 6, \\
& \quad 3x_1 - 2x_3 = 1, \\
& \quad x_1 \geq 0, \ x_2 \geq 0, \ x_3 \geq 0.
\end{align*}
\]

Example of a nonlinear programming problem.

\[
\begin{align*}
\text{minimize} & \quad f(x_1, \ldots, x_n) \\
\text{subject to} & \quad g_1(x_1, \ldots, x_n) \leq 0, \\
& \quad \vdots \\
& \quad g_m(x_1, \ldots, x_n) \leq 0,
\end{align*}
\]

(NLP)

 Tricks for converting to the general form:

- \( \max f \iff \min -f \)
- \( g_i \geq 0 \iff -g_i \leq 0 \)
- \( g_i \leq 2 \iff g_i - 2 \leq 0 \)
- \( g_i = 0 \iff \begin{cases} g_i \leq 0 \\ -g_i \leq 0 \end{cases} \)

where \( f, g_i : \mathbb{R}^n \to \mathbb{R} \). Sometimes it’s useful to explicitly maintain the equality constraints.
rather than convert them to pairs of inequalities. Other times, it helps to convert inequalities into equalities.

The main issues in optimization are as follows:

- problem formulation (and reformulation);
- existence — feasibility, finiteness of optimum, attainment of optimum;
- identification or approximation of candidates for optimality;
- verification or refutation of claimed optimality;
- sensitivity and uniqueness of solutions.

In this course, we examine all of these issues in one context or another.

1.1 Terminology

In this section, we lay out some terminology used when discussing optimization problems. Consider an abstract optimization problem of the form

$$\min \ f(x) \ \text{s.t.} \ x \in S,$$

where $S$ is some set. (In optimization, we use “s.t.” as an abbreviation for “subject to” in short problem statements.) Here are the basic classifications of minimizers.

**Definition 1.1.1 (Global and local minimizers).**

- We say that $\bar{x} \in S$ is a *global minimizer* of $f$ on $S$ if $f(\bar{x}) \leq f(x)$ for all $x \in S$. Equivalently, we say that “$f(\bar{x})$ is a global minimum of $f$ on $S$” or that “$f$ has a global minimum on $S$ at $\bar{x}$.”

- We say that $\bar{x} \in S$ is a *local* minimizer of $f$ on $S$ if $f(\bar{x}) \leq f(x)$ for all $x \in S$ in some neighborhood of $\bar{x}$. Equivalently, we say that “$f(\bar{x})$ is a local minimum of $f$ on $S$” or that “$f$ has a local minimum on $S$ at $\bar{x}$.”

Note that each global minimizer is also a local minimizer.

In the case where $S \subset \mathbb{R}$, a “neighborhood” of $\bar{x}$ is the set of all $x$ for which $|x - \bar{x}| < \delta$, where $\delta > 0$ is some fixed number. In other words, the above definition of a local minimizer in this case is: there exists $\delta > 0$ so that $f(\bar{x}) \leq f(x)$ whenever $x \in S$ satisfies $|x - \bar{x}| < \delta$.

The formal definition of a neighborhood of a vector $\bar{x} \in \mathbb{R}^n$ is given in Section 1.3 later in this chapter.

The above definitions of “global” and “local” minimizer are standard in the mathematical programming community, but they may differ somewhat from the definitions of “relative” and “absolute” minimizer given in some Calculus texts.
The issue of uniqueness of a minimizer is important for certain applications and for various technical purposes, so we want to highlight such cases with a special definition.

**Definition 1.1.2** (Strict minimizers).

- We say that $\bar{x} \in S$ is a strict global minimizer of $f$ on $S$ if $f(\bar{x}) < f(x)$ for all $x \in S$ with $x \neq \bar{x}$.
- We say that $\bar{x} \in S$ is a strict local minimizer of $f$ on $S$ if $f(\bar{x}) < f(x)$ for all $x \in S$ in some neighborhood of $\bar{x}$ with $x \neq \bar{x}$.

The analogous definitions for maximizers (global, local, strict) are given by simply reversing the inequalities $\leq$ and $<$ in the above definitions. Examples are shown in Figure 1.1.

![Figure 1.1 Local and global max/min](image)

Some additional terminology is used to describe the range of $f(x)$ as $x$ varies over $S$.

**Definition 1.1.3** (Lower bounds; infimum).

- We say that $f$ is bounded below on $S$ if there is a number $\alpha$ so that $f(x) \geq \alpha$ for all $x \in S$. In this case, we call $\alpha$ a lower bound for $f$ on $S$.
- We say that $f$ is unbounded below on $S$ if it has no lower bound.
- The infimum of $f$ on $S$, written as

$$\inf_S f \quad \text{or} \quad \inf\{f(x) \mid x \in S\},$$

is the greatest lower bound for $f$ on $S$. As a convention, we have two special cases:

$$\inf_S f = \begin{cases} -\infty, & \text{if } f \text{ is unbounded below on } S; \\ +\infty, & \text{if } S \text{ is empty.} \end{cases}$$

Again, there are analogous notions related to maximizing: we give these next. Examples are shown in Figure 1.2.
Definition 1.1.4 (Upper bounds; supremum). We say that \( f \) is bounded above on \( S \) if there is a number \( \alpha \) so that \( f(x) \leq \alpha \) for all \( x \in S \); in this case, we call \( \alpha \) an upper bound for \( f \) on \( S \). We say that \( f \) is unbounded above on \( S \) if it has no upper bound. The supremum of \( f \) on \( S \), written as \( \sup_S f \) or \( \sup \{ f(x) \mid x \in S \} \), is the least upper bound for \( f \) on \( S \). As a convention, we have two special cases:

\[
\sup_S f = \begin{cases} 
+\infty, & \text{if } f \text{ is unbounded above on } S; \\
-\infty, & \text{if } S \text{ is empty.}
\end{cases}
\]

\[
\sup f = +\infty
\]
\[
\inf f = 0
\]
(no global min)

\[
\sup f = +1
\]
\[
\inf f = -1
\]

Figure 1.2 Infimum and supremum

Definition 1.1.5 (Attainment; minimum and maximum value). When \( f \) has a global minimizer on \( S \), we say that the infimum is attained. In this case, we rewrite \( \inf_S f \) as \( \min_S f \) and refer to the minimum value of \( f \) on \( S \). Likewise, a supremum that is attained is referred to as a maximum value and written as \( \max_S f \).

Notice that although a minimum may not exist, an infimum always exists. This is a defining property of the set \( \mathbb{R} \) of real numbers:

\textit{every set of real numbers that is bounded below must admit a real infimum.}

In particular, this property distinguishes the real numbers from the rational numbers. If a given set of real numbers is unbounded below, or if the set is empty, then our conventions for \( \inf_S f = \pm\infty \) come into play.

In cases where the infimum is not attained or not yet known to be attained, a useful idea is the following:

Definition 1.1.6 (Minimizing sequences). A minimizing sequence for \( f \) on \( S \) is a sequence of points \( \{ x^{(k)} \} \subset S \) for which \( \{ f(x^{(k)}) \} \) is a decreasing sequence converging to \( \inf_S f \).

As long as \( S \) is nonempty, a minimizing sequence must exist.

Example 1.1.7 (Some minimizing sequences).

- For \( f(x) = e^x \) on \( \mathbb{R} \), a minimizing sequence is given by \( x^{(k)} = -k \).
For \( f(x) = 1/x \) on \((0, \infty)\), a minimizing sequence is given by \( x^{(k)} = k \).

For \( f(x) = 1/x \) on \((-\infty, 0)\), a minimizing sequence is given by \( x^{(k)} = -1/k \).

For \( f(x) = \sin x \) on \(\mathbb{R} \), a minimizing sequence is given by \( x^{(k)} = -(\pi/2) + (1/k) \).

### 1.2 Some Facts about Calculus, Vectors, and Matrices

The main tools for this course come from Calculus and Linear Algebra. This section collects some basic facts and introduces the notation used in the course. We begin by recalling the Taylor expansion, which plays an important role in much of what follows.

**Theorem 1.2.1** (Taylor’s formula with remainder). Let \( f : (a, b) \to \mathbb{R} \) and suppose that the derivatives \( f', f'', \ldots, f^{(n)}, f^{(n+1)} \) exist on the interval \((a, b)\). Let \( x, \bar{x} \) be points in \((a, b)\). Then there exists \( z \) strictly between \( x \) and \( \bar{x} \) such that

\[
 f(x) = f(\bar{x}) + f'(\bar{x})(x - \bar{x}) + \frac{1}{2} f''(\bar{x})(x - \bar{x})^2 + \cdots + \frac{1}{n!} f^{(n)}(\bar{x})(x - \bar{x})^n \\
+ \frac{1}{(n+1)!} f^{(n+1)}(z)(x - \bar{x})^{n+1}.
\]

Note that the last term depends on \( z \), but the others don’t.

The setting for most of linear and nonlinear programming is the \( n \)-dimensional real vector space \( \mathbb{R}^n \). In this course, we always think of vectors \( x \in \mathbb{R}^n \) as column matrices

\[
 x = \begin{bmatrix}
 x_1 \\
 x_2 \\
 \vdots \\
 x_n
\end{bmatrix},
\]

although we sometimes write vectors as row matrices \( x = (x_1, \ldots, x_n) \) for convenience. Recall that the dot product of \( x, y \in \mathbb{R}^n \) is defined by

\[
 x \cdot y = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n.
\]

When working with vectors and matrices, we often use the word “scalar” used to mean “number.” In particular, the dot product is also called the “scalar product” because it results in a scalar. Two vectors are said to be orthogonal when their dot product is zero.

**Definition 1.2.2** (Matrix transposes). Let \( A \) be an \( m \times n \) matrix. The transpose of \( A \), denoted \( A^T \), is the \( n \times m \) matrix whose columns (when turned sideways to form row vectors) are the rows of \( A \). An \( n \times n \) matrix \( A \) is said to be symmetric if \( A = A^T \).
Proposition 1.2.3 (Transpose and symmetry rules).

- For matrices $A$ and $B$, the transposes satisfy $(AB)^T = B^T A^T$ if either product is well-defined.
- For column vectors $x, y \in \mathbb{R}^n$ we have $x \cdot y = x^T y = y^T x$.
- A square matrix $A$ is invertible if and only if its inverse $A^{-1}$ is invertible, in which case $(A^T)^{-1} = (A^{-1})^T$.
- The inverse (if it exists) of a symmetric matrix is also symmetric.
- The matrices $A^T A$ and $AA^T$ are always symmetric, whereas the matrices $A^T BA$ and $ACA^T$ are symmetric if $B$ and $C$ are symmetric.

One of the central notions needed in optimization is that of a derivative for multivariable functions.

Definition 1.2.4 (Derivative row-matrix). The derivative of a real-valued function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is the row matrix

$$f'(x) = \left[ \frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right].$$

Using a row matrix makes some calculus rules easier. For example, if $c$ is a row matrix then $f(x) = cx$ is a well-defined function whose derivative is $f'(x) = c$. As another example, see the matrix version of the chain rule below.

Definition 1.2.5 (Gradient). The gradient of $f$ is the column matrix whose transpose is the derivative row-matrix:

$$\nabla f(x) = (f'(x))^T.$$

There are several ways to define what it means for a multivariable function to be “differentiable,” but each definition leads to the derivative row-matrix given above. Furthermore, all the definitions are automatically satisfied by the following idea, which we use throughout this course. More details are provided in the sequel course, MTH 632.

Definition 1.2.6 (Continuous differentiability). The function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is continuously differentiable at $\bar{x}$ if the derivative row-matrix $f'(x)$ exists on a neighborhood of $\bar{x}$ and the entries of $f'(x)$ are continuous at $\bar{x}$.

The notion of “neighborhood” for a vector $\bar{x} \in \mathbb{R}^n$ is defined formally in the next section; the meaning of “continuous” is discussed in more detail in Section 1.4.

Next we consider a function $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ that converts vectors in $\mathbb{R}^n$ into vectors in
It’s often best to think of $g$ as a column matrix of real-valued functions:

$$g(x) = g(x_1, x_2, \cdots, x_n) = \begin{bmatrix} g_1(x) \\ g_2(x) \\ \vdots \\ g_m(x) \end{bmatrix}.$$ 

**Definition 1.2.7** (Derivative matrix; Jacobian). The derivative of $g : \mathbb{R}^n \to \mathbb{R}^m$ is the rectangular matrix of all partial derivatives:

$$g'(x) = \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} & \cdots & \frac{\partial g_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_m}{\partial x_1} & \frac{\partial g_m}{\partial x_2} & \cdots & \frac{\partial g_m}{\partial x_n} \end{bmatrix} = \begin{bmatrix} g'_1(x) \\ g'_2(x) \\ \vdots \\ g'_m(x) \end{bmatrix}.$$ 

This matrix is also called the Jacobian of $g$. We say that $g$ is continuously differentiable if all the entries in $g$ are continuously differentiable.

**Example 1.2.8** (Derivatives of linear and constant mappings). If $g(x) = Ax + b$ for some matrix $A$ and vector $b$, then $g'(x) = A$.

We give two versions of the chain rule for Jacobians: the first is from multivariable calculus, while the second is often easier to remember and use.

**Theorem 1.2.9** (Chain rule — coordinate form). Suppose $w$ is a continuously differentiable function of $n$ variables $u_1, \ldots, u_n$, where each $u_i$ is a continuously differentiable function of $m$ variables $x_1, \ldots, x_m$. Then

$$\frac{\partial w}{\partial x_i} = \frac{\partial w}{\partial u_1} \frac{\partial u_1}{\partial x_i} + \frac{\partial w}{\partial u_2} \frac{\partial u_2}{\partial x_i} + \cdots + \frac{\partial w}{\partial u_n} \frac{\partial u_n}{\partial x_i}, \text{ for each } i = 1, \ldots, m.$$ 

**Theorem 1.2.10** (Chain rule — matrix-vector form). Suppose $g : \mathbb{R}^p \to \mathbb{R}^m$ and $h : \mathbb{R}^n \to \mathbb{R}^p$ are both continuously differentiable. Then $f = g \circ h : \mathbb{R}^n \to \mathbb{R}^m$ is continuously differentiable and the $m \times n$ derivative of $f(x) = g(h(x))$ is given by $f'(x) = g'(h(x))h'(x)$.

### 1.3 Norms and Neighborhoods

The idea of “neighborhoods” for vectors can be stated in terms of “distance” between vectors, whereas distances are most frequently specified in terms of “vector norms.”

**Definition 1.3.1** (Vector norms). The (Euclidean) norm $\|x\|$ of a vector $x \in \mathbb{R}^n$ is defined to be the real number

$$\|x\| = (x \cdot x)^{1/2} = (x_1^2 + \cdots + x_n^2)^{1/2}.$$
In the one-dimensional case ($\mathbb{R}^1 = \mathbb{R}$), the norm is just the absolute value: $|x| = \sqrt{x^2}$.

Here are some fundamental properties of norms used in this course.

**Proposition 1.3.2** (Properties of norms).

- $\|x\| \geq 0$, for all $x$;
- $\|x\| = 0$ if and only if $x = 0$;
- $\|\alpha x\| = |\alpha| \|x\|$, if $\alpha \in \mathbb{R}$;
- $\|x + y\| \leq \|x\| + \|y\|$ (Triangle Inequality);
- $|x \cdot y| \leq \|x\| \|y\|$ (Schwarz Inequality);
- $|x \cdot y| = \|x\| \|y\|$ if and only if $x = \alpha y$ or $y = \beta x$.

The quantity $\|x - y\|$ represents the distance from $x$ to $y$, which motivates the following definition.

**Definition 1.3.3** (Balls and boxes).

- The open ball of radius $r > 0$ about $x \in \mathbb{R}^n$ is $B(x, r) = \{ y \in \mathbb{R}^n \mid \|x - y\| < r \}$.
- The closed ball of radius $r > 0$ about $x \in \mathbb{R}^n$ is $\bar{B}(x, r) = \{ y \in \mathbb{R}^n \mid \|x - y\| \leq r \}$.
- The open box of radius $r > 0$ about $x \in \mathbb{R}^n$ is
  \[ B_\infty(x, r) = \left\{ y \in \mathbb{R}^n \mid |x_i - y_i| < r, \text{ for } i = 1, \ldots, n \right\}. \]
- The closed box of radius $r > 0$ about $x \in \mathbb{R}^n$ is
  \[ \bar{B}_\infty(x, r) = \left\{ y \in \mathbb{R}^n \mid |x_i - y_i| \leq r, \text{ for } i = 1, \ldots, n \right\}. \]

**Examples 1.3.4** (Balls and boxes in low dimensions).

(a) In $\mathbb{R}^1$, where $\|x\| = |x|$, we have
\[
B(x, r) = B(x, r) = (x - r, x + r), \quad \bar{B}(x, r) = \bar{B}_\infty(x, r) = [x - r, x + r].
\]

(b) In $\mathbb{R}^2$, $B(x, r)$ consists of all points inside, but not actually on, the circle of radius $r$ with center at $x$. Likewise, $B_\infty(x, r)$ consists of all points inside, but not actually on, $\bar{B}_\infty(x, r)$.

---

† A scalar-valued function on a real vector space is a “norm” if it satisfies the first four properties in Proposition 1.3.2. In this course, we only use the norm of Definition 1.3.1.
the square centered at \( x \) with edges of length \( 2r \) parallel to the coordinate axes. Their closed counterparts are similar but also include the “boundary” points.

(c) In \( \mathbb{R}^3 \), \( B(x, r) \) is the “inside” of a sphere and \( B_\infty(x, r) \) is the “inside” of a cube.

**Definition 1.3.5 (Neighborhoods).** A neighborhood of a vector \( x \in \mathbb{R}^n \) is any set that contains a (closed or open) ball or box about \( x \). In particular, balls and boxes are themselves examples of neighborhoods.

**Definition 1.3.6 (Local minimizer — revisited).** We say \( \bar{x} \in S \) is a local minimizer of \( f \) on \( S \) when there is some neighborhood \( N \) (i.e., a box or ball, either open or closed) about \( \bar{x} \) so that \( f(\bar{x}) \leq f(x) \) for every \( x \in N \cap S \). Similarly, \( \bar{x} \in S \) is a strict local minimizer of \( f \) on \( S \) if there is some neighborhood \( N \) about \( \bar{x} \) so that \( f(\bar{x}) < f(x) \) for every \( x \in N \cap S \) for which \( x \neq \bar{x} \).

With the notion of neighborhood in hand, we can also give a precise definition of the “interior” of a set.

**Definition 1.3.7 (Interior of a set).** We say that \( \bar{x} \) is an interior point of the set \( S \subseteq \mathbb{R}^n \) if \( S \) contains a neighborhood of \( \bar{x} \). The interior of \( S \) is the set of all interior points of \( S \); it is written as \( \text{int} S \).

**Examples 1.3.8 (Some interiors).**

(a) If \( S \) is a line segment in \( \mathbb{R} \), then “interior point” means “not an endpoint.”

(b) \( B(x, r) \) is the interior of \( \bar{B}(x, r) \).

(c) The interior of \( \{(x_1, x_2) \mid x_1^2 - x_2 \leq 1\} \) is the set \( \{(x_1, x_2) \mid x_1^2 - x_2 < 1\} \).

**Definition 1.3.9 (Boundary of a set).** A point \( \bar{x} \in \mathbb{R}^n \) is called a boundary point of the set \( S \subseteq \mathbb{R}^n \) if it is neither an interior point of \( S \), nor an interior point of the complement of \( S \) in \( \mathbb{R}^n \). The boundary of \( S \) is the set of all boundary points of \( S \).

Note that the boundary of \( S \) is the same as the boundary of the complement of \( S \).

**Examples 1.3.10 (Some boundaries).**

(a) The boundary points of the intervals \((a, b), [a, b), (a, b], [a, b]\) are \( a \) and \( b \).

(b) The only boundary point of the intervals \((a, \infty), [a, \infty), (-\infty, a], and \((-\infty, a)\) is \( a \).

(c) The interval \((-\infty, \infty)\) has no boundary points. Likewise, the full space \( \mathbb{R}^n \) has no boundary points.

(d) The boundary of both \( B(x, r) \) and \( \bar{B}(x, r) \) is the sphere \( \{y \in \mathbb{R}^n \mid \|x - y\| = r\} \).

**Definition 1.3.11 (Open and closed sets).** We say that \( S \) is open if it contains none of its boundary points. We say that \( S \) is closed if it contains all of its boundary points.
Note that $S$ is open if and only if $S$ is its own interior.

**Examples 1.3.12** (Some open and closed sets).

(a) The intervals $(a, b)$, $(-\infty, b)$, $(a, \infty)$, and $(-\infty, \infty)$ are open.
(b) The intervals $[a, b]$, $(-\infty, b]$, $[a, \infty)$, and $(-\infty, \infty)$ are closed.
(c) The intervals $(a, b]$ and $[a, b)$ are neither open nor closed.
(d) The empty set and $\mathbb{R}^n$ itself are the only subsets of $\mathbb{R}^n$ that are both open and closed.
(e) Singletons $\{a\}$ are closed.
(f) $B(x, r)$ and $B_\infty(x, r)$ are open, whereas $\bar{B}(x, r)$ and $\bar{B}_\infty(x, r)$ are closed.
(g) The intersection and union of finitely many closed sets are closed.
(h) The intersection and union of finitely many open sets are open.

Some more examples of closed sets are given in the next section.

1.4 Convergence, Continuity, Closedness and Boundedness

This section outlines concepts underlying the basic existence theory of global optimizers.

**Definition 1.4.1** (Convergence). We say the sequence of points $\{x^{(k)}\} \subseteq \mathbb{R}^n$ converges to $a \in \mathbb{R}^n$, written $a = \lim_{k \to \infty} x^{(k)}$, if for every $\epsilon > 0$ there is some $K$ so that $\|x^{(k)} - a\| < \epsilon$ whenever $k \geq K$. Both of the following are equivalent to $a = \lim_{k \to \infty} x^{(k)}$:

- $a_i = \lim_{k \to \infty} x^{(k)}_i$ for each $i = 1, \ldots, n$;
- $0 = \lim_{k \to \infty} \|x^{(k)} - a\|$.

A sequence that doesn’t converge is said to diverge.

**Example 1.4.2** (Some convergent and divergent sequences).

(a) The sequence in $\mathbb{R}^2$ defined by $x^{(k)} = (1 + 1/k, e^{-k})$ converges to $a = (1, 0)$.
(b) The sequence in $\mathbb{R}^2$ defined by $x^{(k)} = (\cos(k)/k, 3)$ converges to $a = (0, 3)$.
(c) The sequence in $\mathbb{R}^2$ defined by $x^{(k)} = (k, 4/k)$ diverges.

**Definition 1.4.3** (Continuity). We say that $f : S \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is continuous on $S$ at $a \in S$ if, for any sequence $\{x^{(k)}\} \subseteq S$ with $a = \lim_{k \to \infty} x^{(k)}$, one has $f(a) = \lim_{k \to \infty} f(x^{(k)})$. Equivalently, each component $f_i : S \to \mathbb{R}$ is continuous at $a$. We say that $f$ is continuous on $S$ if $f$ is continuous at each point in $S$. 

10
Example 1.4.4 (Some continuous functions).

(a) The elementary functions of calculus are continuous on their domains: constant functions, power functions, trigonometric functions, exponential functions, logarithms.

(b) Any composition of continuous functions is continuous.

(c) Sums, products, and quotients of continuous functions are continuous on their domains. This includes linear mappings and quadratic functions of several variables.

Note: parts (a)–(c) remain true if “continuous” is replaced by “continuously differentiable.”

(d) The norm and the absolute value function are continuous.

(e) Functions of a single variable are continuous at points of differentiability.

(f) Continuity of functions defined piecewise using continuous functions must be verified by checking limits at the break points. This is often straightforward for functions of a single variable, but is more difficult in several dimensions: one must consider all possible approaches to a given break point (lines alone do not suffice).

Proposition 1.4.5 (Sequential characterization of closed sets). A set \( S \subseteq \mathbb{R}^n \) is closed if and only if it contains the limit of every convergent sequence of elements of \( S \).

Examples of closed sets were given in the preceding section. Here are a few more.

Example 1.4.6 (Some more closed sets).

(a) Linear subspaces are closed.

(b) The set of integers is closed.

(c) For a continuous function \( f : \mathbb{R}^n \to \mathbb{R} \), sets of the form

\[
\{ x \in \mathbb{R}^n \mid f(x) \leq a \},
\]

\[
\{ x \in \mathbb{R}^n \mid f(x) \geq a \},
\]

\[
\{ x \in \mathbb{R}^n \mid f(x) = a \}
\]

are all closed.

(d) The graphs of continuous functions are closed.

(e) Cartesian products of closed sets are closed.

Definition 1.4.7 (Bounded sets). A subset of \( \mathbb{R}^n \) is said to be bounded if it lies entirely within some ball (or equivalently, within some box). A set is unbounded if it is not bounded.

Example 1.4.8 (Some bounded and unbounded sets).

(a) Balls, boxes, and polygons are bounded.
(b) Lines and parabolas are unbounded.
(c) Subsets of bounded sets are bounded.
(d) Intersections of bounded sets are bounded.
(e) The union of finitely many bounded sets is bounded.
(f) Cartesian products of bounded sets are bounded.

Definition 1.4.9 (Subsequences). Let \( \{x^{(k)}\}_{k=1}^{\infty} \) be a given sequence and let \( \{k_i\}_{i=1}^{\infty} \) be a strictly increasing sequence of whole numbers. The subsequence of \( \{x^{(k)}\} \) corresponding to \( \{k_i\} \) is the sequence \( \{x^{(k_i)}\}_{i=1}^{\infty} \).

The notation for sequences and subsequences varies widely and is sometimes ambiguous, although the ambiguities can usually made resolved from the context.

Example 1.4.10.

(a) \( \{2, 4, 6, 8, \ldots\} \) is a subsequence of \( \{1, 2, 3, 4, 5, \ldots\} \). [Uses \( k_i = 2i \).]

(b) \( \{2, 10, 6, 8, 12, 14, 16, \ldots\} \) is not a subsequence of \( \{1, 2, 3, 4, 5, \ldots\} \). [Wrong order.]

(c) The sequence \( \{4^k\}_{k=1}^{\infty} \) is a subsequence of \( \{2^k\}_{k=1}^{\infty} \). [Uses \( k_i = 2i \).]
2. UNCONSTRAINED OPTIMIZATION

In this chapter, we develop principles for identifying and validating candidates for optimality in problems with no explicit constraints:

$$\min f(x) \text{ over all } x \in S,$$

where $S$ is typically the whole space $\mathbb{R}^n$ or some simply described subset, such as an interval or neighborhood of a given point. To lay the foundation for these developments, we first review and refine the one-dimensional optimization principles from Calculus. These are then extended in a natural way to higher dimensions, where they are made practical by incorporating ideas from Linear Algebra.

2.1 First-Order Necessary Conditions for Optimality

We start with the one-dimensional case of optimizing a function on an interval in $\mathbb{R}$. The following theorem, due to Fermat, is the best known result in optimization theory.

**Theorem 2.1.1** (Necessary condition for one-dimensional optimality). Suppose $\bar{x}$ is a local minimizer (or maximizer) of $f$ on the interval $I \subseteq \mathbb{R}$. If $\bar{x}$ is not an endpoint of $I$ and $f$ is differentiable at $\bar{x}$, then $f'(\bar{x}) = 0$.

**Definition 2.1.2** (Critical point). We say that $\bar{x}$ is a critical point for $f$ if $f'(\bar{x}) = 0$.

**Observation 2.1.3** (Identifying candidates via necessary conditions). Note that Theorem 2.1.1 provides us with candidates for optimality. If $\bar{x}$ is a local minimizer for $f$ on $I \subseteq \mathbb{R}$, then one of the following conditions must hold:

(a) $\bar{x}$ is an endpoint of $I$;
(b) $f$ is not differentiable at $\bar{x}$; or
(c) $\bar{x}$ is a critical point for $f$.

We may discard any points that do not satisfy at least one of these three conditions. Any rule for selecting candidates in this way is called a necessary condition for optimality.

**Proof of Theorem 2.1.1.** Recall that

$$f'(\bar{x}) = \lim_{x \to \bar{x}} \frac{f(x) - f(\bar{x})}{x - \bar{x}}.$$

Since $\bar{x}$ is a local minimizer on $I$ and not an endpoint of $I$, there must exist $\delta > 0$ so that $f(\bar{x}) \leq f(x)$ for all $x \in (\bar{x} - \delta, \bar{x} + \delta)$. In other words, $f(x) - f(\bar{x}) \geq 0$ for all $x$ close to $\bar{x}$, so

$$\frac{f(x) - f(\bar{x})}{x - \bar{x}} \leq 0, \text{ if } x \in (\bar{x} - \delta, \bar{x}).$$
and
\[
\frac{f(x) - f(\bar{x})}{x - \bar{x}} \geq 0, \text{ if } x \in (\bar{x}, \bar{x} + \delta).
\]

Thus,
\[
\lim_{x \to \bar{x}^-} \frac{f(x) - f(\bar{x})}{x - \bar{x}} \leq 0
\]
and
\[
\lim_{x \to \bar{x}^+} \frac{f(x) - f(\bar{x})}{x - \bar{x}} \geq 0,
\]
so \( f'(\bar{x}) = 0. \)

By considering one-dimensional derivatives along the coordinate axes, we can extend this result to higher dimensions.

**Theorem 2.1.4** (Necessary condition for multidimensional optimality). Consider a set \( S \subseteq \mathbb{R}^n \) and a function \( f : S \to \mathbb{R}^n \). Suppose that \( \bar{x} \) is a local minimizer for \( f \) on \( S \). If \( \bar{x} \) is an interior point of \( S \) and \( \nabla f(\bar{x}) \) exists, then \( \nabla f(\bar{x}) = 0 \).

**Proof.** Given a coordinate \( i \), we need to show that \( \partial f(\bar{x})/\partial x_i = 0 \). Define
\[
\varphi_i(t) = f(\bar{x}_1, \ldots, \bar{x}_{i-1}, t, \bar{x}_{i+1}, \ldots, \bar{x}_n).
\]
This is a function of one variable: \( t \). As an interior point \( S \), \( \bar{x} \) can be varied slightly in any coordinate and stay in \( S \). Thus \( \bar{t} = \bar{x}_i \) minimizes \( \varphi(t) \) on some interval \((\bar{x}_i - \delta_i, \bar{x}_i + \delta_i)\). By the one-dimensional necessary condition (Theorem 2.1.1), we must have
\[
0 = \varphi_i'(0) = \frac{\partial f}{\partial x_i}(\bar{x}),
\]
as desired. \( \square \)

**Observation 2.1.5** (Multidimensional necessary conditions). The above theorem gives a necessary condition for optimality in \( \mathbb{R}^n \). For \( \bar{x} \) to be a local minimizer, it is necessary for \( \bar{x} \) to satisfy one of three conditions:

(a) \( \bar{x} \) is a boundary point of \( S \);

(b) \( \nabla f(\bar{x}) \) does not exist; or

(c) \( \bar{x} \) is a critical point. \( \square \)

**Example 2.1.6** (Using the necessary conditions). Find all candidates on \( \mathbb{R}^2 \) for local minimizers and maximizers of \( f(x_1, x_2) = x_1^3 x_2 - 2x_1 x_2^2 - 2x_1 x_2 \).

**Solution.** First we differentiate \( f \):
\[
\nabla f(x) = \begin{bmatrix} 3x_1^2 x_2 - 2x_2^2 - 2x_2 \\ x_1^3 - 4x_1 x_2 - 2x_1 \end{bmatrix}.
\]
We see that a critical point must satisfy \((3x^2 - 2x_2 - 2)x_2 = 0\) and \(x_1(x_1^2 - 4x_2 - 2) = 0\). By “branching” on \(x_2 = 0\) in the first of these, and then sub-branching on \(x_1 = 0\) in the second, we get four cases:

A. \(x_2 = 0\): The second equation becomes \(x_1(x_1^2 - 2) = 0\).
   1. \(x_1 = 0\): We have \(\bar{x} = (0, 0)\).
   2. \(x_1 \neq 0\): The second equation implies us \(x_1^2 - 2 = 0\), so \(x_1 = \pm \sqrt{2}\). We have \(\bar{x} = (\pm \sqrt{2}, 0)\).

B. \(x_2 \neq 0\): The first equation implies \(3x_1^2 - 2x_2 - 2 = 0\).
   1. \(x_1 = 0\): The first equation becomes \(-2x_2 - 2 = 0, x_2 = -1\). We have \(\bar{x} = (0, -1)\).
   2. \(x_1 \neq 0\): The second equation implies \(x_1^2 - 4x_2 - 2 = 0\). We can combine this with \(3x_1^2 - 2x_2 - 2 = 0\) to get \(x_1^2 = 2/5\) and \(x_2 = -2/5\); in other words, \(\bar{x} = (\pm \sqrt{2/5}, -2/5)\).

The candidates for local optima are: \((\pm \sqrt{2}, 0), (0, 0), (0, -1),\) and \((\pm \sqrt{2/5}, -2/5)\).  

### 2.2 Sufficient Conditions for Optimality — a First Look

Once we’ve found candidates for local optimizers, we examine each to decide if it is a maximizer, minimizer, or neither. Candidates are screened by means of “sufficient conditions.”

Again, we start with the one-dimensional case.

**Theorem 2.2.1** (First-derivative test for sufficiency at endpoints). *If \(\bar{x}\) is the right-hand endpoint of the interval \(I \subseteq \mathbb{R}\) and \(f'(\bar{x}) < 0\), then \(\bar{x}\) is a strict local minimizer for \(f\) on \(I\). Similarly, if \(f'(\bar{x}) > 0\) and \(\bar{x}\) is the left-hand endpoint of the interval \(I\), then \(\bar{x}\) is a strict local minimizer for \(f\) on \(I\).*

**Proof.** We consider only the first statement. Let \(\alpha = f'(\bar{x})/2\), so that \(f'(\bar{x}) < \alpha < 0\). By the definition of derivative as a limit, this implies the existence of \(\delta > 0\) so that

\[
\frac{f(x) - f(\bar{x})}{x - \bar{x}} < \alpha
\]

for all \(x \in (\bar{x} - \delta, \bar{x})\). Noting that \(x - \bar{x} < 0\) on this interval, we can algebraically rearrange the above inequality to obtain \(f(x) > f(\bar{x}) + \alpha(x - \bar{x}) > f(\bar{x})\). Thus, \(\bar{x}\) is a strict minimizer for \(f\) on the interval \((\bar{x} - \delta, \bar{x})\), and so it is a strict local minimizer for \(f\) on \(I\).  

The next result shows that the sign of the first derivative can prove sufficiency over an interval; it’s also useful in proving second-derivative tests for sufficiency.
**Lemma 2.2.2** (Monotonicity and strict monotonicity). Consider an interval $I \subseteq \mathbb{R}$ and a differentiable function $f : I \to \mathbb{R}$.

(a) If $f'(x) \geq 0$ for all points $x \in I$, then $f$ is nondecreasing on $I$: $f(x) \leq f(y)$ for all $x, y \in I$ with $x \leq y$.

(b) If $f'(x) > 0$ for all but finitely many points $x \in I$, then $f$ is strictly increasing on $I$: $f(x) < f(y)$ for all $x, y \in I$ with $x < y$.

**Proof.** We leave (a) as an exercise and prove only (b). Suppose that the finitely many “exceptional” points with $f'(x) \leq 0$ are listed in order as $x^{(1)} < x^{(2)} < \ldots < x^{(k)}$. If $x^{(i)} \leq x < y \leq x^{(i+1)}$, then there exists $z \in (x, y)$ with $f(y) = f(x) + f'(z)(y-x) > f(x)$, where the inequality follows from $f'(z) > 0$ and $x < y$. This shows that $f$ is strictly increasing on each sub-interval $[x^{(i)}, x^{(i+1)}]$, which in turn implies that $f$ is strictly increasing on $[x^{(1)}, x^{(k)}]$. A similar argument shows that $f$ is strictly increasing on $(-\infty, x^{(1)}] \cap I$ and on $I \cap [x^{(k)}, \infty)$. Hence $f$ is strictly increasing on all of $I$, as claimed. □

**Theorem 2.2.3** (Second-derivative test for sufficiency on intervals). Consider a critical point $\bar{x} \in I$ of $f$ and suppose that $f''$ exists on the entire interval $I \subseteq \mathbb{R}$.

(a) If $f''(x) \geq 0$ for all $x \in I$, then $\bar{x}$ is a global minimizer for $f$ on $I$.

(b) If $f''(x) > 0$ for all (except for finitely many) $x \in I$ with $x \neq \bar{x}$, then $\bar{x}$ is a strict global minimizer for $f$ on $I$.

(c) If $f''(\bar{x}) > 0$ and $f''$ is continuous at $\bar{x}$, then $\bar{x}$ is a strict local minimizer for $f$ on $I$.

**Proof.** We use the Taylor formula: if $x \neq \bar{x}$, then there exists $z$ strictly between $x$ and $\bar{x}$ such that $f(x) = f(\bar{x}) + f'(\bar{x})(x-\bar{x}) + \frac{1}{2} f''(z)(x-\bar{x})^2$. Because $\bar{x}$ is a critical point, this becomes $f(x) = f(\bar{x}) + \frac{1}{2} f''(z)(x-\bar{x})^2$.

(a) Here we have $f(x) = f(\bar{x}) + \frac{1}{2} f''(z)(x-\bar{x})^2 \geq f(\bar{x}) + 0$.

(b) If $f''(x) > 0$ for all $x \in I$, then $f''(z) > 0$: in this case, we use the same argument as in (a), but with a strict inequality. The case in which we omit finitely many points is somewhat more complicated and will be proved below.

(c) The continuity of $f''$ implies that $f''(x) > 0$ on some interval $(\bar{x} - \delta, \bar{x} + \delta)$ about $\bar{x}$, so we may restrict attention to this smaller interval and apply part (b).

Now we consider part (b) in its full generality. By Lemma 2.2.2(b), we see that $f'$ is strictly increasing on $I$. Consider any $x \in I$ with $x \neq \bar{x}$, so that there exists $z$ strictly between $x$ and $\bar{x}$ with $f(x) = f(\bar{x}) + f'(\bar{x})(x-\bar{x})$. If $x < z < \bar{x}$, then $f'(z) < f'(\bar{x}) = 0$ and $x-\bar{x} < 0$; on the other hand, if $x > z > \bar{x}$, then $f'(z) > f'(\bar{x}) = 0$ and $x-\bar{x} > 0$. In either case, we have $f'(z)(x-\bar{x}) > 0$, so $f(x) = f(\bar{x}) + f'(z)(x-\bar{x}) > f(\bar{x})$. Hence $\bar{x}$ is a strict global minimizer on $I$. □
Example 2.2.4 (Using local sufficient conditions). Consider \( f(x) = 2x^3 + 3x^2 - 36x + 5 \) on the interval \([-4, 4]\). Find all candidates for optimality and characterize them.

Solution.

Derivatives: \( f'(x) = 6x^2 + 6x - 36, f''(x) = 12x + 6 \)

Necessary conditions:

Endpoints: \( x = -4 \) and \( x = 4 \)
Critical points: \( f'(x) = 0 \) at \( x = 2 \) and \( x = -3 \)

Sufficient conditions:

Endpoints: \( f'(-4) = 36 \Rightarrow \text{strict local minimizer} \) \( f'(4) = 84 \Rightarrow \text{strict local maximizer} \)
Critical points: \( f''(-3) = -30 \Rightarrow \text{strict local maximizer} \) \( f''(2) = 30 \Rightarrow \text{strict local minimizer} \)

The global second-derivative tests are inconclusive, so we must use some other reasoning to determine global optimality. Here is one possible argument:

Note that \( f'(x) = 6(x - 3)(x + 3) > 0 \) for \( x \in [-4, -3) \) and \( x \in (2, 4] \), and \( f'(x) < 0 \) for \( x \in (-3, 2) \). Hence \( f(x) \) is strictly increasing on \([-4, -3] \) and on \([2, 4]\), but strictly decreasing on \([-3, 2]\). It therefore suffices to compare the objective values at the four candidate points: \( f(-4) = 69, \ f(-3) = 86, \ f(2) = -39, \ f(4) = 37 \).

Conclusion: \( x = -3 \) is the strict global maximizer and \( x = 2 \) is the strict global minimizer.

Chapter 3 provides criteria allowing one to compare function values at candidates. □

Example 2.2.5 (Using global sufficient conditions). Find and characterize all candidates for optimality of \( f(x) = x^4 \) on \( \mathbb{R} \).

Solution. The only critical point is \( \bar{x} = 0 \), so there can be at most one local or global optimizer. Because \( f''(0) = 0 \), the local sufficiency test (Theorem 2.2.3(c)) tells us nothing. However, we note that \( f''(x) > 0 \) for all \( x \neq 0 \), so the global sufficiency test (Theorem 2.2.3(b)) tells us that \( \bar{x} = 0 \) is a strict global minimizer. Finally, we note that \( \lim_{x \to \infty} f(x) = +\infty \), which also rules out global maximizers. □

In the preceding section, we extended the one-dimensional critical-point necessary condition to higher dimensions by restricting \( f \) to lines through the “test point” \( \bar{x} \). In that case, we only needed to consider lines parallel to the coordinate axes. The same idea can be used to build a multidimensional sufficient condition, except that lines parallel to the coordinate axes don’t provide a rich enough set of directions to guarantee sufficiency. Below, we develop a basic second-derivative test for sufficiency in multivariable problems.

Consider a function \( f: \mathbb{R}^n \to \mathbb{R} \) and a line \( x = \bar{x} + tv \) (for \( t \in \mathbb{R} \)) passing through a given point \( \bar{x} \) in the direction of a given vector \( v \); see Figure 2.1. Holding \( \bar{x} \) fixed, we define the \( v \)-restriction of \( f \) to be the one-variable function \( \varphi_v(t) = f(\bar{x} + tv) \). Note that \( t = 0 \) corresponds to \( x = \bar{x} \) and that \( \varphi_v(0) = f(\bar{x}) \).
By the one-dimensional Taylor expansion, we have

$$\varphi_v(t) = \varphi_v(0) + \varphi_v'(0)t + \frac{1}{2}\varphi_v''(s)t^2,$$

for some $s$ strictly between 0 and $t$. By applying the chain rule to the composition of $f$ with $t \mapsto \bar{x} + tv$, we obtain the following expressions for the derivatives of $\varphi_v$:

$$\varphi_v'(t) = f'(\bar{x} + tv)v = \nabla f(\bar{x} + tv) \cdot v = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\bar{x} + tv)v_i,$$

$$\varphi_v''(t) = \sum_{i=1}^{n} \left[ \nabla \left( \frac{\partial f}{\partial x_i} \right)(\bar{x} + tv) \cdot v \right] v_i.$$

In these terms, the critical-point necessary condition for local optimality of $\bar{x}$ can be written

$$\varphi_v'(0) = \nabla f(\bar{x}) \cdot v = 0.$$

To derive a sufficient condition, we rewrite $\varphi_v''(t)$ as

$$\varphi_v''(t) = \left[ v \cdot \nabla \left( \frac{\partial f}{\partial x_1} \right), \ldots, v \cdot \nabla \left( \frac{\partial f}{\partial x_n} \right) \right] \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = v \cdot \left[ \nabla \left( \frac{\partial f}{\partial x_1} \right), \ldots, \nabla \left( \frac{\partial f}{\partial x_n} \right) \right] v.$$

Hence, $\varphi_v''(t) = v \cdot [Hf(\bar{x})]v$, where $Hf(x) = [(\nabla f(x))]^T$ denotes the Hessian matrix of mixed second partials of $f$:

$$Hf(x) = \begin{bmatrix}
\frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\
\frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}
\end{bmatrix}.$$

We recall the following fact from Calculus 3.
**Theorem 2.2.6** (Clairaut). If the second mixed partials of $f$ are continuous at $\bar{x}$, then

$$\frac{\partial^2 f(\bar{x})}{\partial x_i \partial x_j} = \frac{\partial^2 f(\bar{x})}{\partial x_j \partial x_i}.$$  

In other words, the Hessian $Hf(\bar{x})$ is a symmetric matrix.

**Example 2.2.7** (A simple Hessian calculation). Consider $f(x) = x_1^3 x_2 - 2x_1x_2^2 - 2x_1x_2$. In the preceding section we used the gradient $\nabla f(\bar{x}) = \begin{bmatrix} 3x_1^2x_2 - 2x_2^2 - 2x_2 \\ x_1^3 - 4x_1x_2 - 2x_1 \end{bmatrix}$ to identify the candidates for optimality of $f$. The Hessian for this function is

$$Hf(x) = \begin{bmatrix} 6x_1x_2 & 3x_1^2 - 4x_2 - 2 \\ 3x_1^2 - 4x_2 - 2 & -4x_1 \end{bmatrix}.$$  

Observe that this matrix is symmetric, as promised by the Clairaut Theorem.

We now return to our discussion of optimizing $f$ on the line through $\bar{x}$. To summarize, we defined $\varphi_v(t) = f(\bar{x} + tv)$ and then calculated that

$$\varphi_v'(t) = \nabla f(\bar{x} + tv) \cdot v, \quad \varphi_v''(t) = v \cdot Hf(\bar{x} + tv)v.$$  

Applying the one-dimensional second-derivative test to $\varphi_v$, we immediately obtain a rough second-derivative test for higher dimensions: $\bar{x}$ is a global minimizer for $f$ if $\varphi_v'(0) = 0$ for all $v$ and $\varphi_v''(t) \geq 0$ for each $v$ and $t$. In other words, we have proved the following.

**Theorem 2.2.8** (Global second-derivative test for sufficiency in $\mathbb{R}^n$). Suppose that $\bar{x}$ is a critical point for $f$ and assume that $Hf$ is continuous.

(a) If $v \cdot Hf(x)v \geq 0$ for all $v$ and $x$, then $\bar{x}$ is a global minimizer for $f$.

(b) If $v \cdot Hf(x)v > 0$ for all $v \neq 0$ and all $x \neq \bar{x}$, then $\bar{x}$ is a strict global minimizer for $f$.

A local second-derivative test is somewhat harder and will be presented later in this chapter. In the meantime, we have a more obvious difficulty to deal with: how can we verify the conditions on $Hf$ needed in these tests? That is the subject of the next section.

### 2.3 Definiteness and Quadratic Forms

The second-derivative test in the previous section is only useful if we can test the following condition on a square matrix $A$: Is it true that $v \cdot Av \geq 0$ for all vectors $v$?

**Definition 2.3.1** (Positive definite and semidefinite matrices).

(a) We say that $A$ is **positive definite** (PD) if $x \cdot Ax > 0$ for all $x \neq 0$.

(b) We say that $A$ is **positive semidefinite** (PSD) if $x \cdot Ax \geq 0$ for all $x$.

In other words, $A$ is positive definite if and only if $x = 0$ is the strict global minimizer of the function $x \cdot Ax$. We can rephrase the global second-derivative test as follows.
Restatement of Theorem 2.2.8 (Global second-derivative test for sufficiency in $\mathbb{R}^n$). Suppose that $\bar{x}$ is a critical point for $f$ and assume that $Hf$ is continuous.

(a) If $Hf(x)$ is positive semidefinite for all $x$, then $\bar{x}$ is a global minimizer for $f$.
(b) If $Hf(x)$ is positive definite for all $x \neq \bar{x}$, then $\bar{x}$ is a strict global minimizer for $f$.

Here are some further definitions that are also useful for testing optimality.

Definition 2.3.2 (Negative definite and semidefinite matrices; indefinite matrices).

(a) We say that $A$ is negative definite (ND) if $x \cdot Ax < 0$ for all $x \neq 0$.  
(b) We say that $A$ is negative semidefinite (NSD) if $x \cdot Ax \leq 0$ for all $x$.  
(c) We say that $A$ is indefinite if $x \cdot Ax > 0 > y \cdot Ay$ for some choice of $x$ and $y$.

Clearly, negative definite and negative semidefinite Hessians provide sufficient conditions for validating maximizers. Less obvious, however, is the use of indefinite matrices. It turns out that an indefinite Hessian refutes a candidate’s optimality. More will be said about this in the next section.

To ease the discussion of practical tests of definiteness, we introduce the following concept, which will be useful in later chapters as well.

Definition 2.3.3 (Quadratic forms). A quadratic form is a function expressible as $Q_A(x) = x \cdot Ax$ for some matrix $A$, which is called the coefficient matrix.

Quadratic forms are the multidimensional generalizations of simple quadratic functions $f(x) = ax^2$ on $\mathbb{R}^1$. In the one-dimensional case, the graph of a quadratic function “opens upward” when $a > 0$ and “opens downward” when $a < 0$. Quadratic forms with positive and negative definite matrices are the multidimensional analogues of these.

Example 2.3.4 (Expanding a quadratic form into a sum of degree-2 terms). Consider the matrix

\[
A = \begin{bmatrix}
-1 & 2 & 0 \\
2 & 1 & 3 \\
0 & 3 & 4
\end{bmatrix}.
\]

This defines a quadratic form on $\mathbb{R}^3$ by the formula

\[
Q_A(x) = x \cdot Ax = x_1^2 + 2(2x_1x_2) + 2(0x_1x_3) + x_2^2 + 2(3x_2x_3) + 4x_3^2,
\]

which is a polynomial of degree two in three variables.

Example 2.3.5 (Packing a sum of degree-2 terms into a quadratic form). Consider the function on $\mathbb{R}^3$ defined by $f(x) = f(x_1, x_2, x_3) = 10x_1^2 - 3x_1x_2 + 2x_1x_3 - x_2^2 + x_2x_3 + 5x_3^2$. We can rewrite $f$ as $f = Q_A$ by choosing the coefficient matrix $A$ to be

\[
A = \begin{bmatrix}
10 & -3/2 & 1 \\
-3/2 & -1 & 1/2 \\
1 & 1/2 & 5
\end{bmatrix}.
\]
Observations 2.3.6 (Facts about quadratic forms).

(a) There is a simple correspondence between the entries in the coefficient matrix and the coefficients of the degree-2 polynomial: the diagonal entries are the coefficients of the squared terms, and off-diagonal entries are the coefficients of the mixed terms.

(b) It is always possible to choose the matrix $A$ to be a symmetric matrix. If $A$ is not already symmetric, then we can replace $A$ by its symmetric part $\frac{1}{2}(A + A^T)$. It is easily verified that $x \cdot Ax = x \cdot \frac{1}{2}(A + A^T)x$ and also that $A + A^T$ is symmetric.

(c) We will make frequent use of the fact that $\nabla Q_A(x) = 2Ax$ and $HQ_A(x) = 2A$ when $A$ is symmetric. These formulas directly generalize the one-dimensional versions.

We now develop a general test of definiteness, starting from two small examples.

Example 2.3.7 (Completing the square).

(a) Consider $A = \begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix}$. We write out $Q_A(x)$ and complete the square:

$$x \cdot Ax = 2x_1^2 - 2x_1x_2 + 3x_2^2 = \left[2x_1^2 - 2\left(\sqrt{2}x_1\right)\left(\frac{1}{\sqrt{2}}x_2\right) + \frac{1}{2}x_2^2\right] - \left[\frac{1}{2}x_2^2\right] + 3x_2^2$$

$$= \left(\sqrt{2}x_1 - \frac{1}{\sqrt{2}}x_2\right)^2 + \frac{5}{2}x_2^2.$$

Clearly, the right-hand side is nonnegative and equals zero only if $x = 0$. Therefore $A$ is positive definite.

(b) Consider $A = \begin{bmatrix} 1 & 2 \\ 2 & 2 \end{bmatrix}$. Again, we write out $Q_A(x)$ and complete the square:

$$x \cdot Ax = x_1^2 + 4x_1x_2 + 2x_2^2 = \left[x_1^2 + 2(x_1)(2x_2) + 4x_2^2\right] - \left[4x_2^2\right] + 2x_2^2$$

$$= (x_1 + 2x_2)^2 - 2x_2^2.$$

From the right-hand side we see that $x = (1, 0)$ gives $Q_A(x) = 1 > 0$, while $x = (2, -1)$ gives $Q_A(x) = -2 < 0$. Therefore, $A$ is indefinite.

In Example 2.3.7(a), we essentially extracted a representation of the form

$$x \cdot Ax = \left\| \begin{bmatrix} \sqrt{2}x_1 - \frac{x_2}{\sqrt{2}} \\ \sqrt{5}/2x_2 \end{bmatrix} \right\|^2 = \|Ux\|^2,$$

where $U$ is an upper triangular matrix given by

$$U = \begin{bmatrix} \sqrt{2} & -1/\sqrt{2} \\ 0 & \sqrt{5}/2 \end{bmatrix}.$$

In fact, we can write this as $x \cdot Ax = (Ux) \cdot (Ux) = x \cdot (U^TU)x$. Moreover, it is easily verified that $A = U^TU$. This motivates the next two results.
Proposition 2.3.8 (Definiteness via symmetric factorization). If $B$ is a $k \times n$ matrix, then $B^T B$ is positive semidefinite. Moreover, $B^T B$ is positive definite if and only if $B$ has full column rank (i.e., the columns of $B$ are linearly independent).

Proof. Exercise.

Theorem 2.3.9 (Definiteness via Cholesky factors and principal minors). Consider a symmetric matrix $A$. The following conditions are equivalent:

(a) $A$ is positive definite;
(b) $A = U^T U$ for some upper triangular matrix $U$ with positive diagonal entries;
(c) all leading principal minors of $A$ are positive.

The statement in part (c) of Theorem 2.3.9 requires an additional definition.

Definition 2.3.10 (Principal minors). The leading principal minors of a square matrix $A$ are the determinants of the upper-left square submatrices of $A$.

Example 2.3.11 (Principal minors of a $4 \times 4$ matrix). The leading principal minors of the $4 \times 4$ matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

are the four determinants

$$a_{11}, \quad \det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad \det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \quad \det \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}.$$

The criterion provided by Theorem 2.3.9(c) is useful for the $2 \times 2$ case, because $2 \times 2$ determinants are very easy to calculate. A complete characterization of definiteness in this case is shown in Figure 2.2; the verification of the procedure is given in the Appendix.

<table>
<thead>
<tr>
<th>PD</th>
<th>ND</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_{11}$</td>
<td>$a_{22}$</td>
</tr>
<tr>
<td>$+$</td>
<td>$+$</td>
</tr>
<tr>
<td>$-$</td>
<td>$-$</td>
</tr>
</tbody>
</table>

Anything else is indefinite.

Figure 2.2
Example 2.3.12 (Testing the definiteness of $2 \times 2$ symmetric matrices).

(a) $A = \begin{bmatrix} 2 & 0 \\ 0 & 1/2 \end{bmatrix}$: $A$ is PD.
(b) $A = \begin{bmatrix} -1 & 2 \\ 2 & 3 \end{bmatrix}$ has mixed signs on diagonal: $A$ is indefinite.
(c) $A = \begin{bmatrix} -2 & 1 \\ 1 & -3 \end{bmatrix}$ has $\det A = 5$: $A$ is ND.
(d) $A = \begin{bmatrix} -2 & 3 \\ 3 & -2 \end{bmatrix}$ has $\det A = -5$: $A$ is indefinite.
(e) $A = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$ has $\det A = 0$: $A$ is PSD but not PD.

Proof of Theorem 2.3.9 ($2 \times 2$ case). First we prove that (a)$\Rightarrow$(c). Assume that $A$ is positive definite and consider the vectors $x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $y = \begin{bmatrix} -a_{12} \\ a_{11} \end{bmatrix}$.
Then $Q_A(x) = a_{11} > 0$ and $Q_A(y) = a_{11} \det A > 0$, so $\det A > 0$. Hence (c) holds.

Next we prove that (c)$\Rightarrow$(b). Assume that $a_{11} > 0$ and $\det A > 0$, and define
$$U = \begin{bmatrix} \sqrt{a_{11}} & a_{12}/\sqrt{a_{11}} \\ 0 & \sqrt{\det A}/\sqrt{a_{11}} \end{bmatrix}.$$ Then $A = U^T U$, so (b) holds.

Finally, for (b)$\Rightarrow$(a), we apply Proposition 2.3.8.

It turns out that higher dimensional determinants are not as easy to work with as the “$A = U^T U$” criterion given in part (b) in Theorem 2.3.9. The upper triangular matrix $U$ is called a Cholesky factor of $A$. When $A$ is positive definite, we can calculate $U$ by a restricted version of the row-reduction procedure from Linear Algebra. Combined with the following result, this gives a complete test for definiteness (see Figure 2.3).

Theorem 2.3.13 (Diagonal test for indefiniteness). A symmetric matrix is indefinite if either of the following holds:

(a) there are both positive and negative entries on the diagonal; or
(b) there is a row with a zero diagonal entry and a nonzero off-diagonal entry.

Proof. Let $e^{(i)}$ denote the $i$th coordinate vector. First we note that $e^{(i)} \cdot Ae^{(j)} = a_{ij}$. In this way, the coordinate vectors allow us to extract specified entries in $A$. Now suppose that condition (a) of the theorem holds: $a_{ii} > 0 > a_{jj}$ for some $i$ and $j$. Then we have $Q_A(e^{(i)}) = e^{(i)} \cdot Ae^{(j)} = a_{ii} > 0 > a_{jj} = e^{(j)} \cdot Ae^{(j)} = Q_A(e^{(j)})$. 

23
so $A$ is indefinite. Next, suppose that condition (b) holds: $a_{ii} = 0 \neq a_{ij}$ for some $i$ and $j$. Then we have

$$Q_A(\alpha e^{(i)} + e^{(j)}) = \left( \alpha e^{(i)} + e^{(j)} \right) \cdot A \left( \alpha e^{(i)} + e^{(j)} \right)$$

$$= \alpha^2 a_{ii} + 2\alpha a_{ij} + a_{jj} = \alpha(2a_{ij}) + a_{jj}.$$

This shows that we can make $Q_A(\alpha e^{(i)} + e^{(j)})$ positive or negative by choosing $\alpha$ to be a sufficiently large positive or negative number. Therefore, $A$ is indefinite.

---

**Testing an $n \times n$ Symmetric Matrix for Definiteness**

- Apply the row-reduction algorithm, but *without* row exchanges and row multiplies.
- At each step, check the diagonal test for indefiniteness. If it holds, then stop reducing: the matrix is indefinite.
- Whenever row-reduction can be completed without detecting indefiniteness, it yields an upper triangular matrix $R$. In this case, the original matrix is:

  - **PD** $\iff$ all diagonal entries $r_{ii} > 0$,
  - **PSD** $\iff$ all diagonal entries $r_{ii} \geq 0$,
  - **ND** $\iff$ all diagonal entries $r_{ii} < 0$,
  - **NSD** $\iff$ all diagonal entries $r_{ii} \leq 0$.

**Figure 2.3**

When the test in Figure 2.3 succeeds in demonstrating positive (semi)definiteness, the Cholesky factor $U$ can be obtained from $R$ by dividing each row of $R$ by the square-root of the diagonal entry in that row. The procedure is justified in the Appendix.

**Example 2.3.14** (Testing the definiteness of $n \times n$ symmetric matrices).

(a) $A = \begin{bmatrix} 4 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 16 \end{bmatrix}$: $A$ is indefinite.

(b) $A = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -3 \end{bmatrix}$: $A$ is NSD but not ND.

(c) $A = \begin{bmatrix} 3 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 1 & 3 \end{bmatrix} \rightarrow \begin{bmatrix} 3 & 1 & 2 \\ 0 & 5/3 & 1/3 \\ 0 & 1/3 & 5/3 \end{bmatrix} \rightarrow \begin{bmatrix} 3 & 1 & 2 \\ 0 & 5/3 & 1/3 \\ 0 & 0 & 24/15 \end{bmatrix}$: $A$ is PD.

(d) $A = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 2 \\ 2 & 2 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 & 2 \\ 0 & 0 & 0 \\ 0 & 0 & -3 \end{bmatrix}$: $A$ is indefinite.
Here are some useful facts for demonstrating the definiteness of a Hessian matrix.

**Proposition 2.3.15** (More facts about definiteness).

(a) If $A$ and $B$ are positive semidefinite, then so is $A + B$. Moreover, if $A$ is actually positive definite, then $A + B$ is positive definite.

(b) Suppose $A$ is positive semidefinite. Then $A$ is positive definite if and only if $A$ is invertible.

(c) If $A$ is positive semidefinite, then $B^TAB$ is positive semidefinite. Moreover, if $A$ is positive definite, then $B^TAB$ is positive definite if and only if $B$ has full column rank.

(d) If $A$ is positive definite, then there exists $\delta > 0$ so that $B$ is positive definite whenever $|a_{ij} - b_{ij}| < \delta$ for all $i, j$.

(e) Suppose $A$ is symmetric. Then $A$ is positive definite if and only if all eigenvalues of $A$ are positive; $A$ is positive semidefinite if and only if all eigenvalues of $A$ are nonnegative; and $A$ is indefinite if and only if the eigenvalues of $A$ have mixed signs.

**Proof.** We leave (a), (b), and (c) as exercises. Statement (d) follows from the continuity of the determinant of a matrix with respect to its entries, along with the characterization of positive definiteness by leading principal minors (Theorem 2.3.9). A determinant-free proof is given in the appendix.

The proof of (e) relies on the “spectral theorem” of linear algebra, which says that $A = QDQ^T$ for some invertible matrix $Q$ and diagonal matrix $D$. Hence, the definiteness of $A$ is equivalent to that of $D$. Moreover, the columns of $Q$ comprise a basis of eigenvectors for $A$, with eigenvalues given by the diagonal entries of $D$. Appendix A.3 gives a proof of the spectral theorem using concepts from Chapter 3.  

2.4 Back to Optimization — Sufficient Conditions

The methods of the previous section help us apply our sufficient conditions for optimality.

**Example 2.4.1.** Find and characterize all candidates for optimality of the function $f(x_1, x_2) = x_1^2 - x_1 x_2 + 2x_2^2 - 2x_1 + e^{x_1} e^{x_2}$ on $\mathbb{R}^2$.

**Solution.** The gradient of $f$ is

$$\nabla f(x) = \begin{bmatrix} 2x_1 - x_2 - 2 + e^{x_1 + x_2} \\ -x_1 + 4x_2 + e^{x_1 + x_2} \end{bmatrix},$$

which is zero only for the vector $\bar{x} \approx (0.3304494983, -0.201730301)$.

Idea of this derivation: Subtracting the two equations gives $x_2 = (3x_1 - 2)/5$; substituting this back into either gradient equation yields $7x_1 - 8 + \exp[(8x_1 - 2)/5] = 0$. The
roots of this equation can be isolated and approximated by the methods of Calculus, such as the Intermediate Value Theorem, Rolle’s Theorem, and Newton’s Method.

Next, the Hessian

\[ Hf(x) = \begin{bmatrix} 2 + e^{x_1+x_2} & -1 + e^{x_1+x_2} \\ -1 + e^{x_1+x_2} & 4 + e^{x_1+x_2} \end{bmatrix} \]

is positive definite for all \( x \), so \( \bar{x} \) is the strict global minimizer for \( f \).

Idea of this verification: We can split \( Hf(x) \) into a sum of simpler matrices, namely,

\[ Hf(x) = \begin{bmatrix} 2 & -1 \\ -1 & 4 \end{bmatrix} + e^{x_1+x_2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \cdot \]

The first matrix is PD, the second is PSD, and the exponential term is positive. Therefore, \( Hf(x) \) is the sum of a PD matrix and a PSD matrix, so \( Hf(x) \) is PD. Alternatively, the leading principal minors of \( Hf(x) \) are both positive for all \( x \).

**Theorem 2.4.2** (Local second-derivative test for sufficiency in \( \mathbb{R}^n \)). Suppose \( f : \mathbb{R}^n \to \mathbb{R} \) has continuous second partial derivatives at a critical point \( \bar{x} \).

(a) If \( Hf(\bar{x}) \) is positive definite, then \( \bar{x} \) is a strict local minimizer for \( f \).

(b) If \( Hf(\bar{x}) \) is negative definite, then \( \bar{x} \) is a strict local maximizer for \( f \).

(c) If \( Hf(\bar{x}) \) is indefinite, then \( \bar{x} \) is neither a local minimizer nor a local maximizer.

**Proof.** Parts (a) and (b) follow from Proposition 2.3.15(d), along with the observation that \( \bar{x} \) is a local minimizer for \( f \) if and only if there exists \( \delta > 0 \) so that \( \bar{t} = 0 \) is a global minimizer on \([-1, 1] \) of \( \varphi_v(t) = f(\bar{x} + tv) \) for all \( v \) with \( \|v\| \leq \delta \). For part (c), we note that the indefiniteness gives us vectors \( u \) and \( v \) with

\[ u \cdot Hf(\bar{x})u > 0 > v \cdot Hf(\bar{x})v. \]

By the chain rule, \( \bar{t} = 0 \) is a strict local minimizer for \( \varphi_u \) and a strict local maximizer for \( \varphi_v \), so \( \bar{x} \) is neither a local maximizer nor a local minimizer for \( f \).

**Definition 2.4.3** (Saddle point). A critical point that is neither a local minimizer nor a local maximizer is called a saddle point.

**Observation 2.4.4** (Failure of local sufficiency test). If \( Hf(\bar{x}) \) is merely semidefinite (but not definite) at the critical point \( \bar{x} \), then the local second-derivative test is inconclusive. Determining whether \( \bar{x} \) is indeed a local optimum requires additional work and perhaps some ingenuity.
Example 2.4.5 (Using local sufficient conditions). Find and characterize all candidates for optimality of the following functions.

(a) \(f(x_1, x_2) = x_1^3 - 2x_1x_2 + x_2^2\). Here we have

\[
\nabla f(x) = \begin{bmatrix} 3x_1^2 - 2x_2 \\ -2x_1 + 2x_2 \end{bmatrix}, \quad Hf(x) = \begin{bmatrix} 6x_1 & -2 \\ -2 & 2 \end{bmatrix}.
\]

Critical points: \(\bar{x} = 0\) and \(\bar{x} = \begin{bmatrix} 2/3 \\ 2/3 \end{bmatrix}\).

Sufficient conditions:

- \(Hf(0, 0) = \begin{bmatrix} 0 & -2 \\ -2 & 2 \end{bmatrix}\) is indefinite \(\Rightarrow\) saddle point.
- \(Hf(2/3, 2/3) = \begin{bmatrix} 4 & -2 \\ -2 & 2 \end{bmatrix}\) is PD \(\Rightarrow\) strict local minimizer.

There are no global maximizers or global minimizers because

\[
\lim_{x_1 \to \infty} f(x_1, 0) = \infty, \quad \lim_{x_1 \to -\infty} f(x_1, 0) = -\infty.
\]

(b) \(f(x_1, x_2) = (x_2 - x_1^2)^2 - x_1^6\). Here we have

\[
\nabla f(x) = \begin{bmatrix} 2(x_2 - x_1^2)(-2x_1) - 6x_1^5 \\ 2(x_2 - x_1^2) \end{bmatrix}, \quad Hf(x) = \begin{bmatrix} -4x_2 + 12x_1^2 - 30x_1^4 & -4x_1 \\ -4x_1 & -4x_1 \end{bmatrix}.
\]

Critical points: \(\bar{x} = 0\).

Sufficient conditions: \(Hf(0) = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}\) is PSD — inconclusive!

A closer examination reveals that \(\bar{x} = 0\) is, in fact, a saddle point. This can be verified examining the paths \(x(t) = (0, t)\) and \(x(t) = (t, t^2)\): as \(t\) moves away from 0, \(f(x(t))\) increases without bound along the first path and decreases without bound along the second. \(\square\)
3. EXISTENCE OF GLOBAL OPTIMIZERS

This chapter presents theoretical tools for demonstrating the existence of global minimizers or maximizers. For practical purposes, such tools enable us to verify global optimality simply by comparing the function values of all candidates for optimality.

3.1 Extreme Value Theory

Our first result underlies most of the existence theory in optimization.

**Theorem 3.1.1** (Weierstrass). Suppose that $S \subset \mathbb{R}^n$ is nonempty, closed and bounded. Then every continuous function on $S$ attains a global maximum and a global minimum.

**Proof.** Deferred until later in this section.

We refer the reader to section 1.4 for definitions and examples of the terms “closed,” “bounded,” and “continuous.” The following special case of the Weierstrass theorem is usually introduced in the first Calculus course.

**Corollary 3.1.2** (One-dimensional extreme value theorem). If $f : [a, b] \to \mathbb{R}$ is continuous then $f$ attains a global maximum and a global minimum on $[a, b]$.

The extreme value theorem provides information even in seemingly more general circumstances, as shown by the next result.

**Corollary 3.1.3** (Existence via one-dimensional sublevel boundedness). Let $I$ be any interval in $\mathbb{R}$. Suppose $f : I \to \mathbb{R}$ is continuous and there is some number $c \in I$ so that the following are satisfied (as applicable):

(a) $\lim_{x \to a^+} f(x) > f(c)$, if $a$ is the left endpoint of $I$;
(b) $\lim_{x \to -\infty} f(x) > f(c)$, if $I$ has no left endpoint;
(c) $\lim_{x \to b^-} f(x) > f(c)$, if $b$ is the right endpoint of $I$;
(d) $\lim_{x \to \infty} f(x) > f(c)$, if $I$ has no right endpoint.

Then $f$ has a global minimizer on $I$.

**Sketch of Proof.** A closed and bounded interval is hiding in the definition of the limit! For example,

- “$\lim_{x \to a^+} f(x) = d$” means “$\forall \epsilon > 0, \exists \delta > 0, \forall x \in (a, a + \delta) : |f(x) - d| < \epsilon.$”
- “$\lim_{x \to a^+} f(x) = \infty$” means “$\forall M, \exists \delta > 0, \forall x \in (a, a + \delta) : f(x) > M.$”
- “$\lim_{x \to -\infty} f(x) = \infty$” means “$\forall M, \exists N, \forall x \in (-\infty, N) : f(x) > M.$”
- “$\lim_{x \to -\infty} f(x) = d$” means “$\forall \epsilon > 0, \exists N, \forall x \in (-\infty, N) : |f(x) - d| < \epsilon.$”

Conditions (a)–(b) grant the existence of a left endpoint $\alpha < c$ and conditions (c)–(d) give a right endpoint $\beta > c$ so that all values of $x$ outside of $[\alpha, \beta]$ yield $f(x) > c$. We therefore can restrict our search to this nonempty, closed, bounded interval $[\alpha, \beta]$. 


Example 3.1.4 (Using sublevel boundedness). The function \( f(x) = (x^4 - 6x^3 + 3x^2 - x + 15)/(100 - x^2) \) has a global minimizer on the open interval \((-10, 10)\) because
\[
\lim_{x \to -10^+} f(x) = \infty = \lim_{x \to 10^-} f(x). \quad \square
\]

Observation 3.1.5 (General use of one-dimensional “coercivity”). Combined with the critical point test, knowledge of the existence of a global minimizer in an interval \( I \) allows us to compare the values of \( f \) at the following candidates:

\( \text{(a) critical points of } f \text{ in } I; \)
\( \text{(b) the endpoints (if any) of } I; \)
\( \text{(c) those points in } I \text{ where } f' \text{ does not exist.} \)

The global minimizers for \( f \) on \( I \) are the candidates giving the smallest value for \( f(\bar{x}) \).

The Weierstrass theorem follows from a special defining property of the set of real numbers (as discussed in Chapter 1), which mathematicians take as an axiom:

Every nonempty set of real numbers which is bounded below has a real infimum.

This axiom fails for the rational numbers: the set of all rationals having \( x^2 \geq 2 \) is easily seen to be bounded below by 1, but has no infimum among the rational numbers. In other words, a “gap” occurs at \( x = \sqrt{2} \). The real numbers, by definition, have no such gaps. The proof of the Weierstrass theorem relies on the following reinterpretation of the above infimum property of the real numbers.

Theorem 3.1.6 (Bolzano-Weierstrass). A bounded sequence in \( \mathbb{R}^n \) always has a convergent subsequence.

Sketch of proof. We use induction on \( n \). First consider the base step, where \( n = 1 \). Consider a bounded sequence \( \{x^{(k)}\} \subset \mathbb{R} \). Defining
\[ b^{(i)} = \sup\{x^{(k)}\}_{k > i} \]
and
\[ b = \inf\{b^{(i)}\}_{i=1}^\infty, \]
it can be verified that \( b^{(i)} \geq b^{(i+1)} \) for all \( i \), and consequently that \( b = \lim_{i \to \infty} b^{(i)} \). For each \( i \), define \( k_i \) to be the smallest integer satisfying
\[ k_i \geq i \text{ and } b^{(i)} < x^{(k_i)}. \]
Then \( \{x^{(k_i)}\} \) defines a subsequence of \( \{x^{(k)}\} \) and
\[ b^{(i)} - \frac{1}{i} < x^{(k_i)} \leq b^{(i)}. \]
The left and right hand sides of the inequality both converge to \( b \), so the Squeeze Theorem (from Calculus) guarantees that \( x^{(k_i)} \to b \) also. This completes the base step for \( n = 1 \).
For the inductive step, assume the theorem is true for \( \mathbb{R}^n \) and consider a bounded sequence \( \{(x^{(k)}, y^{(k)})\} \subset \mathbb{R}^{n+1} \). Then \( \{x^{(k)}\} \) is bounded in \( \mathbb{R}^n \), so by the inductive hypothesis it has a convergent subsequence \( \{x^{(k_i)}\} \rightarrow \bar{x} \). Now \( \{y^{(k_i)}\} \) is bounded in \( \mathbb{R} \), so it has a convergent subsequence \( \{y^{(k_{ij})}\} \rightarrow \bar{y} \). Thus \( \{(x^{(k_{ij})}, y^{(k_{ij})})\} \rightarrow (\bar{x}, \bar{y}) \), which completes the proof. \( \square \)

With this result in hand, we are ready to prove the Weierstrass Theorem.

**Proof of Theorem 3.1.1.** Let \( \{x^{(k)}\} \subset S \) be a minimizing sequence, so that

\[
\lim_{k \to \infty} f(x^{(k)}) = \inf \{f(x) \mid x \in S\}.
\]

Since \( S \) is bounded, the Bolzano-Weierstrass theorem yields a converging subsequence \( \{x^{(k_i)}\} \rightarrow \bar{x} \in \mathbb{R}^n \). Because \( S \) is closed, \( \bar{x} \) must belong to \( S \). By the continuity of \( f \) we have

\[
f(\bar{x}) = \lim_{k \to \infty} f(x^{(k)}) = \inf \{f(x) \mid x \in S\}.
\]

Thus \( \bar{x} \) is a global minimizer for \( f \) on \( S \). The existence of a global maximizer is similar. \( \square \)

We now introduce a concept that allows us to make multidimensional arguments similar to the one-dimensional arguments made possible by Corollary 3.1.3.

**Definition 3.1.7 (Coercivity).** A function \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) is said to be coercive when \( \lim_{\|x\| \to \infty} f(x) = \infty \). Equivalently:

- \( \lim_{t \to \infty} \inf_{\|x\| \geq t} f(x) = \infty \); or
- for each \( M \in \mathbb{R} \), the sublevel set \( \{x : f(x) \leq M\} \) is bounded; or
- for each sequence \( \{x^{(k)}\} \subset \mathbb{R}^n \) with \( \|x^{(k)}\| \to \infty \), one has \( f(x^{(k)}) \to \infty \).

(The proof of equivalence is left as an exercise.)

**Theorem 3.1.8 (Existence via coercivity).** If \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) is continuous and coercive, then \( f \) has a global minimizer on \( \mathbb{R}^n \) and it does not have a global maximizer on \( \mathbb{R}^n \).

**Proof.** For \( M = f(0) \), there exists a number \( r \) so that \( f(x) > M \) whenever \( \|x\| > r \); thus, \( S = \{x \mid f(x) \leq M\} \subseteq B(0, r) \). This sublevel \( S \) is therefore closed, bounded and nonempty. By the Weierstrass theorem, \( f \) has a global minimizer \( \bar{x} \) on \( S \). Now \( x \in S \) implies \( f(x) \geq f(\bar{x}) \), whereas \( x \notin S \) implies \( f(x) > f(0) \geq f(\bar{x}) \). Thus \( \bar{x} \) is a global minimizer over all of \( \mathbb{R}^n \). As for global maximizers, we clearly have \( \sup f = \infty \). \( \square \)

**Observation 3.1.9 (Using coercivity).**

(a) To find the global minimizer of a coercive differentiable function on \( \mathbb{R}^n \), we need only compare the function values at all critical points.

(b) If we can prove coercivity of a continuous function \( g \) by some means, then we know that the set \( \{x \mid g(x) \leq r\} \) is closed and bounded for any choice of \( r \). This is useful for proving existence in constrained optimization problems; see Chapter 7.
Example 3.1.10 (Optimizing a coercive function). Find all maximizers and minimizers for \( f(x_1, x_2) = x_1^4 + x_2^4 - 32x_2^2 \).

Solution. We have

\[
\nabla f(x) = \begin{bmatrix} 4x_1^3 \\ 4x_2^3 - 64x_2 \end{bmatrix}, \quad Hf(x) = \begin{bmatrix} 12x_1^2 & 0 \\ 0 & 12x_2^2 - 64 \end{bmatrix}.
\]

- Necessary conditions: the critical points are \( \bar{x} = 0, \bar{x} = (0, \pm 4) \).
- Sufficient conditions: \( Hf(0) \) is NSD and \( Hf(0, \pm 4) \) is PSD — these are inconclusive!
- Existence theory: \( f \) is coercive (see the next section for why), so some critical point must be a global minimizer; also there can be no global maximizers. We may therefore compare the function values at the critical points, giving us \( f(0) = 0 > -256 = f(0, \pm 4) \). Therefore, \( \bar{x} = (0, \pm 4) \) are the global minimizers and each is a strict local minimizer.
- Further analysis shows that \( \bar{x} = 0 \) is a saddle point: \( f(t, 0) = t^4 \) has a strict global minimizer at \( \bar{t} = 0 \), whereas \( f(0, t) = t^4 - 3t^2 \) has a strict local maximizer at \( \bar{t} = 0 \).

3.2 Verifying Coercivity

This section provides various rules for identifying coercive and non-coercive functions.

Proposition 3.2.1 (Coercivity of polynomials).

(a) Constant functions and polynomials of odd degree on \( \mathbb{R}^n \) are never coercive.

(b) A polynomial on \( \mathbb{R} \) is coercive if and only if its degree is even and its leading coefficient is positive.

(c) The quadratic function \( x \cdot Ax + b \cdot x + c \) is coercive if and only if \( A \) is positive definite.

Proof. Parts (a) and (b) are left as exercises. For part (c), let \( f(x) = x \cdot Ax + b \cdot x + c \). First, suppose \( A \) is positive definite. Take \( \delta = \bar{z} \cdot A \bar{z} > 0 \), where \( \bar{z} \) is a global minimizer of the continuous function \( z \cdot Az \) over the closed and bounded set of \( z \) with \( \|z\| = 1 \). Then we have \( x \cdot Ax \geq \delta \|x\|^2 \) for all \( x \). (To see why, simply consider \( x \neq 0 \) and \( z = x/\|x\| \); then \( \|z\| = 1 \), so \( z \cdot Az \geq \delta \).) By the Schwarz inequality, this implies that \( f(x) \geq \delta \|x\|^2 - \|b\| \|x\| + c \), so

\[
\lim_{\|x\| \to \infty} f(x) \geq \lim_{\|x\| \to \infty} \|x\|((\delta \|x\| - \|b\| + c/\|x\|)) = \infty.
\]

Therefore, \( f \) is coercive.

Conversely, suppose \( A \) is not positive definite. Then there exists a vector \( u \neq 0 \) with \( u \cdot Au \leq 0 \). Without loss of generality, we may assume \( b \cdot u \leq 0 \) (otherwise, we simply replace \( u \) by \( -u \)). The sequence \( x^{(k)} = ku \) gives us \( f(x^{(k)}) \leq 0 \) for all \( k \), so \( f \) is not coercive.

Proposition 3.2.2 (Coercivity of combinations of functions). A function \( f : \mathbb{R}^n \to \mathbb{R} \) is coercive if it has any of the following forms:
(a) \( f(x_1, \ldots, x_n) = p_1(x_1) + \cdots + p_n(x_n) + q(x_1, \ldots, x_n) \), where \( q \) is a polynomial of degree \( k \) and the functions \( p_i(t) - at^k \) are coercive on \( \mathbb{R} \) for every \( i \) and for all real \( a \).

Comment: we describe this case by saying that “the polynomial \( q(x_1, \ldots, x_n) \) is dominated by the separably coercive function \( p_1(x_1) + \cdots + p_n(x_n) \).”

(b) \( f(x, y) = g(x) + h(y), \) where \( g : \mathbb{R}^k \to \mathbb{R}, \ h : \mathbb{R}^{n-k} \to \mathbb{R} \) are coercive and bounded below.

More generally: a finite sum is coercive if every variable appears in at least one summand, and each summand is coercive and bounded below in its variables.

(c) \( f(x, y) = \max\{g(x), h(y)\}, \) where \( g : \mathbb{R}^k \to \mathbb{R} \) and \( h : \mathbb{R}^{n-k} \to \mathbb{R} \) are coercive.

More generally: a point-wise maximum is coercive if every variable appears in at least one operand, and each operand is coercive in its own variables.

(d) \( f(x) = \max\{g(x), h(x)\}, \) where \( g \) is coercive.

(e) \( f(x) = g(h(x)), \) where \( g \) is coercive and \( \lim_{\|x\| \to \infty} \|h(x)\| = \infty. \)

Special cases: \( h(x) = \|x\| \) and \( h(x) = Ax \) where \( A \) has full column rank.

(f) \( f(x) = g(x) + h(x), \) where \( g \) is coercive and \( h \) is bounded below.

(g) \( f(x) = g(x) + h(x), \) where \( g \) is coercive and \( h(x) \geq \alpha g(x) \) whenever \( \|x\| > r \), for some fixed \( \alpha > -1 \) and some \( r \in \mathbb{R}. \)

Proof. See appendix. \( \square \)

**Proposition 3.2.3** (Technical approaches to proving coercivity).

(a) A function \( f(x) = g(x) + h(x) \) is coercive if, for all sufficiently large \( t \), one has

(i) \( g(x) \geq G(t) \) whenever \( \|x\| \geq t \), and

(ii) \( h(x) \geq H(t) \) whenever \( \|x\| \leq t \), and

(iii) \( \lim_{t \to \infty} G(t) + H(t) = \infty. \)

(b) If \( g \) is coercive and \( f(x) \geq g(x) \) whenever \( \|x\| \geq r \), then \( f \) is coercive.

Proof. Exercise. \( \square \)

**Example 3.2.4** (Some coercive functions). Each of the following examples is accompanied by a brief explanation of which rules show it is coercive.

(a) \( \|x\| \) is coercive on \( \mathbb{R}^n \)

Direct application of the definition of coercivity.

(b) \( \|x\|^2 \) is coercive on \( \mathbb{R}^n \)

Composition of the form \( g(\|x\|) \), where \( g(t) = t^2 \) has \( \lim_{t \to \infty} g(t) = \infty. \)

(c) \( e^{x_1^2} + e^{x_2^2} - x_1 \) is coercive on \( \mathbb{R}^2 \)

Sum \( \phi_1(x_1) + \phi_2(x_2) \) covering all variables, where \( \phi_1(x_1) = e^{x_1^2} - x_1 \) is coercive on \( \mathbb{R} \) and \( \phi_2(x_2) = e^{x_2^2} \) is coercive on \( \mathbb{R}. \)
(d) \( x_1^2 + x_2^2 - \sin(x_1x_2) \) is coercive on \( \mathbb{R}^2 \\
\) Sum of the coercive function \( \|x\|^2 \) on \( \mathbb{R}^2 \) and bounded function \(-\sin(x_1x_2)\).

(e) \( x_1^4 + x_4^4 - x_1x_2^2 - x_2x_3^2 \) is coercive on \( \mathbb{R}^3 \\
\) Domination of the polynomial \(-x_1x_2^2 - x_2x_3^2\) by the separably coercive function \( x_1^4 + x_4^4 + x_3^4 \) of all variables.

(f) \( x_1^4 + x_4^4 - 32x_2^2 \) is coercive on \( \mathbb{R}^2 \\
\) Sum \( \phi_1(x_1) + \phi_2(x_2) \) covering all variables, where \( \phi_1(x_1) = x_1^4 \) is coercive on \( \mathbb{R} \) and \( \phi_2(x_2) = x_2^4 - 32x_2^2 \) is coercive on \( \mathbb{R} \).

(g) \( x_1^4 + x_4^4 - 3x_1x_2 \) is coercive on \( \mathbb{R}^2 \\
\) Domination of the polynomial \(-3x_1x_2\) by the separably coercive function \( x_1^4 + x_4^4 \) of all variables.

(h) \( e^{x_1^2} + e^{x_2^2} + e^{x_3^2} - x_1^{100} - x_2^{100} - x_3^{100} \) is coercive on \( \mathbb{R}^3 \\
\) Sum \( \phi(x_1) + \phi(x_2) + \phi(x_3) \) covering all variables, where is \( \phi(t) = e^t - t^{100} \) is coercive on \( \mathbb{R} \).

(i) \( x_1^4 + x_4^4 + x_3^4 - 3x_1x_2x_3 - x_2^2 - x_3^2 \) is coercive on \( \mathbb{R}^3 \\
\) Domination of the polynomial \(-3x_1x_2x_3 - x_2^2 - x_3^2\) by the separably coercive function \( x_1^4 + x_4^4 + x_3^4 \) of all variables.

(j) \( x_1^4 - 4x_1x_2 + x_4^4 \) is coercive on \( \mathbb{R}^2 \\
\) Domination of the polynomial \(-4x_1x_2\) by the separably coercive function \( x_1^4 + x_4^4 \) of all variables.

(k) \( x_1^2 + x_2^2 + x_3^2 - \sin(x_1x_2x_3) \) is coercive on \( \mathbb{R}^3 \\
\) Sum of the coercive function \( \|x\|^2 \) on \( \mathbb{R}^3 \) and the bounded function \(-\sin(x_1x_2x_3)\).

(l) \( 3x_1^2 + 2x_2^2 + 2x_3^2 + 2x_1x_2 + 2x_2x_3 + 2x_1x_3 \) is coercive on \( \mathbb{R}^3 \\
\) Quadratic with positive definite Hessian matrix.

(m) \( (2x_1 - x_2)^2 + (x_2 - x_3)^2 + (x_3 - 1)^2 \) is coercive on \( \mathbb{R}^3 \\
\) Quadratic with positive definite Hessian matrix.

(n) \( x_1^4 + x_2^4 - x_1^2 - x_2^2 + 1 \) is coercive on \( \mathbb{R}^2 \\
\) Sum \( \phi_1(x_1) + \phi_2(x_2) \) covering all variables, where is \( \phi_1(x_1) = x_1^4 - x_1^2 \) is coercive on \( \mathbb{R} \) and \( \phi_2(x_2) = x_2^4 - x_2^2 + 1 \) is coercive on \( \mathbb{R} \).

(o) \( x_1^4 + x_4^4 + x_3^2 - 3x_1x_2 - x_3 \) is coercive on \( \mathbb{R}^3 \\
\) Sum \( \phi_1(x_1, x_2) + \phi_2(x_3) \) covering all variables, where \( \phi_2(x_3) = x_3^2 - x_3 \) is coercive on \( \mathbb{R} \), \( \phi_1(x_1, x_2) = x_1^4 + x_4^4 - 3x_1x_2 \) is coercive on \( \mathbb{R}^2 \) (by domination of the polynomial \(-3x_1x_2\) by the separably coercive function \( x_1^4 + x_4^4 \) of both variables), and the lower boundedness of \( \phi_1 \) and \( \phi_2 \) (because both are continuous and coercive).

(p) \( x_1^2 - 4x_1 + 2x_3^2 + 7 \) is coercive on \( \mathbb{R}^2 \\
\) Sum \( \phi_1(x_1) + \phi_2(x_2) \) covering all variables, where is \( \phi_1(x_1) = x_1^2 - 4x_1 + 7 \) is coercive on \( \mathbb{R} \).
coercive on $\mathbb{R}$ and $\phi_2(x_2) = 2x_2^2$ is coercive on $\mathbb{R}$.

(r) $x_1^4 + 16x_1x_2 + x_2^8$ is coercive on $\mathbb{R}^2$

Domination of the polynomial $16x_1x_2$ by the separably coercive function $x_1^4 + x_2^8$ of all variables.

(s) $x_1^4 + x_1x_2 + x_2^2$ is coercive on $\mathbb{R}^2$

Let $f(x) = x_1^4 + x_1x_2 + x_2^2$. Note that we have

$$f(x) \geq \begin{cases} x_1^2 + x_1x_2 + x_2^2 & \text{if } |x_1| \geq 1, \\ x_1^4 - |x_2| + x_2^2 & \text{if } |x_1| < 1. \end{cases}$$

Notice that both of the functions on the right-hand side are coercive on $\mathbb{R}^2$: one is a PD quadratic and the other is the sum of coercive functions covering both variables. Now consider a sequence $\|x^k\| \to \infty$ and note that the sequence consists of two subsequences satisfying either $|x_1| \geq 1$ or $|x_1| < 1$, and that $f(x)$ is bounded below by a coercive function for either subsequence. Thus, the limit over either subsequence, and hence of the entire sequence, must be $\infty$.

(t) $x_1^4 + (3/2)x_1^2x_2 + x_2^2$ is coercive on $\mathbb{R}^2$

Composition $g(h(x))$, where $g(y_1, y_2) = y_1^2 + (3/2)y_1y_2 + y_2^2$ is a quadratic with positive definite Hessian and $h(x_1, x_2) = (x_1^2, x_2)$ satisfies $\lim_{\|x\| \to \infty} \|h(x)\| = \infty$. The latter is true because $\|h(x_1, x_2)\|^2 = x_1^4 + x_2^2 = \max\{x_1^4, x_2^2\}$, which is the maximum of coercive functions covering all the variables.

To show that a function is not coercive, we typically provide an unbounded sequence $\{x^{(k)}\}$ for which the sequence $\{f(x^{(k)})\}$ of function values is bounded above.

**Example 3.2.5 (Some non-coercive functions).** Each of the following functions is accompanied by a sequence demonstrating that it is not coercive.

(a) $x_1^2 + 2x_1x_2 + x_2^2$ is not coercive on $\mathbb{R}^2$: consider $(k, -k)$

(b) $x_1^4 + 4x_1^2x_2 + x_2^2$ is not coercive on $\mathbb{R}^2$: consider $(k, -k^2)$

(c) $x_1^4 + (5/2)x_1^2x_2 + x_2^2$ is not coercive on $\mathbb{R}^2$: consider $(k, -k^2)$

(d) $x_1^4 + x_1^3 + x_3^2 - 3x_1x_2x_3$ is not coercive on $\mathbb{R}^3$: consider $(\sqrt{k}, \sqrt{k}, k)$

(e) $x_1^3 + x_2^3 + x_3^3 - x_1x_2$ is not coercive on $\mathbb{R}^3$: consider $(-k, 0, 0)$

(f) $x_1^4 + x_1^3 + x_2^3 - 7x_1x_2x_3$ is not coercive on $\mathbb{R}^3$: consider $(1, 1, k)$

(g) $x_1^4 + x_1^3 - 2x_1x_2^2$ is not coercive on $\mathbb{R}^3$: consider $(0, 0, k)$

(h) $x_1^3 + x_2^3 - 3x_1 - 12x_2 + 20$ is not coercive on $\mathbb{R}^2$: consider $(-k, 0)$

(i) $x_1^2 + x_2^2 + x_3^2 - 4x_1x_2$ is not coercive on $\mathbb{R}^3$: consider $(k, k, 0)$

(j) $12x_1^3 - 36x_1x_2 - 2x_2^3 + 9x_2^2 - 72x_1 + 60x_2 + 5$ is not coercive on $\mathbb{R}^2$: consider $(-k, 0)$

(k) $e^{-(x_1^2+x_2^2)}$ is not coercive on $\mathbb{R}^2$: consider $(k, 0)$

(l) $x_1^2 - 2x_1x_2 + (1/3)x_2^3 - 4x_2$ is not coercive on $\mathbb{R}^2$: consider $(0, -k)$ \hfill $\square$
Observation 3.2.6 (Testing coercivity on lines may not work). Note that 3.2.5(d) is an example of a function which is "coercive along all lines through the origin" (meaning that $\lim_{t \to \infty} f(tx) = \infty$ for any $x \neq 0$), but which nevertheless fails to be coercive.

Observation 3.2.7 (Effective boundedness via coercivity). One use of coercivity is to show that the minimization of $f$ on $S$ implies minimization over a bounded region. To do this, we combine the constraints and the inequality $f(x) \leq f(\tilde{x})$ to either (a) bound each coordinate separately or (b) derive an inequality of the form $h(x) \leq \alpha$, where $h$ is some coercive function. Examples are given in Chapter 7.
4. CONVEXITY AND OPTIMIZATION

Convexity plays a natural role in optimization. It has tremendous intuitive appeal and provides powerful and elegant approaches to a wide variety of pure and applied problems.

4.1 Convex Sets and Functions

**Definition 4.1.1** (Convex sets). We say that $S \subseteq \mathbb{R}^n$ is convex if $(1 - \lambda)x + \lambda y \in S$ whenever $x, y \in S$ and $\lambda \in [0, 1]$. (See Figure 4.1.)

**Example 4.1.2** (Some convex sets). The following sets are convex:
(a) points, lines and planes;
(b) boxes (as defined in Definition 1.3.3);
(c) the empty set and all of $\mathbb{R}^n$.

**Proposition 4.1.3** (Intersection property). If $S_1, \ldots, S_k \subseteq \mathbb{R}^n$ are all convex, then $\bigcap_{i=1}^k S_i$ is also convex. (See Figure 4.2.)
**Definition 4.1.4** (Epigraph of a function). The *epigraph* of a function \( f : S \to \mathbb{R} \) is the set of all points in \( S \times \mathbb{R} \) that lie on or above the graph of \( f \) (see Figure 4.3.):

\[
epi f = \{(x, \alpha) \in S \times \mathbb{R} \mid f(x) \leq \alpha\}.
\]

![Figure 4.3 Epigraph of a (non-convex) function](image)

**Definition 4.1.5** (Convex functions). A function on \( S \) is said to be *convex* on \( S \) if its epigraph is a convex set. (See Figure 4.4.) In particular, the domain of a convex function is always a convex set.

![Figure 4.4 Epigraphs of convex functions](image)

**Definition 4.1.5** (Hypographs; concave functions).

(a) The *hypograph* of \( f : S \to \mathbb{R} \) is defined as the set

\[
hypo f = \{(x, \alpha) \in S \times \mathbb{R} \mid f(x) \geq \alpha\}.
\]

(b) A function is said to be *concave* if its hypograph is a convex set. Equivalently, \( f \) is concave if \(-f\) is convex.
Theorem 4.1.6 (Secant criterion for convexity). Suppose $S \subseteq \mathbb{R}^n$ is a convex set. The function $f : S \to \mathbb{R}$ is convex if and only if
\[
f((1 - \lambda)x + \lambda y) \leq (1 - \lambda)f(x) + \lambda f(y)
\]
for all $x, y \in S$ and all $\lambda \in [0, 1]$. (See Figure 4.5.)

Proof. $\Rightarrow$: Suppose that $f$ is convex on $S$ and consider $x, y \in S$ with $\lambda \in [0, 1]$. Because $(x, f(x)) \in \text{epi } f$ and $(y, f(y)) \in \text{epi } f$, we have
\[
(1 - \lambda)(x, f(x)) + \lambda(y, f(y)) = ((1 - \lambda)x + \lambda y, (1 - \lambda)f(x) + \lambda f(y)) \in \text{epi } f,
\]
so $f((1 - \lambda)x + \lambda y) \leq (1 - \lambda)f(x) + \lambda f(y)$.

$\Leftarrow$: Now consider $(x, \alpha), (y, \beta) \in \text{epi } f$ and $\lambda \in [0, 1]$. If
\[
f((1 - \lambda)x + \lambda y) \leq (1 - \lambda)f(x) + \lambda f(y) \leq (1 - \lambda)\alpha + \lambda \beta,
\]
then $((1 - \lambda)x + \lambda y, (1 - \lambda)f(x) + \lambda f(y)) = (1 - \lambda)(x, \alpha) + \lambda(y, \beta) \in \text{epi } f$. \hfill $\square$

Proposition 4.1.7 (Convex inequalities). Any set of the form $\{x \in \mathbb{R}^n \mid g(x) \leq \alpha\}$ or $\{x \in \mathbb{R}^n \mid g(x) < \alpha\}$ is convex if the function $g : \mathbb{R}^n \to \mathbb{R}$ is convex. Consequently, any set of the form $\{x \in \mathbb{R}^n \mid g_i(x) \leq \alpha_i$ for $i = 1, \ldots, m\}$ is convex if the functions $g_i : \mathbb{R}^n \to \mathbb{R}$ are all convex.

Proof. Exercise. \hfill $\square$

Example 4.1.8 (Convexity of linear functions, norms and balls).
(a) The function $f(x) = a \cdot x + \beta$ is convex. [Immediate from the secant criterion.]
(b) The function $f(x) = \|x\|$ is convex. [Use the secant criterion and norm properties.]
(c) Open and closed balls are convex. [Follows from part (b) and Proposition 4.1.7.] \hfill $\square$

The secant criterion for convexity leads us naturally to a related definition.
Definition 4.1.9 (Strict convexity and concavity). A function \( f : S \to \mathbb{R} \) is strictly convex on \( S \) if
\[
f((1 - \lambda)x + \lambda y) < (1 - \lambda)f(x) + \lambda f(y)
\]
for all \( x, y \in S \) with \( x \neq y \) and all \( \lambda \in (0, 1) \). (See figure 4.6.) Equivalently, a strictly convex function is a convex function whose graph contains no line segments. A function \( f \) is strictly concave if \(-f\) is strictly convex.

![strictly convex](image)

![convex and concave (not strict)](image)

![convex, but not strictly convex](image)

Figure 4.6 Strict convexity

The next result is the primary incentive for studying convexity in optimization.

Theorem 4.1.10 (Local minimizers are global under convexity). Assume that \( \bar{x} \) is a local minimizer for \( f \) on the convex set \( S \).

(a) If \( f \) is convex on \( S \), then \( \bar{x} \) is a global minimizer for \( f \) on \( S \).

(b) If \( f \) is strictly convex on \( S \), then \( \bar{x} \) is a strict global minimizer for \( f \) on \( S \).

Proof. We prove only the strict case (b); the proof of (a) is similar and left as an exercise. Because \( \bar{x} \) is a local minimizer, there exists \( r > 0 \) with \( f(x) \geq f(\bar{x}) \) for all \( x \in B(\bar{x}, r) \cap S \). Given \( y \in S \) with \( y \neq \bar{x} \), choose \( \lambda \in (0, 1) \) small enough that \( \bar{x} + \lambda(y - \bar{x}) \in B(\bar{x}, r) \). Then \( \bar{x} + \lambda(y - \bar{x}) \) also belongs to \( S \). Thus, by the strict convexity of \( f \), we have
\[
f(\bar{x}) \leq f(\bar{x} + \lambda(y - \bar{x})) < (1 - \lambda)f(\bar{x}) + \lambda f(y).
\]
This implies that \( \lambda f(\bar{x}) < \lambda f(y) \), so that \( f(\bar{x}) < f(y) \).

Here is a derivative-based counterpart to the secant criterion for convexity.

Theorem 4.1.11 (Tangent criterion for convexity). Assume \( S \subseteq \mathbb{R}^n \) is convex and consider a function \( f : S \to \mathbb{R} \) for which \( \nabla f \) is continuous. Then

(a) \( f \) is convex on \( S \) if and only if \( f(x) + \nabla f(x) \cdot (y - x) \leq f(y) \) for all \( x, y \in S \);

(b) \( f \) is strictly convex on \( S \) if and only if \( f(x) + \nabla f(x) \cdot (y - x) < f(y) \) for all \( x, y \in S \) with \( x \neq y \).

\[1\] For example, \( \lambda = \min\{1, r/\|y - \bar{x}\|\}/2 \) works.
In other words, convexity of \( f \) means that all tangent lines (hyperplanes) lie below the graph of \( f \); see Figure 4.7.

![Figure 4.7 Tangent criterion for convexity](image)

**Proof.** \( \Rightarrow \): Assume \( f \) is convex and consider \( x, y \in S \). For \( 0 < \lambda \leq 1 \) we have \( f(x + \lambda(y - x)) \leq (1 - \lambda)f(x) + \lambda f(y) \), so that

\[
\frac{f(x + \lambda(y - x)) - f(x)}{\lambda} \leq f(y) - f(x).
\]

As \( \lambda \to 0^+ \), we get \( \nabla f(x) \cdot (y - x) \leq f(y) - f(x) \) by the Chain Rule. Thus \( f(x) + \nabla f(x) \cdot (y - x) \leq f(y) \). This proves the desired implication in the context of part (a). The strictly convex case for (b) is a bit trickier because \( "<" \) can become \( "\leq" \) in the limit as \( \lambda \to 0^+ \). However, by the reasoning above, we have

\[
\frac{f(x + \lambda(y - x)) - f(x)}{\lambda} \leq f(y) - f(x)
\]

and

\[
f(x) + \nabla f(x) \cdot (x + \lambda(y - x) - x) \leq f(x + \lambda(y - x)),
\]

so \( \nabla f(x) \cdot (y - x) < f(y) - f(x) \).

\( \Leftarrow \): For this implication, we prove only the strict case (b); case (a) is similar and left as an exercise. Assume the gradient inequality \( f(x) + \nabla f(x) \cdot (y - x) < f(y) \) holds for all \( x, y \in S \) with \( x \neq y \). Consider \( u, v \in S \) with \( u \neq v \) along with \( \lambda \in (0, 1) \). We shall prove \( f((1 - \lambda)u + \lambda v) < (1 - \lambda)f(u) + \lambda f(v) \). Let \( x = (1 - \lambda)u + \lambda v \), then substitute \( y = u \) and \( y = v \) separately into the gradient inequality to get

\[
f(x) + \nabla f(x) \cdot (u - x) < f(u),
\]

\[
f(x) + \nabla f(x) \cdot (v - x) < f(v).
\]

Now note that \( x - u = \lambda(v - u) \) and \( x - v = (1 - \lambda)(u - v) \), so that \( x - u = -[\lambda/(1 - \lambda)](x - v) \). Using this in the above two inequalities gives us

\[
(1 - \lambda) \left[ f(x) + \nabla f(x) \cdot \frac{\lambda}{1 - \lambda} (v - x) \right] < \left[ f(u) \right] (1 - \lambda),
\]

\[
\lambda \left[ f(x) + \nabla f(x) \cdot (v - x) \right] < \left[ f(v) \right] \lambda.
\]
Adding these yields \(f(x) < (1 - \lambda)f(u) + \lambda f(v)\), as desired.

\[\square\]

**Corollary 4.1.12** (Global sufficiency). Suppose \(\bar{x}\) is a critical point for the function \(f : S \to \mathbb{R}\), where \(S \subseteq \mathbb{R}^n\) is convex.

(a) If \(f\) is convex, then \(\bar{x}\) is a global minimizer for \(f\).
(b) If \(f\) is strictly convex, then \(\bar{x}\) is a strict global minimizer for \(f\).

**Proof.** Apply Theorem 4.1.11 with \(x = \bar{x}\).

\[\square\]

### 4.2 Practical Tests for Convexity

As shown in the preceding section, convexity has great potential for optimization. But this potential can be realized only if we can readily verify that a given function is convex. The three equivalent formulations (epigraph, secant, tangent) of the preceding section are not always much help in this regard. Fortunately, there are many other ways to demonstrate convexity.

**Theorem 4.2.1** (Hessian criterion for convexity). Suppose \(f : S \to \mathbb{R}\) has continuous second partial derivatives on the open convex set \(S \subseteq \mathbb{R}^n\). Then

(a) \(f\) is convex if and only if \(Hf(x)\) is positive semidefinite for all \(x \in S\);
(b) \(f\) is strictly convex if \(Hf(x)\) is positive definite for all (except finitely many) \(x \in S\).

**Proof.** First we prove part (b) in the case where \(Hf(x)\) is PD for all \(x \in S\); the “if” statement of part (a) is similar and left as an exercise. By Taylor’s formula, we have

\[
f(y) = f(x) + \nabla f(x) \cdot (y - x) + \frac{1}{2} (y - x) \cdot Hf(z)(y - x)
\]

for some \(z \in S\) on the line segment from \(x\) to \(y\). Therefore, if \(Hf(x)\) is positive definite, then \(f(y) > f(x) + \nabla f(x) \cdot (y - x)\). This shows \(f\) is strictly convex.

Now we prove the “only if” statement of part (a): Suppose that \(Hf(x)\) is not positive semidefinite at some point \(x \in C\). Then there exists \(u\) with \(u \cdot Hf(x)u < 0\), so \(\varphi_u(t) = f(x + tu)\) is strictly concave for \(t \in (-\delta, \delta)\). In particular, \(f\) cannot be convex in this case.

To prove part (b) in its full generality, it suffices to prove the one-dimensional case. Let \(I \subseteq \mathbb{R}\) be an interval and assume that \(f''(x) > 0\) for all but finitely many \(x \in I\). By Lemma 2.2.2(b), we see that \(f'\) is strictly increasing on \(I\). Consider \(x, y \in I\) with \(x \neq y\) and let \(z\) be a point strictly between \(x\) and \(y\) with \(f(y) = f(x) + f'(z)(y - x)\). If \(y < z < x\), then \(f'(z) < f'(x)\) and \(y - x < 0\); on the other hand, if \(x < z < y\), then \(f'(z) > f'(x)\) and \(y - x > 0\). In either case, we have \(f'(z)(y - x) > f'(x)(y - x)\), so \(f(y) = f(x) + f'(z)(y - x) > f(x) + f'(x)(y - x)\). This proves the strict convexity.

\[\square\]

**Example 4.2.2** (Using the Hessian to show convexity).

(a) The function \(f(x) = x^4\) on \(\mathbb{R}\) is an example of strict convexity when the Hessian is positive definite at all but one point: \(f''(0) = 0\) and \(f''(x) > 0\) for all \(x \neq 0\).
(b) For the quadratic function \(f(x) = x \cdot Ax + b \cdot c\), we have \(Hf(x) = 2A\). Thus, \(f\) is convex if and only if \(A\) is positive semidefinite.

\[\square\]
**Theorem 4.2.3** (Combining convex functions). Consider $f : S \to \mathbb{R}$, where $S \subseteq \mathbb{R}^n$ is convex. If $f$ has any of the following forms, then $f$ is convex on $S$.

(a) $f(x) = f_1(x) + \cdots + f_k(x)$, where each $f_i$ is convex on $S$;
(b) $f(x) = \alpha g(x) + \beta$, where $\alpha > 0$, $\beta \in \mathbb{R}$ and $g$ is convex on $S$;
(c) $f(x) = g(Ax)$, where $g : \text{range}(A) \to \mathbb{R}$ is convex;
(d) $f(x) = g(h(x))$, where $h : S \to \mathbb{R}$ is convex and $g : \text{range}(h) \to \mathbb{R}$ is convex and nondecreasing;
(e) $f(x) = \max\{f_1(x), \ldots, f_k(x)\}$, with each $f_i$ convex on $S$.

If $f$ has any of the following forms, then $f$ is strictly convex on $S$.

(a') $f(x) = f_1(x) + \cdots + f_k(x)$, where each $f_i$ is convex on $S$ and at least one $f_i$ is strictly convex on $S$;
(b') $f(x) = \alpha g(x) + \beta$, where $\alpha > 0$, $\beta \in \mathbb{R}$ and $g$ is strictly convex on $S$;
(c') $f(x) = g(Ax)$, where $g : \text{range}(A) \to \mathbb{R}$ is strictly convex and $A$ has full column rank;
(d') $f(x) = g(h(x))$, where $h : S \to \mathbb{R}$ is strictly convex and $g : \text{range}(h) \to \mathbb{R}$ is convex and strictly increasing;
(e') $f(x) = \max\{f_1(x), \ldots, f_k(x)\}$, with each $f_i$ strictly convex on $S$.

**Proof.** We will prove (a')–(e') because they’re slightly more complicated than the corresponding non-strict versions. Let $\lambda \in (0, 1)$ and consider $x, y \in S$ with $x \neq y$.

(a') Since $f_i((1 - \lambda)x + \lambda y) \leq (1 - \lambda)f_i(x) + \lambda f_i(y)$ for each $i$ (with strict inequality for at least one $i$), we have

$$f((1 - \lambda)x + \lambda y) = \sum_{i=1}^{k} f_i((1 - \lambda)x + \lambda y) < \sum_{i=1}^{k} \left[ (1 - \lambda)f_i(x) + \lambda f_i(y) \right] = (1 - \lambda)f(x) + \lambda f(y).$$

(b') We have $f((1 - \lambda)x + \lambda y) = \alpha g((1 - \lambda)x + \lambda y) + \beta < \alpha [ (1 - \lambda)g(x) + \lambda g(y) ] + \beta = (1 - \lambda) [ \alpha g(x) + \beta ] + \lambda [ \alpha g(y) + \beta ] = (1 - \lambda)f(x) + \lambda f(y)$.

(c') Note that $Ax \neq Ay$, so

$$f((1 - \lambda)x + \lambda y) = g((1 - \lambda)Ax + \lambda Ay) < (1 - \lambda)g(Ax) + \lambda g(Ay) = (1 - \lambda)f(x) + \lambda f(y).$$

(d') Here we have

$$f((1 - \lambda)x + \lambda y) = g\left( h((1 - \lambda)x + \lambda y) \right) < g((1 - \lambda)h(x) + \lambda h(y)) \leq (1 - \lambda)g(h(x)) + \lambda g(h(y)) = (1 - \lambda)f(x) + \lambda f(y),$$
where the first inequality uses the strict convexity of \( h \) and strict monotonicity of \( g \).

(e') For each \( i \) we have

\[
\begin{align*}
  f_i((1 - \lambda)x + \lambda y) &< (1 - \lambda)f_i(x) + \lambda f_i(y) \\
  &\leq (1 - \lambda) \max_j \{f_j(x)\} + \lambda \max_j \{f_j(x)\} = (1 - \lambda)f(x) + \lambda f(y).
\end{align*}
\]

Therefore, \( f((1 - \lambda)x + \lambda y) = \max_i \{f_i((1 - \lambda)x + \lambda y)\} < (1 - \lambda)f(x) + \lambda f(y). \)

\[\square\]

**Observation 4.2.4** (Epigraphs of max-functions). Part (e) of Theorem 4.2.3 also follows from this equivalence: \( f(x) = \max_i \{f_1(x), \ldots, f_k(x)\} \) if and only if \( \text{epi } f = \bigcap_{i=1}^{k} \text{epi } f_i \).

\[\square\]

**Example 4.2.5** (Some convex functions).

(a) Prove that \( f(x) = \exp(\|x\|^2) \) is strictly convex on \( \mathbb{R}^n \).

Solution: We can write \( f(x) = g(h(x)) \) for the choice \( g(t) = e^t \) and \( h(x) = \|x\|^2 = x \cdot Ix \). Note that \( g \) is strictly increasing and strictly convex, because \( g'(t) = g''(t) = e^t > 0 \) for all \( t \). Likewise, \( h \) is strictly convex because \( Hh(x) = 2I \) is positive definite. Thus we can use the convex composition rule of Theorem 4.2.3(d').

(b) Prove that \( f(x_1, x_2, x_3) = x_1^2 - 2x_1x_2 + 2x_2^2 + x_2x_3 + 3x_3^2 - x_3^3 \) is strictly convex on \( S = \{x \in \mathbb{R}^3 \mid x_3 \leq 0\} \).

Solution: We can write \( f(x) = f_1(x) + f_2(x) \) for

\[
 f_1(x_1, x_2, x_3) = x_1^2 - 2x_1x_2 + 2x_2^2 + x_2x_3 + 3x_3^2, \quad f_2(x_1, x_2, x_3) = -x_3^3.
\]

Here we have

\[
 Hf_1(x) = 2 \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & 1/2 \\ 0 & 1/2 & 3 \end{bmatrix}, \quad Hf_2(x) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -6x_3 \end{bmatrix}.
\]

We see that \( Hf_1(x) \) is positive definite, so \( f_1 \) is strictly convex on \( \mathbb{R}^3 \). At the same time, \( Hf_2(x) \) is positive semidefinite when \( x_3 \leq 0 \), so \( f_2 \) is convex on \( S \). We may therefore apply the sum rule of Theorem 4.2.3(a').

(c) Prove that \( f(x) = e^{x_1}e^{-3x_2}e^{4x_3} \) is convex on \( \mathbb{R}^3 \), but not strictly convex.

Solution: We can write \( f(x) = g(Ax) \) for the choice \( g(t) = e^t \) and \( A = [2, -3, 4] \), where \( g \) is convex. By the linear composition rule of Theorem 4.2.3(c), this shows that \( f \) is convex. To show that strict convexity fails, we need to exhibit a secant for \( f \) which lies entirely in the graph of \( f \): for example, we leave it to the reader to show that \( f \) is constant on the line segment from \( x = (3, 2, 0) \) to \( x = (6, 4, 0) \).

(d) Prove that \( f(x) = \max\{\|x\|, e^{x_1} + 2x_2, 1/x_3\} \) is convex on \( \{x \in \mathbb{R}^3 \mid x_3 > 0\} \), but not strictly convex.

Solution: In this case, we apply the max rule of Theorem 4.2.3(e). We also note that \( \|x\| \) is convex on \( \mathbb{R}^3 \) (see Proposition 4.1.8), that \( e^t \) and \( 2t \) are convex on \( \mathbb{R} \), and that
1/t is convex for t > 0. These last three are then incorporated into f by composition with a linear function (the projection of x onto one of the coordinates x_i) and by the sum rule.

To demonstrate that f is not strictly convex, we need to exhibit a secant of f that also meets the graph of f somewhere between its endpoints. Typically, this means finding points x ≠ y with f(½(x + y)) = ½[f(x) + f(y)], and this is done by exploiting the linearity of f along some line segment. In the present example, one can take x = (0, 0, 1) and y = (0, 2, 1).

(e) The function f(x_1, x_2) = f_1(x_1) + f_2(x_2) is strictly convex on \( \mathbb{R}^2 \) if and only if f_1 and f_2 are strictly convex on \( \mathbb{R} \).

Solution: exercise.

(f) If g is concave, then \( f(x) = 1/g(x) \) is convex on \( S = \{ x \mid g(x) > 0 \} \).

Solution: exercise.

(g) A function \( f : \mathbb{R}^n \to \mathbb{R} \) is both convex and concave if and only if \( f(x) = c \cdot x + \alpha \) for some choice of \( c \in \mathbb{R}^n \) and \( \alpha \in \mathbb{R} \).

Solution: exercise. [Hint: Consider the function \( g(x) = f(x) - f(0), \) then use the fact from linear algebra that \( g(\lambda x + \mu y) = \lambda g(x) + \mu g(y) \) for all \( x, y \in \mathbb{R}^n \) and all \( \lambda, \mu \in \mathbb{R} \) if and only if \( g(x) = c \cdot x \) for some choice of \( c \in \mathbb{R}^n \).]

**Example 4.2.6** (Using convexity as a sufficient condition.).

(a) The function \( f(x) = x(1 - \ln x) \) is strictly concave on \( S = (0, \infty) \) and its only critical point is \( x = 1 \), which must therefore be its strict global maximum.

(b) Consider minimizing and maximizing \( f(x_1, x_2) = x_1^4 + 17x_1^2 + x_2^6 + 8x_1x_2 - 30x_1 \) on \( \mathbb{R}^2 \). We have

\[
\nabla f(x) = \begin{bmatrix} 4x_1^3 + 34x_1 - 8x_2 - 30 \\ 2x_2 + 6x_2^5 - 8x_1 \end{bmatrix},
\]

so there is a critical point at \( \bar{x} = (1, 1) \). It would take quite a bit of work to determine directly whether f has any other critical points. In this case, however, we see that

\[
H f(x) = \begin{bmatrix} 12x_1^2 + 34 & -8 \\ -8 & 2 + 30x_2^4 \end{bmatrix} = \begin{bmatrix} 34 & -8 \\ -8 & 2 \end{bmatrix} + \begin{bmatrix} 12x_1^2 & 0 \\ 0 & 30x_2^4 \end{bmatrix}.
\]

Hence f is strictly convex, so \( \bar{x} = (1, 1) \) is the strict global minimizer and f has no other critical points. (Note that if f did have other critical points, they would have to be global minimizers as well — but we have already shown there is just one global minimizer!)

**Observation 4.2.7.** In general, it isn’t possible to prove convexity of a function by merely demonstrating convexity on different subregions of its domain. As an example, consider the function \( f(x) = -\sqrt{|x|} \), which is shown in Figure 4.8. This function is convex on \( (-\infty, 0] \) and also convex on \( [0, \infty) \), but it is not convex on the full line \( (-\infty, \infty) \).
4.3 Further Properties of Convex Sets and Functions

This section provides some useful facts and concepts about convexity.

**Definition 4.3.1 (Convex combinations).** A *convex combination* (or *weighted average*) of points \( x^{(1)}, \ldots, x^{(m)} \in \mathbb{R}^n \) is a linear combination

\[
\sum_{i=1}^{m} \lambda_i x^{(i)}
\]

for which the scalars \( \lambda_i \) are nonnegative and satisfy \( \sum_{i=1}^{m} \lambda_i = 1 \).

Convex combinations generalize both the notion of a single point \( (m = 1) \) and the interpolation \( (1 - \lambda)x + \lambda y \) of \( m = 2 \) points, as needed in the definition of convex set and the secant characterization of convexity.

**Observation 4.3.2 (Convex hull of a set).** Given an arbitrary set \( C \subseteq \mathbb{R}^n \), the set of all convex combinations of members of \( C \) is itself a convex set, which we call the *convex hull* of \( C \). In particular, given a finite list of points \( a^{(1)}, \ldots, a^{(m)} \in \mathbb{R}^n \), the set

\[
\left\{ x = \sum_{i=1}^{m} \lambda_i a^{(i)} \left| \sum_{i=1}^{m} \lambda_i = 1; \lambda_i \geq 0 \text{ for all } i \right. \right\}
\]

is convex. It can be shown that the convex hull of a set \( C \) is the “smallest convex set containing \( C \).”

**Proposition 4.3.3 (Convex combinations and convex sets).** If \( C \subseteq \mathbb{R}^n \) is convex, then every convex combination of points in \( C \) also belongs to \( C \).

**Proof.** We prove this by induction on the number \( m \) of points in the convex combination. The case \( m = 1 \) is trivial. Suppose that every convex combination of \( m = k \) or fewer points in \( C \) belongs to \( C \). Consider \( x^{(1)}, \ldots, x^{(k+1)} \in C \) and \( \lambda_1, \ldots, \lambda_{k+1} \geq 0 \), with \( \sum_{i=1}^{k+1} \lambda_i = 1 \). If \( \lambda_{k+1} = 1 \), then \( \lambda_i = 0 \) for \( i \leq k \) and we’re done. If \( \lambda_{k+1} < 1 \), then
\[
\sum_{j=1}^{k} \lambda_j = 1 - \lambda_{k+1} > 0. \text{ Hence }
\sum_{i=1}^{k+1} \lambda_i x^{(i)} = \left( \sum_{j=1}^{k} \lambda_j \right) \left[ \sum_{i=1}^{k} \left( \frac{\lambda_i}{\sum_{j=1}^{m} \lambda_j} \right) x^{(i)} \right] + \lambda_{k+1} x^{(k+1)}
\]

\[
= \left( 1 - \lambda_{k+1} \right) \left[ \sum_{i=1}^{k} \left( \frac{\lambda_i}{\sum_{j=1}^{m} \lambda_j} \right) x^{(i)} \right] + \lambda_{k+1} x^{(k+1)} \in C,
\]

because the inductive hypothesis implies that
\[
\sum_{i=1}^{k} \left( \frac{\lambda_i}{\sum_{j=1}^{m} \lambda_j} \right) x^{(i)} \in C.
\]

**Proposition 4.3.4** (Jensen’s inequality). Suppose \( C \subseteq \mathbb{R}^n \) is convex and \( f : C \to \mathbb{R} \) is convex. If \( x^{(1)}, \ldots, x^{(m)} \in C, \lambda_1, \ldots, \lambda_m \geq 0, \text{ and } \sum_{i=1}^{m} \lambda_i = 1 \), then
\[
f \left( \sum_{i=1}^{m} \lambda_i x^{(i)} \right) \leq \sum_{i=1}^{m} \lambda_i f(x^{(i)}).
\]

**Proof.** The proof of Proposition 4.3.2 can be adapted to this setting by using the secant characterization of convexity for functions. Alternatively, Jensen’s inequality can be proved by applying Proposition 4.3.2 to the epigraphical definition of convexity for functions.

**Corollary 4.3.5** (Maximum of a convex function). Suppose \( C \subseteq \mathbb{R}^n \) is convex and \( f : C \to \mathbb{R} \) is convex. If \( x^{(1)}, \ldots, x^{(m)} \in C, \lambda_1, \ldots, \lambda_m \geq 0, \text{ and } \sum_{i=1}^{m} \lambda_i = 1 \), then
\[
f \left( \sum_{i=1}^{m} \lambda_i x^{(i)} \right) \leq \max_i \left\{ f(x^{(i)}) \right\}.
\]

In particular, the maximum of a convex function on the convex hull of a finite set \( D \) must be attained at some member of \( D \).

**Theorem 4.3.6** (Continuity of convex functions). Suppose \( C \subseteq \mathbb{R}^n \) is convex and open. If \( f : C \to \mathbb{R} \) is convex, then \( f \) is continuous on \( C \).

**Proof.** Consider a point \( \bar{x} \in C \) and a closed ball \( B(\bar{x}, r) \subseteq C \). Note that \( \bar{x} \pm r e^{(i)} \in C \) for any coordinate vector \( e^{(i)} \). Let \( \delta = r/(2n) \) and define \( \alpha = 2f(\bar{x}) - \beta \), where
\[
\beta = \max \left\{ f(\bar{x} + r e^{(1)}), \ldots, f(\bar{x} + r e^{(n)}), f(\bar{x} - r e^{(1)}), \ldots, f(\bar{x} - r e^{(n)}) \right\}.
\]

We shall prove that
\[
|f(x) - f(y)| \leq \frac{\beta - \alpha}{\delta} \|x - y\| \quad (\ast)
\]

46
for all \( x, y \in B(\bar{x}, \delta) \). First, we show that \( \alpha \) and \( \beta \) are bounds on \( f \) near \( \bar{x} \).

Claim: \( \alpha \leq f(x) \leq \beta \) for all \( x \in B(\bar{x}, 2\delta) \).

Proof of claim. Let \( \gamma = \sum_{i=1}^{n} |x_i - \bar{x}_i| \). Because \( (\sum_{i=1}^{n} |x_i - \bar{x}_i|^2)^{1/2} < 2\delta \), we must have \( |x_i - \bar{x}_i| < 2\delta < r \) for each \( i \). Hence, \( \gamma < r \) and \( 0 \leq \gamma / r < 1 \). Defining

\[
\lambda_i^+ = \left(1 - \frac{\gamma}{r}\right) \frac{1}{2n} + \frac{1}{r} \max\{0, x_i - \bar{x}_i\}, \quad \lambda_i^- = \left(1 - \frac{\gamma}{r}\right) \frac{1}{2n} + \frac{1}{r} \max\{0, x_i - \bar{x}_i\},
\]

we see that \( \lambda_i^+ > 0 \) and \( \lambda_i^- > 0 \). As an exercise, it can be verified that

\[
\sum_{i=1}^{n} \lambda_i^+ + \sum_{i=1}^{n} \lambda_i^- = \left(1 - \frac{\gamma}{r}\right) + \frac{1}{r} \sum_{i=1}^{n} |x_i - \bar{x}_i| = 1
\]

and

\[
x = \sum_{i=1}^{n} \lambda_i^+ \left(\bar{x} + re^{(i)}\right) + \sum_{i=1}^{n} \lambda_i^- \left(\bar{x} - re^{(i)}\right).
\]

By Corollary 4.3.5, this proves \( f(x) \leq \beta \). Note that this bound holds for any choice of \( x \in B(\bar{x}, 2\delta) \). We now fix \( x \in B(\bar{x}, 2\delta) \) and let \( z = 2\bar{x} - x \). Then \( \|z - \bar{x}\| = \|\bar{x} - x\| \), so \( z \in B(\bar{x}, 2\delta) \) and \( f(z) \leq \beta \). Moreover, we have \( \bar{x} = (x + z)/2 \), so \( f(\bar{x}) \leq \frac{[f(x) + f(z)]}{2} \leq \frac{[f(x) + \beta]}{2} \). Thus \( \alpha \leq f(x) \), which completes the proof of the claim.

We now use the claim to prove \((*)\). Consider \( x, y \in B(\bar{x}, \delta) \) with \( x \neq y \) and define

\[
\epsilon = \|x - y\|, \quad \lambda = \frac{\epsilon}{\delta + \epsilon}, \quad z = y + \frac{\delta}{\epsilon} (y - x).
\]

Note that \( \|z - \bar{x}\| \leq \|y - \bar{x}\| + \|y - x\| \delta / \epsilon < 2\delta \), so \( z \in B(\bar{x}, 2\delta) \) and hence \( f(z) \leq \beta \). Because \( f(x) \geq \alpha \), this shows that \( f(z) - f(x) \leq \beta - \alpha \). Next, we note that \( y = (1 - \lambda)x + \lambda z \), so \( f(y) \leq (1 - \lambda)f(x) + \lambda f(z) \). Subtracting \( f(x) \) from both sides yields \( f(y) - f(x) \leq \lambda[f(z) - f(x)] \leq \lambda(\beta - \alpha) \). Recalling the definitions of \( \lambda, \epsilon \) and \( \delta \), we see that

\[
f(y) - f(x) \leq \frac{\beta - \alpha}{\delta} \|y - x\|.
\]

By interchanging the roles of \( x \) and \( y \) in this argument, we obtain \((*)\). \( \square \)
5. ITERATIVE METHODS FOR OPTIMIZATION

Our first step in solving an optimization problem is to use some form of necessary condition to identify candidates for optimality. For unconstrained problems, this means finding the critical points of the objective function, which in turn requires solving $\nabla f(x) = 0$ for $x$. In general, a nonlinear system of equations can be difficult (or impossible) to solve exactly, so it is useful to have some scheme for approximating solutions. This chapter informally presents some of the standard methods for unconstrained minimization. However, the theory and numerical analysis of these methods go well beyond the scope of this course; further details can be found in the text of Dennis and Schnabel.\(^1\)

5.1 Descent Methods

Suppose we want to minimize $f(x)$ over all $x \in \mathbb{R}^n$ and $\bar{x}$ is our current best “guess.” In the one-dimensional case ($n = 1$), there is a simple rule for attempting to improve upon $\bar{x}$:

- if $f'(\bar{x}) > 0$, we can decrease $f$ by moving to the left of $\bar{x}$;
- if $f'(\bar{x}) < 0$, we can decrease $f$ by moving to the right of $\bar{x}$.

In higher dimensions, we can do the same thing in each coordinate:

- if $(\partial f/\partial x_i)(\bar{x}) > 0$, we can decrease $f$ by decreasing $x_i$;
- if $(\partial f/\partial x_i)(\bar{x}) < 0$, we can decrease $f$ by increasing $x_i$.

Methods based on this idea are known as coordinate-descent methods. Here is a more general notion of descent.

**Definition 5.1.1 (Descent direction).** A vector $v$ is called a descent direction for $f$ at $\bar{x}$ if there exists $\delta > 0$ so that $f(\bar{x} + tv) < f(\bar{x})$ for all $t \in (0, \delta)$.

**Method 5.1.2 (General framework for descent methods).** Given an initial guess $x^{(0)}$, generate a sequence of iterates $\{x^{(k)}\}$ according to the following procedure:

1. [Search direction] Calculate a descent direction $d^{(k)}$ for $f$ at $x^{(k)}$.
2. [Step-size] Choose a value $t^{(k)}$ so that $f(x^{(k)} + t^{(k)}d^{(k)})$ provides an “acceptable decrease” below the value $f(x^{(k)})$.
3. [Update] Set $x^{(k+1)} = x^{(k)} + t^{(k)}d^{(k)}$.
4. [Check for termination] Stop if some pre-specified criterion has been met; otherwise, replace $k$ by $k + 1$ and go to step 1.

There’s a lot of leeway in steps 1, 2, and 4 of the above framework. Different choices allow one to customize the procedure and perhaps prove convergence of the sequence $\{x^{(k)}\}$ to a point $\bar{x}$ with desirable properties, such as the necessary conditions for optimality.

For step 1, the simplest way to identify a descent direction is to consider the derivative of the $v$-restriction $\varphi_v(t) = f(\bar{x} + tv)$ of $f$ at $\bar{x}$ in the direction $v$. Note that $v$ is a descent direction for $f$ at $\bar{x}$ if $\varphi'_v(0) < 0$, and recall that $\varphi'_v(t) = \nabla f(\bar{x} + tv) \cdot v$.

Proposition 5.1.3 (Derivatives and descent). If $\nabla f(\bar{x}) \cdot v < 0$, then $v$ is a descent direction for $f$ at $\bar{x}$. In particular, $v = -A\nabla f(\bar{x})$ is a descent direction if $\nabla f(\bar{x}) \neq 0$ and $A$ is any given positive definite matrix. Moreover, the choice of $v = -\nabla f(\bar{x}) / \|\nabla f(\bar{x})\|$ gives the direction of steepest descent among all unit vectors $v$.

Proof. The first statement follows from the preceding discussion. The second statement is an immediate consequence of the first because $-v \cdot Av < 0$ for all $v \neq 0$. Finally, the statement about steepest descent is a corollary of the Schwarz inequality: for $\|v\| = 1$ we have

$$-\|u\| = -\|u\|\|v\| \leq u \cdot v,$$

with equality if and only if $v = -u / \|u\|$. In other words, the most negative value of $u \cdot v$ over all $v$ with $\|v\| = 1$ occurs when we choose $v = -u / \|u\|$, so steepest descent corresponds to the case where $u = \nabla f(\bar{x})$.

Among all the choices for descent directions, two main types are used in practice. One is the steepest descent direction $v = -\nabla f(x^{(k)})$, which generally does a good job of finding the rough location of a local minimizer. The second type of direction has the form $v = -H^{(k)} \nabla f(x^{(k)})$, where $H^{(k)}$ is an approximation of the Hessian $Hf(x^{(k)})$. Such directions take the curvature of $f$ into account and typically provide fast local convergence if started near a solution. There are also methods that combine these two in some way.

For the step-size selection (step 2 of framework 5.1.2), there are three widely used choices. The simplest is to always choose $t^{(k)}$ to be some fixed number, a choice that requires careful selection of the descent direction. Another possibility is to perform a complete one-dimensional optimization in $t$ to minimize $\varphi_v(t)$. This is called an exact linesearch, and it yields convergence to a critical point for a wide variety of direction choices under mild assumptions on $f$. In general, however, an exact linesearch is impractical. A less demanding possibility is to attempt a full step ($t = 1$) and then adjust $t$ according to some criterion for “acceptable decrease.”

Method 5.1.4 (Inexact linesearch). Suppose $\bar{x}$ and a descent direction $v$ for $f$ at $\bar{x}$ are both given. Fix a value $\alpha \in (0, 1/2)$ and a value $\beta > 1$. Start with $t = 1$.

1. As long as $\varphi_v(t) \leq \varphi_v(0) + \alpha \varphi'_v(0)t$, repeatedly replace $t$ by $t \beta$.
2. As long as $\varphi_v(t) > \varphi_v(0) + \alpha \varphi'_v(0)t$, repeatedly replace $t$ by $t / \beta$.

(Note that $t = \beta^m$ for some integer $m$.)

The effect of this inexact linesearch is to find a value $t \in (0, \infty)$ satisfying $\varphi_v(t) \leq \varphi_v(0) + \alpha \varphi'_v(0)t < \varphi(t \beta)$. As illustrated in Figure 5.1, the downward slope predicts the rate of decrease, whereas the coefficient $\alpha$ hedges slightly on the prediction, thereby allowing room for acceptable decrease that can actually be realized by some step-size.
Eventually, we need to terminate the descent method (step 4 of framework 5.1.2). Criteria for termination fall roughly into three categories. First, the method is declared convergent if the gradient becomes sufficiently small. Second, we stop iterating if it appears that little progress has been made recently. “Progress” might be measured by the difference between consecutive iterates \(x^{(k)}\) and \(x^{(k+1)}\), where a tiny difference indicates the method is stalling and an enormous difference indicates the iterates may be unbounded. The first and second criteria are usually implemented using “relative” measures, such as

\[
\frac{\|\nabla f(x^{(k+1)})\|}{\max\{\|f(x^{(k+1)})\|,1\}} < \epsilon_1, \quad \frac{\|x^{(k+1)} - x^{(k)}\|}{\max\{\|x^{(k)}\|,1\}} \notin [\epsilon_2, 1/\epsilon_3],
\]

where the choice of tolerances \(\epsilon_i\) is a fairly subtle issue. Third, some limitation is placed on the effort to be expended, such as a maximum allowable number of iterations. In subsequent sections, we forego explicit mention of the termination step.

### 5.2 Classic Examples of Descent Methods

The Method of Steepest Descent, due to Cauchy, is the simplest version of a descent method. It takes the following basic form:

**Method 5.2.1** (Steepest descent — Cauchy). Given an initial guess \(x^{(0)}\), generate a sequence of iterates \(\{x^{(k)}\}\) according to the following procedure:

1. Calculate \(d^{(k)} = -\nabla f(x^{(k)})\).
2. Choose a value \(t^{(k)}\) to minimize \(f(x^{(k)} + td^{(k)})\) over all \(t \geq 0\).
3. Set \(x^{(k+1)} = x^{(k)} + t^{(k)}d^{(k)}\). Replace \(k\) by \(k + 1\) and go to step 1.

According to Proposition 5.1.3, this method fits into the general framework of the preceding section. It is guaranteed to converge to a critical point for any coercive, continuously differentiable function \(f\), even if the exact linesearch of step 2 is replaced by the inexact linesearch described in Method 5.1.4.

**Example 5.2.2** (Using steepest descent). Apply two iterations of the Method of Steepest Descent with exact linesearch to the function \(f(x_1, x_2) = 2x_1^2 - x_1x_2 + 3x_2^2\), beginning from the initial guess \(x = (1, 1)\).
Solution. We have

\[ x^{(0)} = (1, 1), \quad \nabla f(x) = \begin{bmatrix} 4x_1 - x_2 \\ -x_1 + 6x_2 \end{bmatrix}. \]

Also, \( f \) is strictly convex, so any critical point of \( \varphi_v \) is a strict global minimizer for \( \varphi_v \).

Calculate \( x^{(1)} \).

direction: \( d^{(0)} = -\nabla f(x^{(0)}) = \begin{bmatrix} -3 \\ -5 \end{bmatrix} \)

step-size: We need to minimize \( \varphi_{d^{(0)}}(t) := f(x^{(0)} + td^{(0)}) \), so we need to solve

\[
0 = \varphi'_{d^{(0)}}(t) = \nabla f(x^{(0)} + td^{(0)}) \cdot d^{(0)} \\
= \begin{bmatrix} 4(1 - 3t) - (1 - 5t) \\ -(1 - 3t) + 6(1 - 5t) \end{bmatrix} \cdot \begin{bmatrix} -3 \\ -5 \end{bmatrix} = -34 + 156t.
\]

So \( t^{(0)} = 34/156 = 17/78 \).

update: \( x^{(1)} = x^{(0)} + t^{(0)}d^{(0)} = \begin{bmatrix} 9/26 \\ -7/78 \end{bmatrix} \)

Calculate \( x^{(2)} \).

direction: \( d^{(1)} = -\nabla f(x^{(1)}) = \begin{bmatrix} -115/78 \\ 69/78 \end{bmatrix} \)

step-size: We need to minimize \( \varphi_{d^{(1)}}(t) := f(x^{(1)} + td^{(1)}) \), which requires solving

\[
0 = \varphi'_{d^{(1)}}(t) = \nabla f(x^{(1)} + td^{(1)}) \cdot d^{(1)} \approx -2.956788 + (15.99869)t.
\]

So \( t^{(1)} \approx 0.1847826 \).

update: \( x^{(2)} = x^{(1)} + t^{(1)}d^{(1)} \approx \begin{bmatrix} 0.0737179 \\ 0.0737179 \end{bmatrix} \)

**Observation 5.2.3** (Zig-zag effect and merely linear convergence rate). The Method of Steepest Descent suffers from a “zig-zag” effect (see Figure 5.2), in the sense that each direction is orthogonal to the next: \( d^{(k+1)} \cdot d^{(k)} = 0 \). This is caused by the exact linesearch:

\[
d^{(k+1)} \cdot d^{(k)} = -\nabla f(x^{(k+1)}) \cdot d^{(k)} = -\nabla f(x^{(k)} + t^{(k)}d^{(k)}) \cdot d^{(k)} = -\varphi'_{d^{(k)}}(t^{(k)}) = 0.
\]

This behavior forces the Method of Steepest Descent to converge no faster than a *linear rate*: for all large \( k \) and some fixed \( \beta \in (0, 1) \), we have

\[
\|x^{(k+1)} - \bar{x}\| \approx \beta\|x^{(k)} - \bar{x}\|.
\]

Near the optimum, this relation between consecutive iterates means that essentially the same number of new digits of accuracy are added in each iteration. Unfortunately, the “number of new digits” per iteration may be a small fraction, so that dozens or hundreds of
iterations are actually required to obtain each full digit of accuracy. The upshot is that the intended gains in accuracy are more than offset by losses due to round-off error. And even without round-off, this is considered to be a slow form of convergence. For this reason, the Method of Steepest Descent is primarily reserved for the early stages of a descent method, where it makes large gains rapidly.

Figure 5.2 Zig-zag effect of steepest descent

The Method of Steepest Descent is simply too greedy: the linesearch passes up shorter step-sizes that better reflect the curvature of the objective function near the current iterate \(x^{(k)}\). The main difficulty is that there is no “natural” step-size for the method. An approach that makes far better use of curvature information is known as Newton’s Method.

**Method 5.2.4 (Newton’s Method for minimization).** Given an initial guess \(x^{(0)}\), generate a sequence of iterates \(\{x^{(k)}\}\) according to the following procedure:

1. Calculate \(d^{(k)} = -[Hf(x^{(k)})]^{-1} \nabla f(x^{(k)})\).
2. Choose \(t^{(k)} = 1\).
3. Set \(x^{(k+1)} = x^{(k)} + t^{(k)}d^{(k)}\). Replace \(k\) by \(k + 1\) and go to step 1.

Note: the calculation of \(d^{(k)}\) in step 1 should be carried out by solving the linear system \([Hf(x^{(k)})]d = -\nabla f(x^{(k)})\) for \(d\), instead of explicitly inverting the Hessian matrix.

**Example 5.2.8 (Using Newton’s Method for minimization).** Carry out two iterations of Newton’s Method for optimizing the function \(f(x_1, x_2) = x_1^2 - 3x_1x_2 + x_2^4 - x_2\), starting at the initial guess \(x = (0, 1)\).

**Solution.** Here we have

\[
\nabla f(x) = \begin{bmatrix} 2x_1 - 3x_2 \\ -3x_1 + 4x_2^3 - 1 \end{bmatrix}, \quad Hf(x) = \begin{bmatrix} 2 & -3 \\ -3 & 12x_2^2 \end{bmatrix}, \quad x^{(0)} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.
\]

Calculate \(x^{(1)}\).

- direction: solve \([Hf(x^{(0)})]d = -\nabla f(x^{(0)})\) for \(d\); i.e.,
  \[
  \begin{bmatrix} 2 & -3 \\ -3 & 12 \end{bmatrix} d = \begin{bmatrix} 3 \\ -3 \end{bmatrix} \Rightarrow d^{(0)} = \begin{bmatrix} 9/5 \\ 1/5 \end{bmatrix}
  \]

- update: \(x^{(1)} = x^{(0)} + d^{(0)} = \begin{bmatrix} 9/5 \\ 6/5 \end{bmatrix}\)
Calculate $x^{(2)}$.

Direction: solve $[Hf(x^{(1)})]d = -\nabla f(x^{(1)})$ for $d$; i.e.,

$$
\begin{bmatrix}
2 & -3 \\
-3 & 432/25
\end{bmatrix} d = \begin{bmatrix}
0 \\
-64/125
\end{bmatrix} \implies d^{(1)} = \begin{bmatrix}
-64/1065 \\
-128/3195
\end{bmatrix}
$$

Update: $x^{(2)} = x^{(1)} + d^{(1)} = \begin{bmatrix}
1853/1065 \\
747/644
\end{bmatrix}$. 

\[ \square \]

**Theorem 5.2.5** (Convergence of Newton’s Method for minimization). Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ has continuous third partial derivatives, that $Hf(\bar{x})$ is positive definite, and that $\nabla f(\bar{x}) = 0$. If Newton’s method is started at any guess $x^{(0)}$ sufficiently close to $\bar{x}$, then its iterates converge to $\bar{x}$. Moreover, the convergence occurs at a quadratic rate: for all large $k$ and some fixed $\beta \in (0, 1)$, we have

$$
\|x^{(k+1)} - \bar{x}\| \approx \beta\|x^{(k)} - \bar{x}\|^2.
$$

**Sketch of proof** of quadratic convergence. We look only at the one-dimensional case because the notation is simpler. Applying Taylor’s theorem to both $f'$ and $f''$ about $\bar{x}$, we obtain

$$
|x^{(k+1)} - \bar{x}| = \left| \left[ x^{(k)} - \frac{f'(x^{(k)})}{f''(x^{(k)})} \right] - \bar{x} \right| = \left| \frac{|f'(x^{(k)}) + f''(x^{(k)})(\bar{x} - x^{(k)})|}{f''(x^{(k)})} \right|
$$

$$
= \left| \left[ f'(\bar{x}) + f''(\bar{x})(x^{(k)} - \bar{x}) + \frac{1}{2} f'''(z^{(k)})(x^{(k)} - \bar{x})^2 \right] \right|
$$

$$
= \left| \frac{1}{2} f'''(z^{(k)}) - f'''(y^{(k)}) \right| \|x^{(k)} - \bar{x}\|^2 \approx \frac{|f'''(\bar{x})|}{2|f''(\bar{x})|} \|x^{(k)} - \bar{x}\|^2.
$$

\[ \square \]

**Observation 5.2.6** (Self-correcting behavior of Newton’s Method). The quadratic rate of convergence means, roughly, that each iteration of Newton’s Method doubles the number of digits of accuracy. This speed of convergence frequently overcomes the round-off error, so that the method is essentially “self-correcting.”

**Observation 5.2.7** (Newton’s Method as a quadratic approximation to the objective). By the Taylor expansion of $f$ about $x^{(k)}$, we have

$$
f(x) \approx f(x^{(k)}) + \nabla f(x^{(k)}) \cdot (x - x^{(k)}) + \frac{1}{2} (x - x^{(k)}) \cdot Hf(x^{(k)})(x - x^{(k)}).
$$

This quadratic approximation to the objective $f(x)$ near the current iterate $x^{(k)}$ has its gradient and Hessian at $x$ given, respectively, by

$$
\nabla f(x^{(k)}) + Hf(x^{(k)})(x - x^{(k)}) \quad \text{and} \quad Hf(x^{(k)}).
$$
Consequently, if $Hf(x^{(k)})$ is positive definite, then we can minimize the quadratic approximation by finding its critical point, which is the solution $x$ of

$$\nabla f(x^{(k)}) + Hf(x^{(k)})(x - x^{(k)}) = 0.$$ 

Solving for $x$ in this equation gives us $x = x^{(k)} - [Hf(x^{(k)})]^{-1}\nabla f(x^{(k)})$, which is the next iterate $x^{(k+1)}$ in Newton’s Method. Consequently, Newton’s Method can be viewed as finding the minimizer of a quadratic approximation to $f$ near $x^{(k)}$. (See Figure 5.3.)

\[ \text{Figure 5.3 Newton's method as a quadratic approximation of the objective} \]

The unfortunate corollary of this last observation is that Newton’s method is not guaranteed to provide improvement of the objective function: the quadratic approximation may grossly overshoot the local minimum. In fact, the method may move in the wrong direction altogether if the Hessian is not positive definite. For these reasons, Newton’s Method is primarily of value only when we can give it a fairly good starting point, which it will then clean up quickly.

We now summarize the trade-offs between Steepest Descent and Newton’s Method. The Method of Steepest Descent provides the following advantages:

- guaranteed descent (regardless of curvature);
- “clusters” about critical points for coercive functions;
- no need to compute Hessians or to repeatedly solve linear systems.

On the downside, Steepest Descent has very poor local convergence. Newton’s Method excels at local convergence, but is costly and provides no guarantee of global convergence. Because the trade-offs are so evenly split, practical optimization methods provide some sort of combination of the two. A very simple hybrid is to apply Steepest Descent at the beginning, and then switch to Newton’s Method when some criterion indicates that local convergence is required. Alternatively, one can apply Newton’s Method as the primary algorithm and use Steepest Descent only when the Newton step fails to give an acceptable decrease in the objective value.

In practice, two other hybridizations are widely used. The first of these constitutes the general class of quasi-Newton methods. Instead of using the actual Hessian $Hf(x^{(k)})$, quasi-Newton methods use an approximate Hessian $H^{(k)}$ that is required to satisfy the secant equation

$$H^{(k+1)}(x^{(k+1)} - x^{(k)}) = \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}).$$
Many different matrices $H^{(k+1)}$ will work here. The most popular choice is the BFGS update (named for its inventors — Broyden, Fletcher, Goldfarb, and Shanno):

$$H^{(k+1)} = H^{(k)} + \frac{yy^T}{y \cdot s} - \frac{pp^T}{s \cdot p},$$

where

$$y := \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}), \quad s := x^{(k+1)} - x^{(k)}, \quad p := H^{(k)}s.$$

A practical BFGS method calculates the search direction just as in Newton’s method but with the approximate Hessian $H^{(k)}$; the linear system can be solved efficiently for $d^{(k)}$ by methods exploiting the specific way the coefficient matrix is being updated. An inexact linesearch is generally employed, along with various devices for determining the quality of descent. When descent is considered insufficient, the approximate Hessian may be replaced by the identity matrix, thereby giving one iteration of pure Steepest Descent.

Another idea that works very well in practice is the notion of a trust region. The basic idea is that a local quadratic approximation of the objective $f$ can only be “trusted” in a region close to the current iterate $x^{(k)}$. The minimization of the local approximation is therefore constrained to a ball about $x^{(k)}$, so that it has the form

$$\min_{x} \left\{ \frac{1}{2} x \cdot A^{(k)} x + b^{(k)} \cdot x \left| \|x - x^{(k)}\| \leq \Delta^{(k)} \right. \right\}.$$

This minimization can be carried out very efficiently (it is equivalent to a one-dimensional root-finding problem) even if the matrix $A^{(k)}$ is not positive semidefinite! For this reason, trust region methods can actually exploit negative curvature when it’s present. As iterates are generated, the trust region radius $\Delta^{(k)}$ is adjusted according to criteria measuring the recent “trustworthiness” of the local quadratic model in predicting improvement.

### 5.3 Solving Systems of Nonlinear Equations

Root-finding is closely related to optimization. For example, optimizing a function frequently involves solving a system of equations, such as $\nabla f(x) = 0$. Conversely, the problem of solving a general system of the form $g(x) = 0$ can be recast as an optimization: minimize $f(x) := \|g(x)\|^2$ over all $x$. This optimization problem is known as nonlinear least squares. The special case where $g(x) = Ax - b$ is called linear least squares and is the topic of the next chapter.

To solve a nonlinear system of equations, we can apply any of the methods of the preceding section to the nonlinear least squares formulation. However, the Newton and quasi-Newton methods can also be restated directly in terms of the function $g : \mathbb{R}^n \to \mathbb{R}^n$. The main idea is to consider the first-order Taylor expansion of $g$ about an iterate $x^{(k)}$:

$$g(x) \approx g(x^{(k)}) + g'(x^{(k)})(x - x^{(k)}).$$

If $g'(x^{(k)})$ is invertible, then we can let the next iterate $x^{(k+1)}$ be a root of the linearization on the right-hand side; see Figure 5.4. Here is the formal statement of this method.
**Method 5.3.1** (Newton’s Method for root-finding). Given an initial guess \(x^{(0)}\), generate a sequence of iterates \(\{x^{(k)}\}\) according to the following procedure:

1. Calculate \(d^{(k)} = -[g'(x^{(k)})]^{-1}g(x^{(k)})\).
2. Set \(x^{(k+1)} = x^{(k)} + d^{(k)}\). Replace \(k\) by \(k + 1\) and go to step 1.

(Note: the direction \(d^{(k)}\) should be calculated by solving \([g'(x^{(k)})]d = -g(x^{(k)})\) for \(d\).)

![Figure 5.4 Newton’s method as a linear approximation for root-finding](image)

In cases where \(g(x) = \nabla f(x)\), we see that \(g' = HF\) and we recover Newton’s Method for optimization. The convergence analysis is essentially the same in the root-finding case.

**Theorem 5.3.2** (Convergence of Newton’s Method for systems of equations). Suppose that \(g : \mathbb{R}^n \rightarrow \mathbb{R}^n\) has continuous second partial derivatives, that \(g'(\bar{x})\) is invertible, and that \(g(\bar{x}) = 0\). If Newton’s method is started at any guess \(x^{(0)}\) sufficiently close to \(\bar{x}\), then its iterates converge to \(\bar{x}\). Moreover, the convergence occurs at a quadratic rate: for all large \(k\) and some fixed \(\beta \in (0, 1)\), we have

\[\|x^{(k+1)} - \bar{x}\| \approx \beta \|x^{(k)} - \bar{x}\|^2.\]

**Example 5.3.3** (Using Newton’s Method on a system of equations). Carry out two iterations of Newton’s Method for solving the system

\[x_1^2 + x_2^2 = 2, \quad x_1 + x_2 = 2\]

starting at the initial guess \(x = (1, 0)\).

**Solution.** Here we have

\[g(x) = \begin{bmatrix} x_1^2 + x_2^2 - 2 \\ x_1 + x_2 - 2 \end{bmatrix}, \quad g'(x) = \begin{bmatrix} 2x_1 & 2x_2 \\ 1 & 1 \end{bmatrix}, \quad x^{(0)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.

Calculate \(x^{(1)}\).

direction: solve \([g'(x^{(0)})]d = -g(x^{(0)})\) for \(d\); i.e.,

\[
\begin{bmatrix} 2 & 0 \\ 1 & 1 \end{bmatrix} d = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \Rightarrow \quad d^{(0)} = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}
\]
update: \( x^{(1)} = x^{(0)} + d^{(0)} = \begin{bmatrix} 3/2 \\ 1/2 \end{bmatrix} \)

Calculate \( x^{(2)} \).

direction: solve \([g'(x^{(1)})]d = -g(x^{(1)})\) for \( d \); i.e.,

\[
\begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix} d = \begin{bmatrix} -1/2 \\ 0 \end{bmatrix} \implies d^{(1)} = \begin{bmatrix} -1/4 \\ 1/4 \end{bmatrix}
\]

update: \( x^{(2)} = x^{(1)} + d^{(1)} = \begin{bmatrix} 5/4 \\ 3/4 \end{bmatrix} \)

For quasi-Newton methods, the derivative \( g'(x^{(k)}) \) is replaced in Method 5.3.1 by an approximation \( G^{(k)} \). The analogue of the secant equation in this setting is

\[
G^{(k+1)} \left( x^{(k+1)} - x^{(k)} \right) = g(x^{(k+1)}) - g(x^{(k)}).
\]

The preferred way to build such a sequence is Broyden’s update:

\[
G^{(k+1)} = G^{(k)} + \frac{(y - G^{(k)}s) s^T}{s \cdot s},
\]

where

\[
y := g(x^{(k+1)}) - g(x^{(k)}), \quad s := x^{(k+1)} - x^{(k)}.
\]
6. LINEAR LEAST SQUARES AND RELATED PROBLEMS

Linear least squares constitutes one of the most important classes of optimization problems in modern society, primarily because of its central role in statistical data analysis. Within optimization itself, linear least squares provides basic tools needed in constrained optimization and serves as a prototype for more complicated problems.

6.1 A Motivating Application — Curve-Fitting

Suppose we have a finite data set of data pairs (see Figure 6.1)

\[(t_1, s_1), \ldots, (t_m, s_m)\]

meant to represent the values of some function \(s = f(t)\), where \(f\) is assumed to belong to a specified class (e.g., all polynomials of degree less than \(p\), or all sums of trig functions).

![Figure 6.1 Scatterplot of data pairs and a curve of the form \(s = at + bt^2 + c \sin t\)]

**Goal 6.1.1 (Best fit).** Choose \(f\) from the specified class of functions so as to obtain the “best” fit to the given data points. The notion of “best” is viewed subjectively as meaning that

\[|s_i - f(t_i)|\]

is made as small as possible for all \(i\).

A typical context is that the theory in some area of application suggests that the data in Figure 6.1 should be representable in the form

\[s = at + bt^2 + c \sin t,\]

for some choice of the parameters \((a, b, c)\). Our job is to make each of the quantities

\[|s_i - [at_i + bt_i^2 + c \sin t_i]|, \text{ for } i = 1, \ldots, m\]

as small as possible. The difficulty is that a choice of \((a, b, c)\) which makes \(|s_1 - f(t_1)|\) very small may make \(|s_3 - f(t_3)|\) very large. Consequently, we need some sort of aggregate measure over all the deviations \(|s_i - f(t_i)|\). Here are some popular choices:
(1) *l*₁-fit \[ \text{minimize } f_1(a, b, c) = \sum_{i=1}^{m} |s_i - [at_i + bt_i^2 + c\sin t_i]| \]

(2) *least-squares fit* \[ \text{minimize } f_2(a, b, c) = \sum_{i=1}^{m} |s_i - [at_i + bt_i^2 + c\sin t_i]|^2 \]

(∞) *minimax fit* \[ \text{minimize } f_\infty(a, b, c) = \max_{i} |s_i - [at_i + bt_i^2 + c\sin t_i]| \]

Each of these objective functions has desirable qualities. In particular, all are convex functions of \((a, b, c)\):

- \(|x|\) and \(|x|^2\) are convex on \(\mathbb{R}\);
- \((a, b, c) \mapsto s_i - [at_i + bt_i^2 + c\sin t_i]\) is linear (affine) on \(\mathbb{R}^3\).
- \(f_1\) and \(f_2\) are sums of convex functions, whereas \(f_\infty\) is a maximum of convex functions.

The functions \(f_1\) and \(f_\infty\) are special cases of “linear programming,” which is discussed later in the course. The present chapter focuses on least squares and its special properties: \(f_2\) is twice differentiable in \((a, b, c)\) and the optimality conditions for \(f_2\) can be treated directly by linear algebra techniques.

### 6.2 Linear Least Squares

As in the preceding section, we assume we’re given data points \((t_1, s_1), \ldots, (t_m, s_m)\) and a prescribed list of functions \(q_1, \ldots, q_n\). Our goal is to find the coefficients \(x_1, \ldots, x_n\) that minimize the objective

\[ f_2(x_1, \ldots, x_n) = \sum_{i=1}^{m} |s_i - [x_1 q_1(t_i) + \cdots + x_n q_n(t_i)]|^2. \]

Observe that the list of values

\[
\begin{aligned}
x_1q_1(t_1) + \cdots + x_nq_n(t_1), \\
\vdots \\
x_1q_1(t_m) + \cdots + x_nq_n(t_m),
\end{aligned}
\]

can be written more compactly as \(Ax\), where we define \(A\) to be the matrix

\[
A = \begin{bmatrix}
q_1(t_1) & \cdots & q_n(t_1) \\
\vdots & \ddots & \vdots \\
q_1(t_m) & \cdots & q_n(t_m)
\end{bmatrix}.
\]

Taking

\[
b = \begin{bmatrix}
s_1 \\
\vdots \\
s_m
\end{bmatrix}
\]

59
allows us to rewrite the objective as
\[ f_2(x) = \|b - Ax\|^2. \]
Expanding this in terms of dot products and differentiating yields
\[ f_2(x) = b \cdot b - 2(A^T b) \cdot x + x \cdot A^T A x, \]
\[ \nabla f_2(x) = -2A^T b + 2A^T A x, \quad H f_2(x) = 2A^T A. \]
In particular, the Hessian is positive semidefinite as expected, given the convexity of \( f_2 \).
Moreover, it is positive definite if and only if \( A \) has full column rank.

**Theorem 6.2.1 (Normal equations for linear least squares).** The global minimizers for \( \|b - Ax\|^2 \) are precisely the solutions of the normal equations
\[ A^T A x = A^T b. \]
If \( A \) has full column rank, the unique minimizer is given by \( \bar{x} = (A^T A)^{-1} A^T b \).

**Proof.** The necessity and sufficiency of the normal equations for optimality stem from the observation that, for a differentiable convex function, optimality is equivalent to being a critical point. The statement concerning uniqueness then follows from the normal equations and the full-rank assumption.

**Observations 6.2.2 (Interpreting and solving least squares problems).**
(a) One interpretation of least squares is as follows: we can’t solve an overdetermined system \( Ax = b \) exactly, so the next best thing is to minimize the difference \( \|Ax - b\| \).
If \( Ax = b \) actually is solvable, then this achieves the minimum value \( \|Ax - b\| = 0 \).
(b) When solving a least-squares problem by hand, we simply form the normal equations and apply row-reduction. For larger systems solved by a computer, this approach is prone to round-off errors, so other methods are used.

**Definition 6.2.3 (Pseudo-inverse of a matrix).** When \( A \) has full column rank, the matrix \( A^\dagger := (A^T A)^{-1} A^T \) in Theorem 6.2.1 is called the pseudo-inverse of \( A \). Like the inverse matrix, it is a useful theoretical device that is generally avoided in actual computation.

**Example 6.2.4 (Solving a least squares problem).** Find a least squares fit of the data points \((1, 2), (2, 2.5), (3, 2.9), \) and \((4, 3.6)\) to a model of the form \( s = f(t) \), where \( f \) is chosen from the parameterized family \( f(t) = \alpha t + \beta + \gamma e^t \).

**Solution.** We set up the data in matrix form as
\[ A = \begin{bmatrix} 1 & 1 & e^1 \\ 2 & 1 & e^2 \\ 3 & 1 & e^3 \\ 4 & 1 & e^4 \end{bmatrix}, \quad b = \begin{bmatrix} 2 \\ 2.5 \\ 2.9 \\ 3.6 \end{bmatrix}, \quad x = \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix}. \]
We use these to form the matrix $A^T A$ and vector $A^T b$ and then solve the normal equations $(A^T A)x = (A^T b)$ for $x$ to get:

$$
\bar{x} \approx \begin{bmatrix} 0.3918 \\ 1.6091 \\ 0.0076 \end{bmatrix}.
$$

The best fit for the model is therefore $f(t) = 0.3918t + 1.6091 + 0.0076e^t$.  

An important property of the linear least squares problem is that a solution always exists. Before proving this, we recall some concepts from linear algebra.

**Definition 6.2.5 (Linear subspace).** A set $L \subseteq \mathbb{R}^m$ is a linear subspace if it satisfies these conditions:

- $x + y \in L$ whenever $x, y \in L$;
- $\lambda x \in L$ whenever $x \in L$ and $\lambda \in \mathbb{R}$.

Equivalently, a linear subspace contains all linear combinations of its elements:

$$
\sum_{i=1}^{k} \lambda^{(i)} x^{(i)} \in L \text{ whenever } \lambda^{(i)} \in \mathbb{R}, x^{(i)} \in L.
$$

**Examples 6.2.6 (Some subspaces).** The following are subspaces (verification is an exercise):

(a) the range (also called column space or image) of a matrix, given by

$$
\text{range}(A) = \{ y \mid y = Ax \text{ for some } x \};
$$

(b) the null space (also called kernel) of a matrix, given by

$$
\text{null}(A) = \{ x \mid Ax = 0 \};
$$

(c) the orthogonal complement of a set $S$, given by

$$
S^\perp = \{ y \mid x \cdot y = 0 \text{ for all } x \in S \}.
$$

**Definition 6.2.7 (Basis of a linear subspace).** A basis for a linear subspace $L$ is a set $\{ x^{(1)}, \ldots, x^{(k)} \} \subseteq L$ that:

- spans $L$, so that each $y \in L$ is expressible as a linear combination of the $x^{(i)}$;
- is linearly independent, so that the only way to write the vector 0 as a linear combination $\sum_{i=1}^{k} \lambda^{(i)} x^{(i)}$ is to have all $\lambda^{(i)} = 0$.

Equivalently, $L = \text{range}(A)$ for some matrix $A = [x^{(1)}, \ldots, x^{(k)}]$ of full column rank.
Theorem 6.2.8 (Existence of a basis). Each linear subspace $L \subseteq \mathbb{R}^m$ has a basis. Moreover, if $S \subseteq L$ spans $L$ then a basis for $L$ can be selected from the elements of $S$.

Theorem 6.2.9 (Existence of solutions for linear least squares problems). The linear least squares problem $\min_x \|b - Ax\|^2$ always admits a global minimizer.

Proof. Without loss of generality, we may assume that $A$ is partitioned as $A = [B, N]$, where $B$ has full column rank and rank $B = \text{rank } A$. This is possible because we can identify a basis for the column space of $A$ from among the columns of $A$ (by Theorem 6.2.8), and then permute the columns so the basis columns come first. Corresponding to this partition, we can write

$$x = \begin{bmatrix} u \\ v \end{bmatrix},$$

so that

$$Ax = [B, N] \begin{bmatrix} u \\ v \end{bmatrix} = Bu + Nv.$$

The assumption that rank $B = \text{rank } A$ implies that every column in $N$ is a linear combination of columns in $B$, so $N = BW$ for some matrix $W$. In other words, $Ax = B(u + Wv)$, which shows we can effectively ignore $v$ altogether. For instance, we can assume $v = 0$ and just carry out the optimization with respect to the $u$-portion of $x$:

$$\min_x \|b - Ax\|^2 = \min_u \|b - Bu\|^2.$$

For this new problem over $u$, the normal equations are $B^T Bu = B^T b$. Since $B$ has full column rank, we can solve these to get $\bar{u} = (B^T B)^{-1} B^T b$. This corresponds to a solution

$$\bar{x} = \begin{bmatrix} \bar{u} \\ 0 \end{bmatrix}$$

of the original problem. 

6.3 A Very Brief Look at Nonlinear Least Squares

Consider fitting a model of the form $s = \sin(\alpha t) + \cos(\beta t)$. This is nonlinear in the parameters $(\alpha, \beta)$, so the corresponding least squares problem will not be convex and we won’t be able to solve it directly by linear algebra methods.

In general, the nonlinear least squares problem attempts to solve $g(x) = 0$ in the “least-squares sense:”

$$\min_x f(x) = \|g(x)\|^2.$$

Here we have

$$\nabla f(x) = 2[g'(x)]^T g(x), \quad Hf(x) = 2[g'(x)]^T g'(x) + \sum_{i=1}^m [g_i(x) Hg_i(x)].$$

62
From this we see that the objective function is convex near a minimizer $\bar{x}$ if $g'(\bar{x})$ has full column rank and the magnitudes $|g_i(\bar{x})|$ are small. In such cases, we expect the descent methods of Chapter 5 to perform pretty well. Here’s another approach that is popular, although it’s often inferior to the methods of Chapter 5.

**Method 6.3.1 (Gauss-Newton).** Given an initial guess $x^{(0)}$, generate a sequence of iterates $\{x^{(k)}\}$ according to the following procedure:

1. Calculate a linear least squares solution $d^{(k)}$ to “$[g'(x^{(k)})]d = -g(x^{(k)})$.”
2. Set $x^{(k+1)} = x^{(k)} + d^{(k)}$. Replace $k$ by $k + 1$ and go to step 1.

6.4 Nearest-Point Projections and Complementary Subspaces

Many applications of optimization are a variation on the following nearest-point problem:

Given a set $S \subseteq \mathbb{R}^m$ and a point $z \notin S$, find a point $\bar{x} \in S$ that is nearest to $z$. The special case where $S$ is a linear subspace is fundamental and can be treated by the methods of linear least squares.

**Theorem 6.4.1 (Nearest points in subspaces).** Suppose that $L \subseteq \mathbb{R}^m$ is a linear subspace and let $z \in \mathbb{R}^m$ be given. Then there exists a unique point $\bar{x} \in L$ that is nearest to $z$. Moreover, the following are equivalent:

(a) $\bar{x}$ minimizes $\|z - x\|$ over all $x \in L$;
(b) $z - \bar{x} \in L^\perp$ and $\bar{x} \in L$;

If $L = \text{range}(A)$, then (a) and (b) are also equivalent to

(c) $\bar{x} = A\bar{y}$ for some $\bar{y}$ satisfying $A^TA\bar{y} = A^Tz$.

If, in addition, $A$ has full column rank, then (a)–(c) are also equivalent to

(d) $\bar{x} = AA^Tz$.

**Proof.** By Theorem 6.2.8, we may assume $L = \text{range}(A)$ for some $m \times n$ matrix $A$ of full column rank. Consider $\bar{x} \in L$, so that $\bar{x} = A\bar{y}$ for some $\bar{y} \in \mathbb{R}^n$. Then

(a) holds $\iff \|z - \bar{x}\| \leq \|z - x\|$, $\forall x \in L$
$\iff \|z - A\bar{y}\| \leq \|z - Ay\|$, $\forall y \in \mathbb{R}^n$
$\iff A^TA\bar{y} = A^Tz$ [by the normal equations; this gives (c)]
$\iff A^TA\bar{y} - A^Tz = 0$
$\iff (A^TA\bar{y} - A^Tz) \cdot y = 0$, $\forall y \in \mathbb{R}^n$
$\iff (A\bar{y} - z) \cdot A\bar{y} = 0$, $\forall y \in \mathbb{R}^n$
$\iff (A\bar{y} - z) \cdot x = 0$, $\forall x \in L = \text{range}(A)$
$\iff A\bar{y} - z \in L^\perp$
$\iff \bar{x} - z \in L^\perp$
$\iff (b)$ holds.
The equivalence of (c) to (d) comes from Theorem 6.2.1, which also proves the uniqueness, regardless of the choice of basis. Finally, the existence follows from Theorem 6.2.9.

Observations 6.4.2 (Interpreting the nearest-point optimality conditions). Part (b) of Theorem 6.4.1 expresses \( z \) uniquely as \( z = \bar{x} + \bar{w} \) for some \( \bar{x} \in L \) and some \( \bar{w} \in L^\perp \). Part (c) allows us to calculate \( \bar{x} = Ay \) by solving “\( Ay = z \)” in the least squares sense.

Definition 6.4.3 (Orthogonal projections). The unique nearest point in a linear subspace \( L \) to \( z \) is called the projection of \( z \) onto \( L \). (It is sometimes called the orthogonal projection or metric projection.) The projection is written as \( P_L(z) \).

Observation 6.4.4 (Projection matrices). Part (d) of Theorem 6.4.1 shows that the projection mapping can be expressed as \( P_L(z) = AA^\dagger z \), for any choice of \( A \) whose columns form a basis for \( L \). We can therefore consider \( P_L \) a matrix, representable as \( P_L = AA^\dagger \).

The next result is useful in building optimality conditions for more general problems.

Theorem 6.4.5 (Orthogonal complementarity of subspaces). For any linear subspace \( L \subseteq \mathbb{R}^m \), we have \( (L^\perp)^\perp = L \).

Proof. To prove \( L \subseteq (L^\perp)^\perp \), consider \( x \in L \). Then \( x \cdot y = 0 \) for all \( y \in L^\perp \), so \( x \in (L^\perp)^\perp \).

To prove the converse, \( L \supseteq (L^\perp)^\perp \), we need two main facts, namely,

\[
\begin{align*}
&\bullet \ x \cdot (x - P_Lx) = 0 \text{ whenever } x \in (L^\perp)^\perp; \\
&\bullet \ (P_Lx) \cdot (x - P_Lx) = 0. 
\end{align*}
\]

[because \( x - P_Lx \in L^\perp \) \[because \( P_Lx \in L \) and \( x - P_Lx \in L^\perp \)]

Now, if \( x \in (L^\perp)^\perp \), we must have

\[
\|x - P_Lx\|^2 = (x - P_Lx) \cdot (x - P_Lx) = x \cdot (x - P_Lx) - (P_Lx) \cdot (x - P_Lx) = 0 - 0,
\]

so \( \|x - P_Lx\|^2 = 0 \). Hence \( x = P_Lx \in L \).

Corollary 6.4.6 (Transpose complementarity of range and null space). For any matrix \( A \) we have

\[
\text{range}(A)^\perp = \text{null}(A^T), \quad \text{null}(A)^\perp = \text{range}(A^T).
\]

Proof. For the first statement, we have

\[
x \in \text{range}(A)^\perp \iff x \cdot y = 0, \ \forall y \in \text{range}(A)
\]

\[
\iff x \cdot Aw = 0, \ \forall w
\]

\[
\iff (A^T x) \cdot w = 0, \ \forall w
\]

\[
\iff A^T x = 0
\]

\[
\iff x \in \text{null}(A^T).
\]

The second statement now follows from Theorem 6.4.5 by replacing \( A \) with \( A^T \).
6.5 Optimality Conditions for Linear Least Squares

The nearest-point theorem and complementarity theorems of the preceding section provide a unified approach to deriving usable optimality conditions for a larger class of related problems. The main idea is as follows:

1. reformulate the given problem into one of finding a nearest point in some subspace \(L\);
2. identify a usable representation of \(L^\perp\);
3. apply the optimality conditions \(\bar{x} \in L\) and \(z - \bar{x} \in L^\perp\);
4. translate the result into optimality conditions for the original problem.

In this section we provide examples of how this works. The goal in each case is to derive necessary and sufficient conditions in the form of linear systems, with existence of an optimum guaranteed by Theorem 6.4.1. A couple of the examples simply use the four-step procedure above to recover the results of earlier sections.

**Example 6.5.1** (Nearest point in the range of a matrix). Given \(A\) and \(z\), minimize \(\|x - z\|\) over all \(x \in \text{range}(A)\).

**Solution.** We take \(L = \text{range}(A)\) and \(L^\perp = \text{null}(A^T)\). Then we have

- \(x \in L \iff x = Ay\), for some \(y\);
- \(z - x \in L^\perp \iff A^T(z - x) = 0\).

So this problem requires solving the system consisting of \(A^T A y = A^T z\) and \(x = Ay\) for \(x\) and \(y\). [This recovers formula (c) of Theorem 6.4.1.]

**Example 6.5.2** (Nearest point in the null space of a matrix). Given \(A\) and \(z\), minimize \(\|x - z\|\) over all \(x \in \text{null}(A)\).

**Solution.** We take \(L = \text{null}(A)\) and \(L^\perp = \text{range}(A^T)\). Then we have

- \(x \in L \iff Ax = 0\);
- \(z - x \in L^\perp \iff z - x = A^T y\), for some \(y\).

So this problem requires solving the system consisting of \(AA^T y = Az\) and \(x = z - A^T y\) for \(x\) and \(y\).

**Example 6.5.3** (Unconstrained linear least squares). Given \(A\) and \(b\), minimize \(\|Ax - b\|\) over all \(x\).

**Solution.** First we recast this in terms of finding a nearest point in a subspace: make the substitution \(y = Ax\) and then minimize \(\|y - b\|\) over all \(y \in \text{range}(A)\). Now we take \(L = \text{range}(A)\) and \(L^\perp = \text{null}(A^T)\) to get the optimality conditions:

- \(y \in L \iff y = Ax\), for some \(x\);
- \(b - y \in L^\perp \iff A^T(b - y) = 0\).

So this problem requires solving the system

\[ A^T y = A^T b, \quad y = Ax \]

for \(x\) and \(y\). In this case, we can eliminate the vector \(y\) altogether, so that in fact we only
Need to solve
\[ A^T Ax = A^T b \]
for \( x \). [This recovers the normal equations of Section 6.2.]

**Example 6.5.4** (Minimum-norm solution of an underdetermined system). Given \( A \) and \( b \), minimize \( \|x\| \) over all \( x \) satisfying \( Ax = b \).

**Solution.** This is the special case of the next example in which \( z = 0 \).

**Example 6.5.5** (Nearest point among solutions of an underdetermined system). Given \( A \), \( b \) and \( z \), minimize \( \|x - z\| \) over all \( x \) satisfying \( Ax = b \).

**Solution.** Consider a fixed vector \( \bar{x} \) with \( A\bar{x} = b \). Using the notation \( x = \bar{x} + y \) gives us \( Ax = b \iff Ay = 0 \). This has the effect of translating the given problem into one about subspaces. Specifically, we see that \( \bar{x} \) solves \( \min_x \{ \|x - z\| \mid Ax = b \} \) if and only if \( 0 \) solves \( \min_y \{ \|y - (z - \bar{x})\| \mid Ay = 0 \} \). Taking \( L = \text{null}(A) \), the optimality conditions become

- \( 0 \in L \) (trivially satisfied);
- \( (z - \bar{x}) - 0 \in L^\perp = \text{range}(A^T) \iff z - \bar{x} = A^T w \), for some \( w \).

So our optimality conditions amount to saying that we must solve the system consisting of \( Ax = b \) and \( z - x = A^T w \) for \( x \) and \( w \).

Note that the optimality conditions just derived give a 2-step procedure: first solve \( A A^T w = A z - b \) for \( w \), and then set \( x = z - A^T w \). Here’s an illustration of this procedure.

**Example 6.5.6** (Solving a minimum-norm solution problem). Find the minimum-norm solution of the system
\[
\begin{bmatrix}
2 & 0 & 1 & \\
1 & 1 & -1 & \\
\end{bmatrix}
\begin{bmatrix}
x & \\
y & \\
\end{bmatrix}
= 
\begin{bmatrix}
6 & \\
4 & \\
\end{bmatrix}.
\]

**Solution.** According the final sentence of the preceding example (we have \( z = 0 \) here), we first solve \( A A^T w = -b \) for \( w \):
\[
\begin{bmatrix}
5 & 1 & \\
1 & 3 & \\
\end{bmatrix}
w = - \begin{bmatrix}
6 & \\
4 & \\
\end{bmatrix} \implies w = \begin{bmatrix}
1 & \\
1 & \\
\end{bmatrix}.
\]

Now we set \( x = -A^T w \) to get
\[
\begin{bmatrix}
x & \\
\end{bmatrix}
= \begin{bmatrix}
3 & \\
1 & \\
0 & \\
\end{bmatrix}.
\]

The reader can verify that \( Ax = b \).  

We close this section by deriving a result that generalizes all of the others in this chapter and provides a prototype for the optimality conditions in Chapter 7. The proof follows the same scheme as the preceding examples, but uses the more complicated choice of subspace discussed in the following lemma.

**Lemma 6.5.7** (Image of a nullspace). For matrices \( A \) and \( C \), the set \( \{Cy \mid y \in \text{null}(A)\} \) is a linear subspace whose orthogonal complement is \( \{z \mid C^T z \in \text{range}(A^T)\} \).
Proof. Let $L = \{Cy \mid y \in \text{null}(A)\}$. It is an exercise to show that $L$ is a linear subspace. Because null$(A)$ is a linear subspace, there exists a matrix $E$ such that null$(A) = \text{range}(E)$. (In other words, null$(A)$ is spanned by the columns of $E$.) Therefore, $y \in \text{null}(A)$ if and only if $y = Ew$ for some $w$, which allows us to write $L = \text{range}(CE)$. Hence $L^\perp = \text{null}(E^T C^T)$, so that $z \in L^\perp$ if and only if $C^T z \in \text{null}(E^T)$. On the other hand, we also have null$(E^T) = \text{range}(E)^\perp = \text{null}(A)^\perp = \text{range}(A^T)$. Thus $z \in L^\perp$ if and only if $C^T z \in \text{range}(A^T)$, as claimed.

Theorem 6.5.8 (Equality-constrained linear least squares). Consider matrices $A$ and $C$, along with vectors $b$ and $d$. Then $\bar{x}$ solves

$$
\min_x \|Cx - d\| \quad \text{s.t.} \quad Ax = b
$$

if and only if there exists a vector $\bar{u}$ so that $(\bar{x}, \bar{u})$ solves the linear system

$$
C^T Cx - C^T d = A^T u, \quad Ax = b.
$$

Proof. Consider a fixed vector $\bar{x}$ with $A\bar{x} = b$ and use the notation $x = \bar{x} + y$ to get $Ax = b \iff Ay = 0$. Then $\bar{x}$ solves (*) if and only if 0 solves $\min_y \{\|Cy - (d - C\bar{x})\| \mid Ay = 0\}$. Taking $L = \{Cy \mid y \in \text{null}(A)\}$, the substitution $w = Cy$ shows that $\bar{x}$ solves (*) if and only if 0 solves $\min_w \{\|w - (d - C\bar{x})\| \mid w \in L\}$. By Lemma 6.5.7 and the nearest-point theorem, the optimality conditions for the problem in $w$ are

- $0 \in L$ (trivially satisfied);
- $(d - C\bar{x}) - 0 \in L^\perp \iff C^T (d - C\bar{x}) = A^T u$, for some $u$.

So our optimality conditions for (*) amount to saying that we must solve

$$
Ax = b, \quad C^T (d - Cx) = A^T u
$$

for $x$ and $u$. Replacing $u$ by $-u$ gives the desired result.

In the terminology of Chapter 7, $u$ is the vector of Lagrange multipliers for the constraints $Ax = b$.

6.6 Nonnegative Linear Least Squares

We can include nonnegativity constraints in a least squares problem with little difficulty. Besides extending the results of earlier sections, this lays the groundwork for dealing with inequality constraints in more general problems.

Theorem 6.6.1. (Nonnegative least squares). Consider a matrix $A$ and vector $b$. Then $\bar{x}$ is a global minimizer for

$$
\min_x \|b - Ax\| \quad \text{s.t.} \quad x \geq 0
$$

if and only if there exists a vector $\bar{u}$ so that $(\bar{x}, \bar{u})$ solves the system

$$
x \geq 0, \quad u \geq 0, \quad x \cdot u = 0, \quad A^T (Ax - b) = u.
$$

67
**Proof.** To prove the sufficiency, suppose \((\bar{x}, \bar{u})\) satisfies the conditions stated in the theorem. The equation \(A^T(A\bar{x} - b) = \bar{u}\) implies that \(\bar{x}\) is a critical point (and hence a global minimizer) for the convex function \(\frac{1}{2}\|Ax - b\|^2 - \bar{u} \cdot x\). Given any \(x \geq 0\), we therefore have

\[
\frac{1}{2}\|Ax - b\|^2 \geq \frac{1}{2}\|Ax - b\|^2 - \bar{u} \cdot x \geq \frac{1}{2}\|A\bar{x} - b\|^2 - \bar{u} \cdot \bar{x} = \frac{1}{2}\|A\bar{x} - b\|^2.
\]

Hence \(\bar{x}\) is a global minimizer for \(\|Ax - b\|\) over \(x \geq 0\).

Now we prove the necessity. Suppose \(\bar{x}\) is optimal and define a diagonal matrix \(D\) by

\[
d_{ii} = \begin{cases} 0 & \text{if } \bar{x}_i > 0, \\ 1 & \text{if } \bar{x}_i = 0. \end{cases}
\]

Taking \(f(x) = \frac{1}{2}\|Ax - b\|^2\), we see that \(\bar{x}\) solves the equality-constrained problem

\[
\min_x f(x) \text{ s.t. } Dx = 0.
\]

By Theorem 6.5.8, there exists \(\bar{u}\) satisfying \(A^T A\bar{x} - A^T b = D\bar{u}\). Note that we may assume that \(D\bar{u} = \bar{u}\), so that \(\bar{u}_i = 0\) for any \(i\) with \(\bar{x}_i > 0\). This gives us \(\bar{x} \cdot \bar{u} = 0\). Hence, we have verified all the conditions in (*) except \(\bar{u} \geq 0\). If \(\bar{u}_i < 0\) for some \(i\), then

\[
\nabla f(\bar{x}) \cdot e^{(i)} = (A^T A\bar{x} - A^T b) \cdot e^{(i)} = \bar{u} \cdot e^{(i)} = \bar{u}_i < 0.
\]

Hence \(e^{(i)}\) is a feasible descent direction at \(\bar{x} \geq 0\), contradicting the optimality of \(\bar{x}\). \(\square\)

Here’s the most popular method for solving nonnegative least-squares problems.

**Method 6.6.2 (NNLS; Lawson and Hanson).** Start with the vector \(x^{(0)} = 0\) and \(k = 0\).

1. **[Optimality test]** Let \(u^{(k)} = A^T (Ax^{(k)} - b)\). If \(u^{(k)} \geq 0\), stop: \(x^{(k)}\) is optimal.
2. **[Build-up]** Choose \(j\) with \(u_j^{(k)} < 0\) and define \(D^{(k)}\) to be the diagonal matrix with

\[
d_{jj}^{(k)} = \begin{cases} 0 & \text{if } j = j \text{ or } x_i^{(k)} > 0, \\ 1 & \text{otherwise}. \end{cases}
\]

3. **[Solve LS subproblem]** Let \(z^{(k)}\) be a solution to

\[
\min_z \left\{ \frac{1}{2}\|Az - b\|^2 \bigg| \quad D^{(k)} z = 0 \right\}.
\]

If \(z^{(k)} \geq 0\), then set \(x^{(k+1)} = z^{(k)}\), replace \(k\) by \(k + 1\), and go to step 1.

4. **[Build-down]** Let \(t = \min_i \left\{ \frac{x_i^{(k)}}{x_i^{(k)} - z_i^{(k)}} \bigg| \quad z_i^{(k)} < 0 \right\}.

Set \(x^{(k+1)} = x^{(k)} + t(z^{(k)} - x^{(k)})\) and define \(D^{(k+1)}\) to be the diagonal matrix with

\[
d_{ii}^{(k+1)} = \begin{cases} 0 & \text{if } x_i^{(k+1)} > 0, \\ 1 & \text{if } x_i^{(k+1)} = 0. \end{cases}
\]

Replace \(k\) by \(k + 1\), and go to step 3.
Theorem 6.6.3 (Validation of the NNLS method). Method 6.6.2 finds an optimal solution (and confirms its optimality) in a finite number of iterations. In particular, the nonnegative linear least squares problem always admits a solution.

Proof. To see why the method must terminate, consider these two observations:

(a) The build-down cycle consisting of step 4 followed by step 3 can be executed at most $n$ times before returning to step 1, because each pass adds new positive entries to the diagonal of $D^{(k)}$. Moreover, no pass through this cycle can increase the objective value $\frac{1}{2} \|Ax^{(k)} - b\|^2$, because the objective is convex and $z^{(k)}$ is optimal in the subproblem.

(b) Each execution of step 3 following steps 1 and 2 decreases the objective value, because the build-up in step 2 puts a guaranteed descent direction into the subproblem.

Together, (a) and (b) imply that the objective value eventually decreases after each instance in which the termination criterion is denied, and that each such decrease is eventually followed by yet another test of the termination criterion. On the other hand, only finitely many possible subproblems are solved in step 3, because there are only finitely many possible matrices $D^{(k)}$. The repeated decrease of the objective value shows that none of these subproblems occurs more than once, so the method cannot run forever. Hence, the termination criterion must be satisfied eventually.

To see that termination implies optimality, note that $x^{(k)} \geq 0$ for all $k$ during Method 6.6.2. If the method stops in step 1 when $k = 0$, then we clearly have $x^{(0)} \cdot u^{(0)} = 0$. Otherwise, step 1 can only follow step 3, in which case we have $A^T Ax^{(k)} - A^T b = u^{(k)} = D^{(k-1)} u^{(k)}$. This again shows that $x^{(k)} \cdot u^{(k)} = 0$. In either case, the claim of optimality in step 1 is justified by Theorem 6.6.1.

In Section 6.4, we proved $\text{null}(A) \perp = \text{range}(A^T)$, which says that $c \cdot d = 0$ for all $d$ with $Ad = 0$ if and only if $c = A^T u$ for some $u$. In other words, the equation $c^T d = 0$ is a "consequence" of the system $Ad = 0$ if and only if $c^T$ is expressible as a linear combination of the rows of $A$. Here is the analogous result for systems of inequalities.

Lemma 6.6.4 (Farkas). Given a matrix $A$ and vector $c$, we have $c \cdot d \leq 0$ for all $d$ with $Ad \leq 0$ if and only if $c = A^T u$ for some $u \geq 0$. In other words, the inequality $c^T d \leq 0$ is a consequence of the system $Ad \leq 0$ if and only if $c^T$ can be expressed as a nonnegative combination of the rows of $A$.

Proof. The implication $\Leftarrow$ follows immediately from the fact that $d \cdot A^T u = (Ad) \cdot u$. To prove the converse, consider the nonnegative least squares problem

$$\min_u \{\|A^T u - c\| \mid u \geq 0\}.$$  

From Theorems 6.6.1 and 6.6.3, we know that there exists an optimizer $u \geq 0$ satisfying

$$A(A^T u - c) \geq 0, \quad u \cdot A(A^T u - c) = 0.$$  

Now consider $d = c - A^T u$. Because $Ad \leq 0$, we have $c \cdot d \leq 0$. Thus

$$0 \leq \|c - A^T u\|^2 = c \cdot (c - A^T u) - u \cdot A(c - A^T u) = c \cdot d - 0 \leq 0.$$
Hence $\|c - A^T u\| = 0$, so $c = A^T u$.  

The Farkas Lemma is a basic tool in the theory of linear inequalities. We use it in later chapters to derive optimality conditions for linear and nonlinear programming problems.
7. GENERAL NONLINEAR PROGRAMMING

In this chapter we develop necessary and sufficient conditions for optimality for the general nonlinear programming problem

\[
\text{(NLP)} \quad \text{minimize } f(x) \text{ subject to } g_i(x) = 0, \text{ for } i = 1, \ldots, m, \\
g_i(x) \leq 0, \text{ for } i = m + 1, \ldots, p.
\]

7.1 Equality-Constrained Problems — Lagrange Multipliers

We begin with the special case of linear-equality constrained problems:

\[
\text{(LEP)} \quad \text{minimize } f(x) \text{ subject to } Ax = b.
\]

This is the case of \text{(NLP)} in which \(g(x) = Ax - b\) and \(p = m\). It provides a simple prototype of the results in this chapter.

**Theorem 7.1.1.** Suppose \(f : \mathbb{R}^n \to \mathbb{R}\) is differentiable at \(\bar{x}\) and assume that \(A\bar{x} = b\).

(a) If \(\bar{x}\) is a local minimizer for \text{(LEP)}, then \(\nabla f(\bar{x}) \in \text{range}(A^T)\).

(b) If \(f\) is convex and \(\nabla f(\bar{x}) \in \text{range}(A^T)\), then \(\bar{x}\) is a global minimizer for \text{(LEP)}.

**Observation 7.1.2.** The condition \(\nabla f(\bar{x}) \in \text{range}(A^T)\) can be rewritten as

\[
\nabla f(\bar{x}) + A^T \bar{\lambda} = 0.
\]

In other words,

\[
\nabla f(\bar{x}) + \sum_{i=1}^{m} \bar{\lambda}_i a^{(i)} = 0,
\]

where

\[
A^T = [a^{(1)}, \ldots, a^{(m)}].
\]

We call \(\bar{\lambda}_i\) the Lagrange multiplier for the constraint \(a^{(i)} \cdot x = b_i\).

**Proof of Theorem 7.1.1.** Part (a): Assume \(\bar{x}\) is a local minimizer for \text{(LEP)}. To show \(\nabla f(\bar{x}) \in \text{range}(A^T) = \text{null}(A)^\perp\), we must show that \(\nabla f(\bar{x}) \cdot d = 0\) whenever \(Ad = 0\). (See Figure 7.1.) Consider \(d \in \text{null}(A)\) and define \(\varphi_d(t) = f(\bar{x} + td)\). Because \(A(\bar{x} + td) = A\bar{x} + Ad = b + 0 = b\), we see that \(\bar{x} + td\) satisfies the constraints of \text{(LEP)} for all \(t\). Thus \(\bar{t} = 0\) is a local minimizer for \(\varphi_d\), and hence \(0 = \varphi_d'(0) = \nabla f(\bar{x}) \cdot d\) as desired.

Part (b): Assume that \(f\) is convex, \(A\bar{x} = b\), and \(\nabla f(\bar{x}) = A^T \bar{\lambda}\). By convexity, each vector \(x\) must satisfy \(f(x) \geq f(\bar{x}) + \nabla f(\bar{x}) \cdot (x - \bar{x}) = f(\bar{x}) + (A^T \bar{\lambda}) \cdot (x - \bar{x}) = f(\bar{x}) + \bar{\lambda} \cdot A(x - \bar{x})\). If \(x\) is feasible (so \(Ax = b = A\bar{x}\)), then this implies that \(f(x) \geq f(\bar{x}) + \bar{\lambda} \cdot 0 = f(\bar{x})\). Hence \(\bar{x}\) is a global minimizer for \text{(LEP)}.

\[\square\]
Example 7.1.3 (Equality-constrained linear least squares). Consider the problem of minimizing \( \|Cx - d\| \) subject to \( Ax = b \), which was discussed at the end of Chapter 6. This is equivalent to minimizing \( f(x) = \frac{1}{2}\|Cx - d\|^2 \) subject to \( Ax = b \). We write \( f(x) \) out as

\[
f(x) = \frac{1}{2}\|Cx - d\|^2 = \frac{1}{2}(Cx - d) \cdot (Cx - d) = \frac{1}{2}[x \cdot C^TCx - 2(C^Td) \cdot x + d \cdot d]
\]

and differentiate to get \( \nabla f(x) = C^TCx - C^Td \) and \( Hf(x) = C^TC \). The necessary and sufficient conditions for optimality are therefore \( A\bar{x} = b \) and \( C^T(C\bar{x} - d) = A^Tu \) for some \( u \), which recovers the result found in Section 6.5.

Next, we consider problems with general nonlinear equality constraints

\((\text{EP})\)

minimize \( f(x) \) subject to \( g_i(x) = 0 \), for \( i = 1, \ldots, m \).

The classical theorem in this case is very similar to that for the linearly constrained problem, but has an additional requirement.

Definition 7.1.4 (Regular points). We say that \( \bar{x} \) is a regular point if the constraint gradients \( \nabla g_i(\bar{x}) \) (for \( i = 1, \ldots, m \)) are linearly independent. Equivalently, \( \bar{x} \) is regular if the derivative matrix \( g'(\bar{x}) \) has full row rank.

Theorem 7.1.5 (Lagrange multipliers). Suppose \( f : \mathbb{R}^n \to \mathbb{R} \) and \( g : \mathbb{R}^n \to \mathbb{R}^m \) are continuously differentiable and consider a local minimizer or maximizer \( \bar{x} \) for \((\text{EP})\). If \( \bar{x} \) is a regular point, then there exists \( \bar{\lambda} \in \mathbb{R}^m \) such that

\[
\nabla f(\bar{x}) + \sum_{i=1}^{m} \bar{\lambda}_i \nabla g_i(\bar{x}) = 0.
\]

Proof. This theorem is a special case of Theorem 7.3.2, which is proved in Section 7.3.

Observation 7.1.6 (Linearization of the constraints). This theorem can be restated as saying that \( \nabla f(\bar{x}) \in \text{range}[g'(\bar{x})]^T \) at any local minimizer that is also regular. If we
compare this with Theorem 7.1.1, we see that the Lagrange multiplier theorem amounts to saying that optimality in the problem \( \min_x \{ f(x) \mid g(\bar{x}) + g'(\bar{x})(x - \bar{x}) = 0 \} \) with “linearized constraints” is a necessary condition for optimality in the original nonlinear problem \( \min_x \{ f(x) \mid g(x) = 0, \} \), provided that \( g'(\bar{x}) \) has full row rank.

**Observation 7.1.7** (Constraint qualification). The regularity condition is a *constraint qualification*. It guarantees that the linearization of the constraint function \( g \) provides reliable information about tangents to the feasible set.

**Example 7.1.8** (The trouble with irregular points). Consider the constraint \( g_1(x_1, x_2) = x_1^3 - x_2^3 = 0 \). Note that the feasible set is \( \{ x \in \mathbb{R}^2 \mid x_1 = x_2 \} \) and that the origin \( x = 0 \) is an irregular point:

\[
\nabla g_1(x) = \begin{bmatrix} 3x_1^2 \\ -3x_2^2 \end{bmatrix} = 0 \text{ when } x = 0.
\]

The linearization of the constraint function is \( g(0) + \nabla g(0) \cdot (x - 0) = 0 \), which says that \( 0 = 0 \). This provides no information at all about tangents to the feasible set at \( x = 0 \).

**Observation 7.1.9** (Identifying candidates for optimality). Because of the regularity assumption, the Lagrange multiplier theorem identifies candidates for optimality as all feasible points that are either

- irregular points (if any \( g_i \) is nonlinear), or
- solutions to \( \nabla f(x) + \sum \lambda_i \nabla g_i(x) = 0 \).

Accordingly, our treatment of example problems in the next few sections follows the general outline shown in Figure 7.2.

---

**Analyzing a Constrained Optimization Problem**

I. Consider the problem statement for clues about its structure.
   - Sketch graphs if possible.
   - Reformulate algebraically to improve structure or eliminate variables.
   - Calculate derivatives.

II. Check preliminary criteria concerning global optimizers.
   - Coercivity? Boundedness? Convexity?

III. Find all feasible irregular points. (Skip this step if all constraints are linear.)

IV. Solve for all feasible points satisfying the Lagrange multiplier equation (or KKT conditions — see Section 7.4).

V. Test candidates for sufficiency and draw conclusions.

Steps III and IV often involve a numerical procedure such as Newton’s Method.
7.2 Equality-Constrained Problems — Examples

This presents examples illustrating the direct use of the Lagrange multiplier theorem. In each case, however, it’s also possible to use the constraints to eliminate variables and reduce the dimension of the problem.

Example 7.2.1. \( \min / \max \{ x_1^2 + x_2^2 + x_3^2 \mid x_1^2 + \frac{1}{4} x_2^2 + \frac{1}{9} x_3^2 = 1 \} \)

Solution.

I. Problem considerations:
\( f(x) = x_1^2 + x_2^2 + x_3^2, \quad g_1(x) = x_1^2 + \frac{1}{4} x_2^2 + \frac{1}{9} x_3^2 = 1. \)

• problem structure: seeks the points on an ellipsoid (centered at the origin) that closest to, or farthest from, the origin.

• derivatives: \( \nabla f(x) = \begin{bmatrix} 2x_1 \\ 2x_2 \\ 2x_3 \end{bmatrix}, \quad \nabla g_1(x) = 2 \begin{bmatrix} x_1 \\ x_2/4 \\ x_3/9 \end{bmatrix} \)

II. Existence:

• constraint \( \Rightarrow |x_1| \leq 1, |x_2| \leq 2, |x_2| \leq 3 \Rightarrow \) feasible region is bounded

• continuity of constraint function \( \Rightarrow \) feasible region is closed

Therefore, global max/min exist by Weierstrass Theorem.

III. Irregular points: \( \nabla g_1(x) = 0 \) only at \( x = 0 \), which is infeasible.

IV. Lagrange multiplier equations:
\( \begin{bmatrix} 2x_1 \\ 2x_2 \\ 2x_3 \end{bmatrix} + \lambda \begin{bmatrix} 2x_1 \\ x_2/2 \\ 2x_3/9 \end{bmatrix} = 0 \)

This has the form
\( \begin{bmatrix} 2(1 + \lambda)x_1 \\ 2(1 + \lambda/4)x_2 \\ 2(1 + \lambda/9)x_3 \end{bmatrix} = 0. \)

Since we do not have \( x = 0 \), we must have one of \( \lambda = -1, -4, -9. \)

1. \( \lambda = -1 \Rightarrow x_2 = x_3 = 0 \Rightarrow x_1 = \pm 1 \) (by constraint). Here we have \( f(x) = 1. \)

2. \( \lambda = -4 \Rightarrow x_1 = x_3 = 0 \Rightarrow x_2 = \pm 2 \) (by constraint). Here we have \( f(x) = 4. \)

3. \( \lambda = -9 \Rightarrow x_1 = x_2 = 0 \Rightarrow x_3 = \pm 3 \) (by constraint). Here we have \( f(x) = 9. \)

V. Conclusions: global max at \( \pm \begin{bmatrix} 0 \\ 0 \\ 3 \end{bmatrix} \), global min at \( \pm \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}. \)
Example 7.2.2. \( \min / \max \{ x_1^2 + 4x_1x_2 + x_2^2 \mid x_1^2 + x_2^2 = 1 \} \)

Solution.

I. Problem considerations: \( f(x) = x_1^2 + 4x_1x_2 + x_2^2 \), \( g_1(x) = x_1^2 + x_2^2 - 1 \).
   - problem structure: minimize or maximize a quadratic form (elliptical level curves) over the unit circle
   - derivatives: \( \nabla f(x) = \begin{bmatrix} 2x_1 + 4x_2 \\ 4x_1 + 2x_2 \end{bmatrix}, \quad \nabla g_1(x) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} \)

II. Existence:
   - constraint \( \Rightarrow |x_1| \leq 1, |x_2| \leq 1 \Rightarrow \) feasible region is bounded
   - continuity of constraint function \( \Rightarrow \) feasible region is closed
Therefore global max/min exist by Weierstrass Theorem.

III. Irregular points: \( \nabla g_1(x) = 0 \) only at \( x = 0 \), which is infeasible.

IV. Lagrange multiplier equations: \( \begin{bmatrix} 2x_1 + 4x_2 \\ 4x_1 + 2x_2 \end{bmatrix} + \lambda \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} = 0 \)
To solve: for fixed \( \lambda \) this is linear in \( x \), with solutions satisfying

\[ [4 - (1 + \lambda)^2]x_2 = 0, \quad x_1 = -\frac{1}{2}(1 + \lambda)x_2. \]

Our cases depend on the coefficient of \( x_2 \) in the first of these.

A. \( 4 - (1 + \lambda)^2 \neq 0 \Rightarrow x_2 = 0 \Rightarrow x_1 = 0 \): infeasible!

B. \( 4 - (1 + \lambda)^2 = 0 \Rightarrow \lambda = 1, -3 \).
   1. \( \lambda = 1 \Rightarrow x_1 = -x_2 \Rightarrow x = \pm \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} \) (by constraint). In this case \( f(x) = -1 \).
   2. \( \lambda = -3 \Rightarrow x_1 = x_2 \Rightarrow x = \pm \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \) (by constraint). In this case \( f(x) = 3 \).

V. Conclusions: global max at \( \pm \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \), global min at \( \pm \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} \). \( \square \)
Example 7.2.3. \( \min / \max \{3x_1 - x_2 - 3x_3 \mid x_1 + x_2 - x_3 = 0, \ x_1^2 + 2x_3^2 = 1\} \)

Solution.

I. Problem considerations:
\[
f(x) = 3x_1 - x_2 - 3x_3, \quad g_1(x) = x_1 + x_2 - x_3, \quad g_2(x) = x_1^2 + 2x_3^2 - 1.
\]
- problem structure: optimize a linear function over the intersection of a plane (through the origin) with an elliptical cylinder (centered at the origin)
- derivatives:
  \[
  \nabla f(x) = \begin{bmatrix} 3 \\ -1 \\ -3 \end{bmatrix}, \quad \nabla g_1(x) = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}, \quad \nabla g_2(x) = \begin{bmatrix} 2x_1 \\ 0 \\ 4x_3 \end{bmatrix}
  \]

II. Existence:
- 2nd constraint \( \Rightarrow |x_1| \leq 1, \ |x_3| \leq 1/\sqrt{2} \)
- 1st constraint \( \Rightarrow |x_2| = |x_1 - x_3| \leq |x_1| + |x_3| \leq 1 + 1/\sqrt{2} \)

conclude that feasible region is bounded
- continuity of constraint functions \( \Rightarrow \) feasible region is closed.

Therefore, global max/min exist by Weierstrass Theorem.

III. Irregular points: linear dependence for two vectors means that one is a linear combination of the other. Because the second entries of \( \nabla g_1, \nabla g_2 \) allow this only if \( \nabla g_2(x) = 0, \) which requires \( x = 0. \) But \( x = 0 \) is infeasible.

IV. Lagrange multiplier equations:
\[
\begin{bmatrix} 3 \\ -1 \\ -3 \end{bmatrix} + \lambda_1 \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} + \lambda_2 \begin{bmatrix} 2x_1 \\ 0 \\ 4x_3 \end{bmatrix} = 0
\]

The second row gives \( \lambda_1 = 1, \) reducing the system to
\[
4 + 2\lambda_2 x_1 = 0, \quad -4 + 4\lambda_2 x_3 = 0.
\]

Solving in terms of \( \lambda_2 \) and substituting into the constraints yields equations
\[
\frac{3}{\lambda_2} = x_2, \quad \frac{6}{\lambda_2^2} = 1.
\]

This leads to \( x = \frac{\pm 1}{\sqrt{6}} \begin{bmatrix} 2 \\ -3 \\ -1 \end{bmatrix}, \) depending on \( \lambda_2 = \mp \sqrt{6}. \) The objective values are \( \pm 2\sqrt{6}. \)

V. Conclusions: global max at \( \frac{1}{\sqrt{6}} \begin{bmatrix} 2 \\ -3 \\ -1 \end{bmatrix}, \) global min at \( \frac{-1}{\sqrt{6}} \begin{bmatrix} 2 \\ -3 \\ -1 \end{bmatrix}. \) \( \Box \)
Example 7.2.4. \( \min / \max \{ x_1^2 - x_1 x_2 + x_2^2 + 2x_3^2 \mid x_1 + x_2 + x_3 = 1, \ x_1 - x_2 - 2x_3 = 2 \} \)

Solution.

I. Problem considerations:

\[ f(x) = x_1^2 - x_1 x_2 + x_2^2 + 2x_3^2, \quad g_1(x) = x_1 + x_2 + x_3 - 1, \quad g_2(x) = x_1 - x_2 - 2x_3 - 2. \]

- problem structure: optimize quadratic form (PSD \( \Rightarrow \) ellipsoidal level sets) over a line (intersection of two planes)

- derivatives: \( \nabla f(x) = \begin{bmatrix} 2x_1 - x_2 \\ -x_1 + 2x_2 \\ \frac{4x}{3} \end{bmatrix}, \quad \nabla g_1(x) = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \nabla g_2(x) = \begin{bmatrix} 1 \\ -1 \\ -2 \end{bmatrix} \)

II. Existence: The constraints represent the intersection of two planes, which must be a line or a plane—this isn’t bounded. However, the objective function is coercive: it is a quadratic with positive definite Hessian. A continuous coercive function attains a global minimum over a closed set; because the feasible region is unbounded, coercivity also tells us there are no global maximizers.

III. Irregular points: The constraints are linear, so we don’t need to worry about regularity.

IV. Lagrange multiplier equations:

\[ \begin{bmatrix} 2x_1 - x_2 \\ -x_1 + 2x_2 \\ \frac{4x}{3} \end{bmatrix} + \lambda_1 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \lambda_2 \begin{bmatrix} 1 \\ -1 \\ -2 \end{bmatrix} = 0 \]

These equations are linear in \( x \) and \( \lambda \) together, so we may combine them with the constraints to get one large linear system of equations:

\[ \begin{bmatrix} 2 & -1 & 0 & 1 & 1 \\ -1 & 2 & 0 & 1 & -1 \\ 0 & 0 & 4 & 1 & -2 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & -1 & -2 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 2 \end{bmatrix}. \]

The unique solution is \( x = \begin{bmatrix} 26/21 \\ 2/7 \\ -11/21 \end{bmatrix}, \ \lambda = \begin{bmatrix} -16/21 \\ -10/7 \end{bmatrix} \) with \( f(x) = 370/147. \)

V. Conclusions: Global minimum at \( x = \begin{bmatrix} 26/21 \\ 2/7 \\ -11/21 \end{bmatrix} \); no global maximum.

[Note: lack of global max is also indicated by linearity of constraints and uniqueness of solution to Lagrange multiplier equations.]
7.3 Inequality-Constrained Problems — Necessary Conditions

First, suppose we have a single inequality constraint: \( \min_x f(x) \) s.t. \( g_1(x) \leq 0 \). There are two cases, depending on whether the local minimizer \( \bar{x} \) is a boundary or interior point:

- If \( g_1(\bar{x}) = 0 \), the Lagrange Theorem gives us \( \nabla f(\bar{x}) = -\lambda_1 \nabla g_1(\bar{x}) \), for some \( \lambda_1 \geq 0 \).
- If \( g_1(\bar{x}) = 0 \), the critical point condition gives us \( \nabla f(\bar{x}) = 0 \), so that \( \lambda_1 = 0 \).

In fact, we can cover both cases by incorporating them in a pair of equations:

\[
\nabla f(x) + \lambda_1 \nabla g_1(x) = 0, \quad \text{(multiplier equation)} \\
\lambda_1 g_1(x) = 0. \quad \text{(complementary slackness)}
\]

We can extend this idea to the general nonlinear programming problem:

(NLP) \[ \text{minimize } f(x) \text{ subject to } g_i(x) = 0, \text{ for } i = 1, \ldots, m, \\
g_i(x) \leq 0, \text{ for } i = m + 1, \ldots, p. \]

**Definition 7.3.1** (Active constraints and regularity). The active constraints at a point \( \bar{x} \) are those with indices in the set

\[ J(\bar{x}) = \{ i \mid 1 \leq i \leq p, \ g_i(\bar{x}) = 0 \}. \]

We say \( \bar{x} \) is a regular point if the gradients \( \nabla g_i(\bar{x}) \) for \( i \in J(\bar{x}) \) are linearly independent.

**Theorem 7.3.2** (Karush, Kuhn, and Tucker). Suppose \( f : \mathbb{R}^n \to \mathbb{R} \) and \( g : \mathbb{R}^n \to \mathbb{R}^p \) are continuously differentiable. Assume that \( \bar{x} \) is feasible for (NLP) and that either

- (a) the active constraints at \( \bar{x} \) are all linear, or
- (b) \( \bar{x} \) is a regular point for (NLP).

If \( \bar{x} \) is a local minimizer for (NLP), then there exists \( \bar{\lambda} \in \mathbb{R}^p \) such that

1. \( \bar{\lambda}_i \geq 0 \) for all \( i > m \);
2. \( \bar{\lambda}_i g_i(\bar{x}) = 0 \) for all \( i \);
3. \( \nabla f(\bar{x}) + \sum_{i=1}^p \bar{\lambda}_i \nabla g_i(\bar{x}) = 0. \)

If \( \bar{x} \) is a local maximizer, then condition (1) is replaced by \( \bar{\lambda}_i \leq 0 \) for all \( i > m \).

**Proof.** We may assume \( J(\bar{x}) = \{1, \ldots, p\} \), because the continuity of \( g_i \) for \( i \notin J(\bar{x}) \) prevents \( g_i \) from taking the value 0 on some neighborhood of \( \bar{x} \): for these \( i \), (2) amounts to choosing \( \bar{\lambda}_i = 0 \). Also, note that condition (3) is equivalent to \( f'(\bar{x}) + \bar{\lambda}^T g'(\bar{x}) = 0 \). We shall prove the theorem in three cases separately.

**Case 1: assumption (a) holds.** In this case, we may write the active constraints as \( Ax - b \leq 0 \), which accounts for equality constraints as well as inequality constraints. We must show that \( \bar{\lambda} \geq 0 \) exists such that \( \nabla f(\bar{x}) + A^T \bar{\lambda} = 0 \). By the Farkas Lemma (6.6.4), it suffices to show that \( -\nabla f(\bar{x}) \cdot d \leq 0 \) for all \( d \) with \( Ad \leq 0 \). So assume \( Ad \leq 0 \). If \( -\nabla f(\bar{x}) \cdot d > 0 \), then \( d \) is a descent direction for \( f \) at \( \bar{x} \). Moreover, \( A(\bar{x} + td) \leq b \) for
all $t > 0$. So any small motion along the direction $d$ from $\bar{x}$ yields a feasible point with a smaller value of the objective than at $\bar{x}$, which contradicts the local optimality of $\bar{x}$.

**Case 2: assumption (b) holds and there are no equality constraints.** Our proof of this case relies on the following.

Claim: For every vector $z < 0$, there are no solutions $d$ to the linear system

$$
\begin{bmatrix}
    f'(\bar{x}) \\
g'(\bar{x})
\end{bmatrix} d = z.
$$

Proof of claim: Suppose the linear system does have a solution $d$ for some negative vector $z$. By the assumed differentiability, there exist functions $u : \mathbb{R} \to \mathbb{R}$ and $w : \mathbb{R} \to \mathbb{R}^p$ satisfying $\lim_{t \to 0} |u(t)|/t = 0$, $\lim_{t \to 0} \|w(t)\|/t = 0$, and

$$
\begin{bmatrix}
f(\bar{x} + td) \\
g(\bar{x} + td)
\end{bmatrix} = \begin{bmatrix}
f(\bar{x}) + tf'(\bar{x})d + u(t) \\
g(\bar{x}) + tg'(\bar{x})d + w(t)
\end{bmatrix} = \begin{bmatrix}
f(\bar{x}) \\
0
\end{bmatrix} + t \begin{bmatrix}
z + \begin{bmatrix} u(t)/t \\
w(t)/t \end{bmatrix}
\end{bmatrix}
$$

for all $t \neq 0$. Thus, for all sufficiently small $t > 0$, we have $g(\bar{x} + td) < 0$ and $f(\bar{x} + td) < f(\bar{x})$. This contradicts the minimality of $\bar{x}$, and so the claim is true. This claim shows that some $\bar{\lambda} \in \mathbb{R}^p$ must satisfy condition (3): otherwise, the matrix

$$
\begin{bmatrix}
f'(\bar{x}) \\
g'(\bar{x})
\end{bmatrix}
$$

would have full row rank and the linear system in the claim would admit a solution $d$ for *any* vector $z$. We now show that such a multiplier vector $\bar{\lambda}$ must actually satisfy $\bar{\lambda} \geq 0$. Suppose, to the contrary, that some entry in $\bar{\lambda}$ is negative: then we can choose a vector $q \in \mathbb{R}^p$ satisfying $q < 0$ and $\bar{\lambda} \cdot q > 0$.$\dagger$ By the full row-rank assumption (b), we know that the linear system $g'(\bar{x})d = q$ has a solution $d$. Combining this with (3) yields

$$
\begin{bmatrix}
f'(\bar{x}) \\
g'(\bar{x})
\end{bmatrix} d = \begin{bmatrix} -\bar{\lambda}^T g'(\bar{x})d \\ -\bar{\lambda} \cdot q \end{bmatrix} < 0.
$$

Therefore, the linear system in the claim has a solution $d$ for a specific vector $z < 0$. Hence, we must have $\bar{\lambda} \geq 0$ after all, which concludes the proof of the theorem in this case.

**Case 3: assumption (b) holds and there are equality constraints.** Suppose $\bar{x}$ is a minimizer for \((\text{NLP})\) over the neighborhood $\bar{B}(\bar{x}, \delta)$, where $\delta > 0$. We may choose $\delta$ so that $\{\nabla g_i(x) \mid i \in I(\bar{x})\}$ remains linearly independent for all $x \in \bar{B}(\bar{x}, \delta)$ (apply Proposition A.1.7 to the transpose). Consider the sequence of inequality-constrained problems

\[(\text{NLP}^{(j)})\]

\[
\begin{align*}
\text{minimize } & f(x) + \|x - \bar{x}\|^2 \\
\text{subject to } & g_i(x) - 1/j \leq 0, \text{ for } i = 1, \ldots, m, \\
& -g_i(x) + 1/j \leq 0, \text{ for } i = 1, \ldots, m, \\
& g_i(x) \leq 0, \text{ for } i = m + 1, \ldots, p, \\
& \|x - \bar{x}\|^2 - \delta^2 \leq 0.
\end{align*}
\]

$\dagger$ For example, let $q = \bar{\lambda}_i \mathbf{1}_p - (1 + |\mathbf{1}_p^T \bar{\lambda}|) \mathbf{e}^{(i)}$ for some index $i$ with $\bar{\lambda}_i < 0$, where $\mathbf{e}^{(i)} \in \mathbb{R}^p$ is the $i^{th}$ coordinate vector.
Note that \( \bar{x} \) is feasible for (\( \text{NLP}^{(j)} \)) and has the same objective value as in (\( \text{NLP} \)). Because the feasible region for (\( \text{NLP}^{(j)} \)) is closed and bounded, the Weierstrass theorem guarantees the existence of an optimal solution \( \bar{x}^{(j)} \). Moreover, the optimality of \( \bar{x}^{(j)} \) implies that 
\[
 f(\bar{x}^{(j)}) + \| \bar{x}^{(j)} - \bar{x} \|^2 \leq f(\bar{x}).
\]
By the Bolzano-Weierstrass theorem (Theorem 3.1.6), there exists a subsequence \( \{\bar{x}^{(jk)}\} \) which converges to some point \( z \). By the continuity of \( f \) and each \( g_i \), we see that \( z \) is feasible for (\( \text{NLP} \)) and 
\[
 f(z) + \| z - \bar{x} \|^2 \leq f(\bar{x}) < f(z).
\]
Hence \( \| z - \bar{x} \|^2 = 0 \), so that \( z = \bar{x} \). We therefore have \( \bar{x}^{(jk)} \to \bar{x} \).

We now show that \( \bar{x}^{(jk)} \) satisfies the necessary conditions for (\( \text{NLP}^{(jk)} \)). For all \( k \) sufficiently large, we have \( \| \bar{x}^{(jk)} - \bar{x} \|^2 - \delta^2 < 0 \), so the last constraint in (\( \text{NLP}^{(jk)} \)) is inactive at \( \bar{x}^{(jk)} \). Also, for each \( i = 1, \ldots, m \), at most one of the two constraints \( g_i(x) - 1/j_k \leq 0 \) and \( -g_i(x) + 1/j_k \leq 0 \) is active at \( \bar{x}^{(jk)} \). Therefore, \( \bar{x}^{(jk)} \) is a regular point for (\( \text{NLP}^{(jk)} \)), so we may apply Case 2 to obtain multipliers \( \bar{\nu}_i^{(k)} \geq 0 \) and \( \bar{\mu}_i^{(k)} \geq 0 \) for \( i = 1, \ldots, m \), along with multipliers \( \bar{\lambda}_i^{(k)} \geq 0 \) for \( i = m+1, \ldots, p \), so that
\[
 \bar{\nu}_i^{(k)}(g_i(\bar{x}^{(jk)}) - 1/j_k) = 0, \quad \bar{\mu}_i^{(k)}(-g_i(\bar{x}^{(jk)}) + 1/j_k) = 0, \quad \bar{\lambda}_i^{(k)}g_i(\bar{x}^{(jk)}) = 0,
\]
and
\[
 \left[ \nabla f(\bar{x}^{(jk)}) + 2(\bar{x}^{(jk)} - \bar{x}) \right] + \sum_{i=1}^m \left[ \bar{\nu}_i^{(k)} - \bar{\mu}_i^{(k)} \right] \nabla g_i(\bar{x}^{(jk)}) + \sum_{i=m+1}^p \bar{\lambda}_i^{(k)} \nabla g_i(\bar{x}^{(jk)}) = 0.
\]

Finally, we use these optimality conditions to derive the desired conditions for \( \bar{x} \) in (\( \text{NLP} \)). Define \( \bar{\lambda}_i^{(k)} \in \mathbb{R} \) for \( i = 0, \ldots, m \) by
\[
 \bar{\lambda}_i^{(k)} = \begin{cases} 
 1, & \text{if } i = 0, \\
 \bar{\nu}_i^{(k)} - \bar{\mu}_i^{(k)}, & \text{if } i = 1, \ldots, m
 \end{cases}
\]
and let \( \lambda^{(k)} \in \mathbb{R}^{1+p} \) denote the unit vector \( \lambda^{(k)} = \bar{\lambda}^{(k)}/\| \bar{\lambda}^{(k)} \| \). Note that
\[
 \lambda_0^{(k)} \left[ \nabla f(\bar{x}^{(jk)}) + 2(\bar{x}^{(jk)} - \bar{x}) \right] + \sum_{i=1}^p \lambda_i^{(k)} \nabla g_i(\bar{x}^{(jk)}) = 0.
\]

By invoking the Bolzano-Weierstrass theorem if necessary (to extract a further subsequence of \( \{ \bar{x}^{(jk)} \} \)), we may assume \( \{ \lambda^{(k)} \} \) converges to some unit vector \( \lambda \). Letting \( k \to \infty \) yields
\[
 \lambda_0 \nabla f(\bar{x}) + \sum_{i=1}^p \lambda_i \nabla g_i(\bar{x}) = 0,
\]
along with \( \lambda_i g_i(\bar{x}) = 0 \) for \( i = 1, \ldots, p \) and \( \lambda_i \geq 0 \) for \( i = 0 \) and \( i = m+1, \ldots, p \). We must have \( \lambda_0 > 0 \), since \( \lambda_0 = 0 \) would imply linear dependence of the active constraints at \( \bar{x} \). Thus we may define \( \bar{\lambda}_i = \lambda_i/\lambda_0 \) for each \( i \) to obtain the desired conditions (1)–(3) at \( \bar{x} \). \( \square \)
7.4 Inequality-Constrained Problems — Examples

Here are three more examples.

Example 7.4.1. \( \min / \max \{ x_1 x_2 \mid x_1^2 + x_2^2 \leq 1 \} \)

Solution.

I. Problem considerations: \( f(x) = x_1 x_2, \quad g_1(x) = x_1^2 + x_2^2 - 1. \)
   - problem structure: optimize an indefinite quadratic form (hyperbolic level sets) over the unit disk
   - derivatives: \( \nabla f(x) = \begin{bmatrix} x_2 \\ x_1 \end{bmatrix}, \quad \nabla g_1(x) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} \)

II. Existence: Feasible region is closed, bounded and nonempty; objective is continuous. So Weierstrass Theorem guarantees existence of global max/min.

III. Irregular points: \( \nabla g_1(x) = 0 \iff x = 0, \) but the inequality constraint is inactive at this point. So all feasible points are regular.

IV. KKT conditions: \( \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} + \lambda_1 \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} = 0, \quad \lambda_1 (x_1^2 + x_2^2 - 1) = 0. \)

By the complementary slackness, we can branch on \( \lambda_1 = 0. \) In other words, we consider two main cases:

A. \( \lambda_1 = 0 \Rightarrow \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} = 0 \Rightarrow x = 0 \)
   - \( Hf(0) \) is indefinite \( \Rightarrow \) interior saddlepoint \( \Rightarrow \) neither max nor min

B. \( \lambda_1 \neq 0 \Rightarrow x_1^2 + x_2^2 - 1 = 0 \), \( \begin{bmatrix} 2\lambda_1 \\ 1 \\ 2\lambda_1 \end{bmatrix} x = 0 \Rightarrow \begin{bmatrix} 0 \\ 1 \\ 2\lambda_1 \end{bmatrix} \)
   1. \( 1 - 4\lambda_1^2 \neq 0 \Rightarrow x = 0 \) (handled above)
   2. \( 1 - 4\lambda_1^2 = 0 \Rightarrow \lambda_1 = \pm 1/2, \quad x_1 = -2\lambda_1 x_2 \Rightarrow 4\lambda_1^2 x_2^2 + x_2^2 - 1 = 0 \Rightarrow x_2^2 = \frac{1}{1 + 4\lambda_1^2} \)
      a. \( \lambda_1 = 1/2 \Rightarrow x = \pm \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}, \quad f(x) = -1/2 \)
      b. \( \lambda_1 = -1/2 \Rightarrow x = \pm \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}, \quad f(x) = 1/2 \)

V. Conclusions: Global maxima at \( x = \pm \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} ; \) global minima at \( x = \pm \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \). \( \square \)
Example 7.4.2. \( \min / \max \{ x_1 \mid x_1^3 - x_2^2 = 3, \ x_1^2 + x_2^2 \leq 3 \} \)

Solution.

I. Problem considerations: \( f(x) = x_1, \ m = 1, \ p = 2, \)
\( g_1(x) = x_1^3 - x_2^2 - 3, \ g_2(x) = x_1^2 + x_2^2 - 3. \)

- problem structure: max/min the first coordinate on an algebraic curve inside a circle
- derivatives: \( \nabla f(x) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \ \nabla g_1(x) = \begin{bmatrix} 3x_1^2 \\ -2x_2 \end{bmatrix}, \ \nabla g_2(x) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} \)

II. Existence: By the second constraint, the feasible region lies in the ball of radius \( \sqrt{3}. \)
Continuity of the constraint functions implies the feasible region is closed. The Weierstrass theorem guarantees that global maxima and minima exist.

III. Irregular points: Linear dependence of the constraint gradients means
\[ \begin{bmatrix} 3x_1^2 & 2x_1 \\ -2x_2 & 2x_2 \end{bmatrix} \] singular \( \Rightarrow 6x_1^2x_2 + 4x_1x_2 = 0 \Rightarrow x_1 = 0 \) or \( x_2 = 0 \) or \( x_1 = -2/3. \)

Combining these possibilities with the first constraint gives us

A. \( x_1 = 0 \Rightarrow -x_2^2 = 3: \) contradiction.
B. \( x_2 = 0 \Rightarrow x_1 = \sqrt{3} \Rightarrow \) second constraint is inactive \( \Rightarrow \) regular.
C. \( x_1 = -2/3 \Rightarrow -x_2^2 = 3 + (2/3)^3: \) contradiction.

Hence all feasible points are regular.

IV. KKT conditions:
\[ \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \lambda_1 \begin{bmatrix} 3x_1^2 \\ -2x_2 \end{bmatrix} + \lambda_2 \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} = 0, \quad \lambda_2(x_1^2 + x_2^2 - 3) = 0 \]

The second gradient equation tells us that either \( x_2 = 0 \) or \( \lambda_1 = \lambda_2: \)

A. \( x_2 = 0 \Rightarrow x_1 = \sqrt{3} \approx 1.44224957. \)
B. \( x_2 \neq 0 \Rightarrow \lambda_1 = \lambda_2 \Rightarrow 1 + \lambda_1(3x_1^2 + 2x_1) = 0. \) Thus \( \lambda_1 
eq 0, \) so complementary slackness gives \( x_1^3 + x_2^2 - 3 = 0. \) We need to solve the system comprising \( x_1^3 - x_2^2 = 3 \)
and \( x_1^2 + x_2^2 = 3. \) Substituting \( x_2^2 \) from one of these in the other yields \( x_1^3 + x_1^2 - 6 = 0. \)
Calculus (Intermediate Value Theorem and Rolle’s Theorem) shows this has exactly one root. Newton’s method computes it as \( x_1 \approx 1.53766, \) with \( x_2 \approx \pm 0.79725. \)

V. Conclusions: global min at \( \begin{bmatrix} \sqrt{3} \\ 0 \end{bmatrix}, \) global max at \( \begin{bmatrix} 1.53766 \\ \pm 0.79625 \end{bmatrix}. \)
Example 7.4.3. \( \min / \max \{ x_1^2 - x_1x_2 + x_2^2 + 3x_1 - 2x_2 \mid x_1 \geq 0, x_2 \geq 0, x_1 + x_2 \leq 1 \} \)

Solution.

I. Problem considerations: \( f(x) = x_1^2 - x_1x_2 + x_2^2 + 3x_1 - 2x_2, \)
\( g_1(x) = -x_1, \quad g_2(x) = -x_2, \quad g_3(x) = x_1 + x_2 - 1, \)
\( m = 0, \quad p = 3. \)

- convex quadratic objective on triangle with vertices \((0,0), (1,0), (0,1)\)
- \( \nabla f(x) = \begin{bmatrix} 2x_1 - x_2 + 3 \\ -x_1 + 2x_2 - 2 \end{bmatrix}, \nabla g_1(x) = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \nabla g_2(x) = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \nabla g_3(x) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \)

II. Existence: Constraints imply that \( 0 \leq x_1, x_2 \leq 1, \) i.e. bounded feasible region. Closedness is evident. Weierstrass guarantees existence of global max/min.

III. Irregular points: not applicable—all constraints are linear.

IV. KKT conditions:
\[
\begin{bmatrix} 2x_1 - x_2 + 3 \\ -x_1 + 2x_2 - 2 \end{bmatrix} + \lambda_1 \begin{bmatrix} -1 \\ 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 0 \\ -1 \end{bmatrix} + \lambda_3 \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 0,
\lambda_1(-x_1) = 0, \quad \lambda_2(-x_2) = 0, \quad \lambda_3(x_1 + x_2 - 1) = 0.
\]

Eight cases depending on which of \( \lambda_i \) are zero, with each case yielding a linear system:

<table>
<thead>
<tr>
<th>( \lambda_1 )</th>
<th>( \lambda_2 )</th>
<th>( \lambda_3 )</th>
<th>( \lambda_1 )</th>
<th>( \lambda_2 )</th>
<th>( \lambda_3 )</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( f(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( z )</td>
<td>( z )</td>
<td>( z )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-4/3</td>
<td>1/3</td>
<td>infeasible</td>
</tr>
<tr>
<td>( z )</td>
<td>( z )</td>
<td>( n )</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1/3</td>
<td>4/3</td>
<td>infeasible</td>
</tr>
<tr>
<td>( z )</td>
<td>( n )</td>
<td>( z )</td>
<td>0</td>
<td>-1/2</td>
<td>0</td>
<td>-3/2</td>
<td>0</td>
<td>infeasible</td>
</tr>
<tr>
<td>( z )</td>
<td>( n )</td>
<td>( n )</td>
<td>0</td>
<td>-8</td>
<td>-5</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>( n )</td>
<td>( z )</td>
<td>( z )</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>( n )</td>
<td>( z )</td>
<td>( n )</td>
<td>3</td>
<td>-2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>mixed signs</td>
</tr>
<tr>
<td>( n )</td>
<td>( n )</td>
<td>( z )</td>
<td>3</td>
<td>-2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>mixed signs</td>
</tr>
<tr>
<td>( n )</td>
<td>( n )</td>
<td>( n )</td>
<td>3</td>
<td>-2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>mixed signs</td>
</tr>
</tbody>
</table>

Here we have written \( z \) for “zero” and \( n \) for “nonzero.”

V. Conclusions: global max at \( \begin{bmatrix} 1 \\ 0 \end{bmatrix} \), global min at \( \begin{bmatrix} 0 \\ 1 \end{bmatrix} \).
7.5 Sufficient Conditions

Recall our necessary and sufficient conditions for optimality in the unconstrained problem \( \min_x f(x) \).

Necessary: If \( \bar{x} \) is a local minimizer for \( f \), then \( \bar{x} \) is a critical point.

Sufficient:

[Global] If \( \bar{x} \) is a critical point and \( f \) is convex, then \( \bar{x} \) is a global minimizer.

[Local] If \( \bar{x} \) is a critical point and \( Hf \) is continuous and positive definite at \( \bar{x} \), then \( \bar{x} \) is a strict local minimizer.

In this section, we provide analogous conditions for the general nonlinear program

\[
\begin{align*}
\text{(NLP)} & \quad \text{minimize } f(x) \text{ subject to } g_i(x) = 0, \text{ for } i = 1, \ldots, m, \\
& \quad g_i(x) \leq 0, \text{ for } i = m + 1, \ldots, p.
\end{align*}
\]

**Definition 7.5.1** (Lagrangian function). The Lagrangian for (NLP) is the function

\[
L(x, \lambda) = f(x) + \sum_{i=1}^{p} \lambda_i g_i(x).
\]

**Definition 7.5.2** (KKT points). We say that \((x, \lambda)\) is a KKT point for (NLP) if the following all hold:

1. \( g_i(x) = 0 \) for \( 1 \leq i \leq m \), and \( g_i(x) \leq 0 \) for \( m < i \leq p \);
2. \( \lambda_i \geq 0 \) for all \( i > m \);
3. \( L(x, \lambda) = f(x) \);
4. \( \nabla_x L(x, \lambda) = 0 \).

The Karush-Kuhn-Tucker Theorem (7.3.2) gives us the following.

Necessary: If \( \bar{x} \) is a local minimizer for (NLP) and \( \bar{x} \) is regular (or the active constraints are linear), then \((\bar{x}, \bar{\lambda})\) is a KKT point for some \( \bar{\lambda} \).

Here are the corresponding sufficient conditions.

Sufficiency:

[Global] If \((\bar{x}, \bar{\lambda})\) is a KKT point and \( L(x, \bar{\lambda}) \) is convex in \( x \), then \( \bar{x} \) is a global minimizer for (NLP).

[Local] If \((\bar{x}, \bar{\lambda})\) is a KKT point and \( H_x L(x, \bar{\lambda}) \) continuous and positive definite at \( \bar{x} \), then \( \bar{x} \) is a strict local minimizer for (NLP).

To see why these are valid, note that the stated sufficient conditions are precisely the global and local sufficiency conditions for unconstrained minimization of \( L(x, \lambda) \) over \( x \). Consequently, the claimed sufficiency for (NLP) follows from this lemma.
Lemma 7.5.3 (Lagrangian sufficiency). Suppose that $(\bar{x}, \bar{\lambda})$ satisfies conditions (1)–(3) of the KKT conditions in Definition 7.6.2.

(a) If $\bar{x}$ minimizes for $L(x, \bar{\lambda})$ over $x \in D$, then $\bar{x}$ is a minimizer for (NLP) over $x \in D$.
(b) If $\bar{x}$ strictly minimizes $L(x, \bar{\lambda})$ over $x \in D$, then $\bar{x}$ is a strict minimizer for (NLP) over $x \in D$.

Proof. Suppose $x \in D$ is feasible for (NLP). By (1), (2), and (3) we can prove (b) as

$$f(x) \geq f(x) + \sum_{i=1}^{p} \bar{\lambda}_i g_i(x) = L(x, \bar{\lambda}) > L(\bar{x}, \bar{\lambda}) = f(\bar{x}).$$

Example 7.5.4. \(\min/\max \{x_1^2 - 2x_1 + x_2 \mid x_1 + x_2 \leq 0, x_1 + x_2^2 \leq 0\}\)

Solution.

I. Problem considerations: $f(x) = x_1^2 - 2x_1 + x_2$, $g_1(x) = x_1 + x_2$, $g_2(x) = x_1 + x_2^2$.

**problem structure:**
- objective function is convex; the level curves (in decision space) for $f$ are parallel downward-opening parabolas
- feasible region is the convex unbounded intersection of a halfplane and the area bounded by a left-opening parabola

Therefore, we expect the existence of a global minimizer in the 3rd quadrant, and no other local optima.

**derivatives:**

$$\nabla f(x) = \begin{bmatrix} 2x_1 - 2 \\ 1 \end{bmatrix}, \quad \nabla g_1(x) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \nabla g_2(x) = \begin{bmatrix} 1 \\ 2x_2 \end{bmatrix}$$

II. Existence: To show there is no global maximum, take $x(t) = (-t^2, t)$ as $t \to \infty$. Existence of a global minimizer is more subtle: the origin is feasible with $f(0) = 0$, so it suffices to show that the intersection of $\{x \mid f(x) \leq 0\}$ with the constraint region is bounded. Add the inequalities $f(x) \leq 0$, $g_1(x) \leq 0$ and $g_2(x) \leq 0$ to get $x_1^2 + x_2^2 + 2x_2 \leq 0$, which can be rewritten as $x_1^2 + (x_2 + 1)^2 \leq 1$. Consequently, any points in this set are within distance 1 of the point $(0, -1)$, so the set is bounded. Now apply Weierstrass.

[Under convexity & regularity (see step III) global minimizers are precisely the KKT candidates for local minimizers, so existence is also established via step IV.]

III. Irregular points: Each of the constraint gradients is nonzero everywhere, so all points with only one active constraint are regular. The constraints are both active only at $(−1, 1)$ and $(0, 0)$, and their gradients are linearly independent at these points.

IV. KKT conditions:

$$\begin{bmatrix} 2x_1 - 2 \\ 1 \end{bmatrix} + \lambda_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \lambda_2 \begin{bmatrix} 1 \\ 2x_2 \end{bmatrix} = 0,$$

$$\lambda_1(x_1 + x_2) = 0, \quad \lambda_2(x_1 + x_2^2) = 0.$$
A. $\lambda_2 = 0 \Rightarrow \lambda_1 = -1$ by the second gradient equation. The first gradient equation gives $x_1 = 3/2$, which is infeasible.

B. $\lambda_2 \neq 0 \Rightarrow x_1 + x_2^2 = 0$ by complementarity.

1. If $\lambda_1 = 0$ then we have the system

   $x_1 + x_2^2 = 0$, $2x_1 + \lambda_2 = 2$, $1 + 2\lambda_2x_2 = 0$.

   By eliminating first $x_1$ and then $\lambda_2$ we obtain

   $x_1 = -x_2^2$, $\lambda_2 = 2(1 + x_2^2)$, $1 + 4x_2 + 4x_2^3 = 0$.

   The last equation has a unique root at $x_2 \approx -0.2367329$, which then gives $x_1 = -0.0560425$, $\lambda_2 = 2.112085$. (Candidate for a local minimizer.)

2. If $\lambda_1 \neq 0$ then $x_1 + x_2 = 0$, by complementarity. The only points satisfying this are $x = 0$ and $x = (-1, 1)$. In the former case we have $\lambda_1 = -1$ and $\lambda_2 = 3$; in the latter $\lambda_1 = 9$ and $\lambda_2 = -5$. In either case these have mixed signs, so are neither local max nor local min.

V. Conclusions: global min at $[-0.0560425, -0.2367329]$; no global max.
8. LINEAR PROGRAMMING

Linear programming is the most widely used class of constrained optimization problems. Its special structure provides for strong theoretical results and efficient algorithms.

8.1 Linear Programming Formulations

We begin this chapter with some standard formulations of linear programming problems.

**Definition 8.1.1** (Linear function; linear constraints; linear programs).

(a) A *linear function* \( f : \mathbb{R}^n \to \mathbb{R} \) has the form \( f(x) = c \cdot x = c_1 x_1 + \cdots + c_n x_n \).

(b) A *linear equation* is an equation expressible as \( g(x) = b \) for some linear function \( g \).

(c) A *weak linear inequality* is an inequality expressible as \( g(x) \leq b \) (or, equivalently, as \( g(x) \geq b \)) for some linear function \( g \).

(d) A *linear program* (LP) is an optimization problem in which a linear function is minimized (or maximized) subject to a finite number of constraints written as linear equations and weak linear inequalities.

**Example 8.1.2** (Some small LPs).

(a) \[
\begin{align*}
\text{max} \quad & 5x_1 + 4x_2 - x_3 \\
\text{subject to} \quad & x_1 - 2x_2 + 3x_3 \leq 5 \\
& 4x_1 + x_2 - x_3 \leq 7 \\
& 3x_1 + x_2 \leq 8 \\
& x_1 \geq 0, \ x_2 \geq 0, \ x_3 \geq 0
\end{align*}
\]

(b) \[
\begin{align*}
\text{min} \quad & x_3 \\
\text{subject to} \quad & x_1 + x_2 - x_3 \leq 2 \\
& 2x_1 - x_2 + x_3 = 1 \\
& 0 \leq x_1 \leq 1 \\
& 0 \leq x_2 \\
& -2 \leq x_3 \leq 2
\end{align*}
\]

**Definition 8.1.3** (Standard-form LPs). Here are three *standard forms* for linear programs. The right column shows them in matrix-vector notation.

(a) \[
\begin{align*}
\text{max} \quad & \sum_{j=1}^{n} c_j x_j \\
\text{subject to} \quad & \sum_{j=1}^{n} a_{ij} x_j \leq b_i \ (i = 1, \ldots, m) \\
& x_j \geq 0 \ (j = 1, \ldots, n)
\end{align*}
\]

(b) \[
\begin{align*}
\text{min} \quad & \sum_{j=1}^{n} c_j x_j \\
\text{subject to} \quad & \sum_{j=1}^{n} a_{ij} x_j \geq b_i \ (i = 1, \ldots, m) \\
& x_j \geq 0 \ (j = 1, \ldots, n)
\end{align*}
\]

(c) \[
\begin{align*}
\text{min} \quad & \sum_{j=1}^{n} c_j x_j \\
\text{subject to} \quad & \sum_{j=1}^{n} a_{ij} x_j = b_i \ (i = 1, \ldots, m) \\
& x_j \geq 0 \ (j = 1, \ldots, n)
\end{align*}
\]

\[87]
Observation 8.1.4 (Converting LPs into standard forms). Every linear program can be reformulated into any of the standard forms above by one or more of the following devices:

- (negate orientation of optimization) \( \max c \cdot x \leftrightarrow \min -c \cdot x \);
- (split equalities into inequalities) \( Ax = b \leftrightarrow Ax \leq b, -Ax \leq -b \);
- (introduce slack variables) \( Ax \leq b \leftrightarrow Ax + s = b, s \geq 0 \);
- (split variables into positive and negative parts) substitute \( x = x^+ - x^- \) and require \( x^+ \geq 0, x^- \geq 0 \).

Example 8.1.5 (Converting an LP into a standard form). To reformulate the LP

\[
\begin{align*}
\text{min} & \quad x_3 \\
\text{subject to} & \quad x_1 + x_2 - x_3 \leq 2 \\
& \quad 2x_1 - x_2 + x_3 = 1 \\
& \quad 0 \leq x_1 \leq 1 \\
& \quad 0 \leq x_2 \\
& \quad -2 \leq x_3 \leq 2
\end{align*}
\]

in the standard form of Definition 8.1.3(c), we need to:

- split the variable \( x_3 \) into negative and positive parts;
- introduce nonnegative slack variables \( x_4, x_5, x_6, x_7 \) for some of the inequalities.

The result is as follows:

\[
\begin{align*}
\text{min} & \quad x_3^+ - x_3^- \\
\text{subject to} & \quad x_1 + x_2 - x_3^+ + x_3^- + x_4 = 2 \\
& \quad 2x_1 - x_2 + x_3^+ - x_3^- = 1 \\
& \quad x_1 + x_5 = 1 \\
& \quad x_3^+ - x_3^- + x_6 = 2 \\
& \quad x_3^- - x_3^- - x_7 = -2 \\
& \quad 0 \leq x_1, 0 \leq x_2, 0 \leq x_3^+, 0 \leq x_3^-, 0 \leq x_4, 0 \leq x_5, 0 \leq x_6, 0 \leq x_7
\end{align*}
\]

A linear program can have one solution, infinitely many solutions, or no solutions. The following example shows that the possibility of “no solutions” can occur in two ways.

Example 8.1.6 (LPs without solutions).

(a) A linear program program can be infeasible:

\[
\begin{align*}
\text{min} & \quad x_1 - x_2 \\
\text{subject to} & \quad 2x_1 + x_2 \leq 3 \\
& \quad -4x_1 - 2x_2 \leq -8
\end{align*}
\]

(b) A linear program program can be unbounded:

\[
\begin{align*}
\text{min} & \quad x_1 \\
\text{subject to} & \quad x_1 \leq 1
\end{align*}
\]

88
8.2 Optimality Conditions — The Dual of a Linear Program

The motivation of this section is to get bounds on the optimal value in a given linear programming problem. For example, if we have a feasible solution in hand, we’d like to know if it’s worth the additional effort to solve the problem completely. Moreover, such bounds pave an alternative route to optimality conditions and provide leverage for solving complicated problems.

Example 8.2.1 (Bounds and optimality). Consider this system of constraints:

\[
\begin{align*}
3x_1 + 2x_2 − x_3 &\leq 5 \\
−x_1 − x_2 + 2x_3 &\leq −2 \\
x_1 &\geq 0, x_2 \geq 0, x_3 \geq 0
\end{align*}
\]

(a) Consider the objective: \( \max 3x_1 + 2x_2 − x_3 \). Then \( x = (1, 1, 0) \) is optimal.

Why? It’s feasible and attains the bound imposed on the objective by the first constraint.

(b) Consider the objective: \( \max 2x_1 + x_2 + x_3 \). Then \( x = (1, 1, 0) \) is optimal.

Why? It’s feasible and attains the bound imposed on the objective by adding the first two constraints.

(c) Consider the objective: \( \max x_1 + 2x_3 \). Then \( x = (1, 1, 0) \) is optimal.

Why? It’s feasible and attains the bound on the objective given by \( x_1 + 2x_3 \leq x_1 + 0x_2 + 3x_3 \leq 1 \), which can be deduced by adding the first constraint to twice the second constraint and noting that \( x_3 \geq 0 \).

Observation 8.2.2 (Valid inequalities). In the above example, any nonnegative linear combination of the constraints yields a valid inequality satisfied by all feasible points. If the coefficients in a valid inequality are higher than (or equal to) the corresponding objective coefficients, then we get an upper bound on the optimal value for the problem because each \( x_i \geq 0 \).

Definition 8.2.3 (Dual of an LP). The dual of a linear program \( (P) \) is an optimization problem in which an optimal bound is sought for \( (P) \) over all linear combinations of constraints in \( (P) \) that give valid inequalities bounding the objective of \( (P) \).

Example 8.2.4 (Dual of a small LP). Consider the linear program

\[
\begin{align*}
\text{max} & \quad x_1 − x_2 + 2x_3 \\
\text{subject to} & \quad 3x_1 + 2x_2 − x_3 \leq 5 \\
& \quad −x_1 − x_2 + 2x_3 \leq −2 \\
& \quad x_1 \geq 0, x_2 \geq 0, x_3 \geq 0.
\end{align*}
\]

We construct the dual problem in three steps.

1. Multiply the first constraint by \( y_1 \) and the second by \( y_2 \) with

\[
\begin{bmatrix}
y_1 \geq 0, y_2 \geq 0
\end{bmatrix}
\]
to get a valid inequality:

\[
\begin{align*}
&y_1 (3x_1 + 2x_2 - x_3) \leq (5)y_1 \\
&+ y_2 (-x_1 - x_2 + 2x_3) \leq (-2)y_2 \\
&\frac{3y_1 - y_2}{(3y_1 - y_2)x_1} + \frac{2y_1 - y_2}{(2y_1 - y_2)x_2} + (-y_1 + 2y_2)x_3 \leq 5y_1 - 2y_2.
\end{align*}
\]

2. To obtain an upper bound on the given LP, we need these coefficients of \(x_i \geq 0\) to be at least as large as the objective coefficients:

\[
\begin{align*}
3y_1 - y_2 &\geq 1 \\
2y_1 - y_2 &\geq -1 \\
-y_1 + 2y_2 &\geq 2.
\end{align*}
\]

3. To get the best possible bound, make the right-hand side in step 1 as low as possible:

\[
\min 5y_1 - 2y_2.
\]

Combining the boxed items above gives us the dual problem:

\[
\begin{align*}
\text{min} & \quad 5y_1 - 2y_2 \\
\text{subject to} & \quad 3y_1 - y_2 \geq 1 \\
& \quad 2y_1 - y_2 \geq -1 \\
& \quad -y_1 + 2y_2 \geq 2 \\
& \quad y_1 \geq 0, y_2 \geq 0.
\end{align*}
\]

Notice that the dual is just another linear program.

**Example 8.2.5** (Dual of a standard-form LP). Repeating the preceding example in matrix terms, we find that the dual of

\[
\begin{align*}
\text{max} & \quad c \cdot x \\
\text{s.t.} & \quad Ax \leq b \\
& \quad x \geq 0
\end{align*}
\]

is

\[
\begin{align*}
\text{min} & \quad b \cdot y \\
\text{s.t.} & \quad A^T y \geq c \\
& \quad y \geq 0.
\end{align*}
\]

**Example 8.2.6** (Dual of the dual). Now let’s find the dual of the dual in the preceding example. We start by introducing a vector

\[
z \geq 0
\]
to provide a valid inequality $z \cdot A^T y \geq z \cdot c$, or equivalently, $(Az) \cdot y \geq (c \cdot z)$. We need the coefficients $b$ to be at least as large as $Az$, so we require

$$Az \leq b.$$  

We also want to make the lower bound $c \cdot z$ as high as possible:

$$\max c \cdot z.$$  

Putting the boxes together gives us the LP

$$\max c \cdot z$$

s.t. $Az \leq b$

which is exactly the problem we started with in the preceding example. 

**Observation 8.2.7** (Reflexive duality). Every LP is the dual of its own dual. 

Some general guidelines for forming a dual linear program are given in Table 8.1.

<table>
<thead>
<tr>
<th>Correspondences for Dual LPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>sign-restricted variable $\leftrightarrow$ inequality constraint</td>
</tr>
<tr>
<td>free variable $\leftrightarrow$ equality constraint</td>
</tr>
<tr>
<td>objective $\leftrightarrow$ right-hand side</td>
</tr>
<tr>
<td>matrix $\leftrightarrow$ transpose matrix</td>
</tr>
<tr>
<td>maximize $\leftrightarrow$ minimize</td>
</tr>
</tbody>
</table>

Note: the orientation of a dual inequality depends on the orientation of the bound needed for the objective value in the original LP.

### Table 8.1

**Example 8.2.8** (Some more duals).

(a) Find the dual of

$$\begin{align*}
\text{max} & \quad 2x_1 - 3x_2 + 4x_3 \\
\text{subject to} & \quad 2x_1 + x_2 + 2x_3 \leq 5 \\
& \quad x_1 - x_2 + 3x_3 \geq 4 \\
& \quad -x_1 - 2x_2 + 4x_3 = 1 \\
& \quad x_1 \geq 0, \ x_2 \leq 0, \ x_3 \text{ free}
\end{align*}$$

Solution: The dual is

$$\begin{align*}
\text{min} & \quad 5y_1 + 4y_2 + y_3 \\
\text{subject to} & \quad 2y_1 + y_2 - y_3 \geq 2 \\
& \quad y_1 - y_2 - 2y_3 \leq -3 \\
& \quad 2y_1 + 3y_2 + 4y_3 = 4 \\
& \quad y_1 \geq 0, \ y_2 \leq 0, \ y_3 \text{ free}
\end{align*}$$
(b) Find the dual of
\[
\begin{align*}
\min \quad & x_1 + x_2 - 3x_3 \\
\text{subject to} \quad & -x_1 + x_2 + x_3 = 7 \\
& 2x_1 + 3x_2 - 2x_3 \geq 1 \\
& x_1 \geq -4 \\
& x_1 \leq 10, \\
& x_1 \text{ free, } x_2 \geq 0, x_3 \text{ free}
\end{align*}
\]

First we restate the problem so that the only restrictions on variables are for nonnegativity or nonpositivity:
\[
\begin{align*}
\min \quad & x_1 + x_2 - 3x_3 \\
\text{subject to} \quad & -x_1 + x_2 + x_3 = 7 \\
& 2x_1 + 3x_2 - 2x_3 \geq 1 \\
& x_1 \geq -4 \\
& x_1 \leq 10, \\
& x_1 \text{ free, } x_2 \geq 0, x_3 \text{ free}
\end{align*}
\]

Solution: The dual is
\[
\begin{align*}
\max \quad & 7y_1 + y_2 - 4y_3 + 10y_4 \\
\text{subject to} \quad & -y_1 + 2y_2 + y_3 + y_4 = 1 \\
& y_1 + 3y_2 \leq 1 \\
& y_1 - 2y_2 = -3 \\
& y_1 \text{ free, } y_2 \geq 0, y_3 \geq 0, y_4 \leq 0
\end{align*}
\]

(c) Find the dual of
\[
\begin{align*}
\max \quad & c \cdot x \\
\text{subject to} \quad & Ax \leq b \\
& Dx = f
\end{align*}
\]

Solution: The dual is
\[
\begin{align*}
\min \quad & b \cdot y + f \cdot z \\
\text{subject to} \quad & A^T y + D^T z = c \\
& y \geq 0, z \text{ free}
\end{align*}
\]

Now we lay out the main theoretical results concerning linear programming duality. For simplicity, these are stated in terms of a particular dual pair of linear programs. Still, the results are completely general because any linear program can be reformulated as one of these (as in §8.1).

**Lemma 8.2.9** (Weak duality). Consider the dual pair of linear programs
\[
(P) \quad \min \ c \cdot x \text{ s.t. } Ax \geq b; \quad (D) \quad \max \ b \cdot y \text{ s.t. } A^T y = c, \ y \geq 0.
\]

If \( x \) is feasible for \((P)\) and \( y \) is feasible for \((D)\), then \( c \cdot x \geq b \cdot y \).

**Proof.** Feasibility implies that \( c \cdot x = (A^T y) \cdot x = y \cdot (Ax) \geq b \cdot y \).
Theorem 8.2.10 (Duality-based necessary and sufficient conditions). Consider the dual pair of linear programs

\[(\mathcal{P}) \quad \min \, c \cdot x \text{ s.t. } Ax \geq b; \quad (\mathcal{D}) \quad \max \, b \cdot y \text{ s.t. } A^T y = c, \, y \geq 0.\]

Suppose \(\bar{x}\) is feasible for \(\mathcal{P}\). Then the following are equivalent:

(a) [optimality] \(\bar{x}\) is an optimal solution for \(\mathcal{P}\);
(b) [complementary slackness] there exists feasible \(\bar{y}\) for \(\mathcal{D}\) with \(\bar{y} \cdot (A\bar{x} - b) = 0\);
(c) [strong duality] there exists optimal \(\bar{y}\) for \(\mathcal{D}\) with \(b \cdot \bar{y} = c \cdot \bar{x}\).

Proof. The equivalence of (a) and (b) is just a special case of the linear-inequality constrained Karush-Kuhn-Tucker conditions in Chapter 7. The equivalence of (a) and (b) with (c) then follows from weak duality (Lemma 8.2.9).

The complementary slackness condition (b) of Theorem 8.2.10 provides an important test of optimality: each nonzero entry of \(A(\bar{x} - b)\) forces a corresponding zero entry in \(\bar{y}\). Along with the dual equality constraints, these provide a system of linear equations that \(\bar{y}\) must satisfy at optimality. By the Theorem, \(\bar{x}\) is optimal if and only if some solution \(\bar{y}\) of this linear system also satisfies \(\bar{y} \geq 0\).

Theorem 8.2.11 (Existence of LP solutions). Consider the dual pair of linear programs

\[(\mathcal{P}) \quad \min \, c \cdot x \text{ s.t. } Ax \geq b; \quad (\mathcal{D}) \quad \max \, b \cdot y \text{ s.t. } A^T y = c, \, y \geq 0.\]

If either

- (\(\mathcal{P}\)) is feasible and bounded (below, in objective value), or
- (\(\mathcal{D}\)) is feasible and bounded (above, in objective value),

then both problems admit optimal solutions and strong duality holds.

Proof. Because \(\mathcal{P}\) and \(\mathcal{D}\) are each completely general, it suffices to show that if \(\mathcal{P}\) is feasible with its objective value bounded below, then \(\mathcal{D}\) admits a feasible solution with objective value at least as high as the bound on \(\mathcal{P}\). So assume \(\gamma \leq \inf_x \{c \cdot x \mid Ax \geq b\}\) and that \(A\hat{x} \geq b\) for some \(\hat{x}\). Then \(c \cdot x \geq \gamma\) whenever \(Ax \geq b\).

Claim: \(\begin{bmatrix} -c \\ \gamma \end{bmatrix} \cdot \begin{bmatrix} \xi \\ t \end{bmatrix} \leq 0\) whenever \(\begin{bmatrix} -A & b \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \xi \\ t \end{bmatrix} \leq \begin{bmatrix} 0 \\ 0 \end{bmatrix}\).

Proof of claim: Consider \((\xi, t)\) for which \(-A\xi + bt \leq 0\) and \(-t \leq 0\), so that \(A\xi \geq bt\) and \(t \geq 0\). There are two possibilities, depending on whether \(t > 0\) or \(t = 0\). If \(t > 0\), then \(x = \xi/t\) satisfies \(Ax \geq b\). Hence \(c \cdot x \geq \gamma\), and so \(-c \cdot \xi + t \leq 0\). If \(t = 0\), then \(A\xi \geq 0\). Suppose, to the contrary, that \(-c \cdot \xi + \gamma t = -c \cdot \xi > 0\). Then \(x = \hat{x} + \tau \xi\) would satisfy \(Ax \geq b\) and \(c \cdot x < \gamma\) for any \(\tau > (\gamma - c \cdot \hat{x})/(c \cdot \xi)\), which would violate the definition of \(\gamma\).

By the Farkas Lemma (6.6.4), this claim yields the existence of \(y\) and \(s\) satisfying
\[
\begin{bmatrix} y \\ s \end{bmatrix} \geq \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{with} \quad \begin{bmatrix} -c \\ \gamma \end{bmatrix} = \begin{bmatrix} -A^T & 0 \\ b^T & -1 \end{bmatrix} \begin{bmatrix} y \\ s \end{bmatrix},
\]

93
so that $A^Ty = c$, $y \geq 0$, and $b \cdot y \geq \gamma$. The vector $y$ is therefore feasible for $(D)$ with objective value at least $\gamma$. \hfill \Box

Combined with weak duality, Theorem 8.2.11 immediately yields the following.

**Corollary 8.2.12** (Four possibilities in LP duality). Suppose that $(P)$ and $(D)$ are any two linear programs which are dual to each other. Then exactly one of the following holds:

(a) Both $(P)$ and $(D)$ have optimal solutions and they share a common optimal value.
(b) $(P)$ is unbounded and $(D)$ is infeasible.
(c) $(P)$ is infeasible and $(D)$ is unbounded.
(d) Both $(P)$ and $(D)$ are infeasible.

### 8.3 Basic Solutions of Linear Programs

In Chapter 7, we solved nonlinear programming problems by (essentially) considering all possible combinations of active and inactive constraints, and looking for those which satisfied the KKT conditions. Such massive enumeration can generally be avoided in linear programming problems. For example, consider this linear program:

\[
\begin{align*}
\text{maximize} & \quad 3x_1 + 2x_2 \\
\text{subject to} & \quad -x_1 + x_2 \leq 1 \\
& \quad 2x_1 - x_2 \leq 4 \\
& \quad 3x_1 + 4x_2 \leq 12 \\
& \quad x_1 + 2x_2 \leq 6 \\
& \quad x_1 \geq 0 \\
& \quad x_2 \geq 0.
\end{align*}
\]

The feasible region and objective contours for this problem are shown in Figures 8.1 and 8.2.

Figure 8.2 shows that the optimum occurs at a “corner” (or *vertex* or *extreme point*) of the feasible region. It turns out that this is always true: if a linear program has an optimal solution and its feasible region has any vertices at all, then at least one of the vertices is an optimal solution. In particular, we can solve the linear program by examining only the vertices and picking the one with the best objective value. However, doing so requires an algebraic representation of what it means to be a vertex. For example, if the constraint region is described by linear inequalities $Ax \leq b$, Figure 8.1 suggests that the vertices are those feasible points $x \in \mathbb{R}^n$ for which the active constraint gradients span all of $\mathbb{R}^n$. 94
Figure 8.1 LP feasible region

Figure 8.2 LP feasible region and objective contours
Modern methods for solving linear programs generally assume that the problem has first been reformulated so that all constraints are either equations or simple bounds on individual variables. To simplify the discussion of such methods, we shall assume that the feasible region has the form \( \{ x \in \mathbb{R}^n \mid Ax = b, \ x \geq 0 \} \). In this setting, the vertices admit a useful representation: they are the “basic” feasible points for \( A \).

**Definition 8.3.1 (Basic solutions).** Consider an \( m \times n \) matrix \( A \).

(a) We say that \( B \subseteq \{1, \ldots, n\} \) is a basis for \( A \) if the columns \( A_j \) for \( j \in J \) form a linearly independent set spanning the column space of \( A \).

(b) A solution \( x \in \mathbb{R}^n \) of \( Ax = b \) is said to be basic if the indices \( \{1, \ldots, n\} \) can be partitioned into sets \( B \) and \( N \) so that \( B \) is a basis for \( A \) and \( x_j = 0 \) for all \( j \in N \). In this case, the variables \( x_j \) with \( j \in B \) are called basic and those with \( j \in N \) are nonbasic.

**Example 8.3.2 (Basic solutions).** Consider the following system in two variables:

\[
\begin{align*}
x_1 + 2x_2 & \leq 3 \\
x_1 + x_2 & \leq 2 \\
x_1, x_2 & \geq 0
\end{align*}
\]

Adding slack variables gives us this equivalent system in four variables:

\[
\begin{align*}
x_1 + 2x_2 + x_3 & = 3 \\
x_1 + x_2 + x_4 & = 2 \\
x_1, x_2, x_3, x_4 & \geq 0
\end{align*}
\]

The two-dimensional feasible region (in the original variables) is illustrated in Figure 8.3, which shows vertices (basic feasible solutions) at the points labeled A, B, D, and E along with infeasible basic solutions at the points C and F. The corresponding basic representations are indicated in Table 8.2.

**Observation 8.3.3 (Facts about basic solutions).** The example above illustrates the following general facts about basic solutions:

- Each matrix representation is, in effect, row-reduced with respect to the corresponding basic variables; in particular, the basic variables can therefore be calculated from the nonbasic variables.
- The intersection points (i.e., basic solutions) correspond to setting the nonbasic variables to zero; the basic variables are then given by the corresponding right-hand sides.
- Basic solutions are feasible (i.e., nonnegative) if and only if the right-hand sides are all nonnegative.
- Adjacent basic solutions differ by exactly one basic/nonbasic pair.
- To move from one solution to an adjacent one requires a single pivot: use the column of the variable entering the basis and the row of the variable leaving the basis.

† Formal definitions of vertex and extreme point are not given here, but both concepts are equivalent to that of a basic feasible solution for sets of the form \( \{ x \mid Ax = b, \ x \geq 0 \} \).
Table 8.2 Basic solutions in Figure 8.3

<table>
<thead>
<tr>
<th>Point</th>
<th>$x_1$, $x_2$</th>
<th>$x_3$, $x_4$</th>
<th>B</th>
<th>N</th>
<th>Matrix Representation</th>
</tr>
</thead>
</table>
| A     | 0.0, 0.0     | 3.0, 2.0     | {3, 4} | {1, 2} | \[
|       |              |              |       |       | \begin{bmatrix}
|       |              |              |       |       | 1.0 & 2.0 & 1.0 & 0.0 & 3.0 \\
|       |              |              |       |       | 1.0 & 1.0 & 0.0 & 1.0 & 2.0 \\
|       |              |              |       |       | \end{bmatrix} feas. |
| B     | 0.0, 1.5     | 0.0, 0.5     | {2, 4} | {1, 3} | \[
|       |              |              |       |       | \begin{bmatrix}
|       |              |              |       |       | 0.5 & 1.0 & 0.5 & 0.0 & 1.5 \\
|       |              |              |       |       | 0.5 & 0.0 & -0.5 & 1.0 & 0.5 \\
|       |              |              |       |       | \end{bmatrix} feas. |
| C     | 0.0, 2.0     | -1.0, 0.0    | {2, 3} | {1, 4} | \[
|       |              |              |       |       | \begin{bmatrix}
|       |              |              |       |       | 1.0 & 1.0 & 0.0 & 1.0 & 2.0 \\
|       |              |              |       |       | -1.0 & 0.0 & 1.0 & -2.0 & -1.0 \\
|       |              |              |       |       | \end{bmatrix} infeas. |
| D     | 1.0, 1.0     | 0.0, 0.0     | {1, 2} | {3, 4} | \[
|       |              |              |       |       | \begin{bmatrix}
|       |              |              |       |       | 0.0 & 1.0 & 1.0 & -1.0 & 1.0 \\
|       |              |              |       |       | 1.0 & 0.0 & -1.0 & 2.0 & 1.0 \\
|       |              |              |       |       | \end{bmatrix} feas. |
| E     | 2.0, 0.0     | 1.0, 0.0     | {1, 3} | {2, 4} | \[
|       |              |              |       |       | \begin{bmatrix}
|       |              |              |       |       | 0.0 & 1.0 & 1.0 & -1.0 & 1.0 \\
|       |              |              |       |       | 1.0 & 1.0 & 0.0 & 1.0 & 2.0 \\
|       |              |              |       |       | \end{bmatrix} feas. |
| F     | 3.0, 0.0     | 0.0, -1.0    | {1, 4} | {2, 3} | \[
|       |              |              |       |       | \begin{bmatrix}
|       |              |              |       |       | 0.0 & -1.0 & -1.0 & 1.0 & -1.0 \\
|       |              |              |       |       | 1.0 & 2.0 & 1.0 & 0.0 & 3.0 \\
|       |              |              |       |       | \end{bmatrix} infeas. |
Observation 8.3.4 (Uniqueness and degeneracy). A given basis may have several matrix representations, but these may differ only by the order of the rows. On the other hand, it is possible that a given point is representable by more than one basic solution: such a point is said to be degenerate. Degeneracy only occurs when some basic variable takes the value zero. An example of degeneracy is shown in Figure 8.1, where the basic infeasible solution $(0,3)$ has three different basic representations.

Figure 8.3 again suggests that any linear function will attain its optimum at a basic feasible solution (if at all). The next result verifies this.

Theorem 8.3.5 (Existence of a basic optimal solution). If the linear programming problem

$$\max \ c \cdot x \ \text{s.t.} \ Ax = b, \ x \geq 0$$

has an optimal solution, then it also has a basic optimal solution.

Proof. Let $\bar{x}$ be an optimal solution with the smallest possible number of positive entries and define $J = \{j \mid \bar{x}_j > 0\}$. We shall show that $\bar{x}$ is basic; more specifically, we shall show that the columns $A_j$ for $j \in J$ form a linearly independent (or empty) set. Suppose, to the contrary, that they are linearly dependent. Then there exist scalars $d_j$, not all zero, with $\sum_{j \in J} A_j d_j = 0$. Without loss of generality, we may assume that $d_j = 0$ for all $j \notin J$ and that $d_j > 0$ for at least one $j \in J$. Now define

$$\alpha = \max \left\{ -\frac{\bar{x}_j}{d_j} \mid j \in J \text{ with } d_j > 0 \right\}.$$  

If $d_j \geq 0$ for all $j \in J$, define $\beta = 1$; otherwise, define

$$\beta = \min \left\{ -\frac{\bar{x}_j}{d_j} \mid j \in J \text{ with } d_j < 0 \right\}.$$  

Then $\alpha < 0 < \beta$ and $\bar{x} + td$ is feasible for all $t \in [\alpha, \beta]$. Hence, the optimality of $\bar{x}$ implies that $c \cdot (\bar{x} + td) \geq c \cdot \bar{x}$ for all $t \in [\alpha, \beta]$. For $t = \alpha$, this shows that $c \cdot d \leq 0$, whereas $t = \beta$ yields $c \cdot d \geq 0$. Consequently, $c \cdot d = 0$ and we see that $\bar{x} + td$ is optimal for all $t \in [\alpha, \beta]$. In particular, $\bar{x} + ad$ is optimal and has fewer positive entries than $\bar{x}$. This contradicts the choice of $\bar{x}$, thereby completing the proof.

The basic-solution representation of a vertex is useful for several reasons. First, the $x_j$ values at each vertex are easily read off the augmented matrix. Second, when we move from a vertex to one of its neighbors, the representation is updated by simply “pivoting” in the sense of a row-reduction operation. Third, the objective value each vertex can be monitored as we go along. The next section demonstrates how these features may be employed.
8.4 The Simplex Method — A Motivating Example

The previous section showed that a linear program can be solved by focusing attention on the basic solutions (vertices). A wide variety of algorithms build on this principle. One of the most popular, the simplex method, is developed and illustrated in the following example. A systematic statement of the simplex method and concise examples of its use are provided in the next two sections.

Example 8.4.1 (Introduction to the simplex method). In this example we solve the linear program

$$\begin{align*}
\text{max} & \quad 3x_1 + 2x_2 \\
\text{subject to} & \quad x_1 + 2x_2 \leq 3 \\
& \quad x_1 + x_2 \leq 2 \\
& \quad x_1, x_2 \geq 0.
\end{align*}$$

The feasible region and basic solutions for this problem are the same as in Example 8.3.2, Figure 8.3, and Table 8.2. As before, we introduce slack variables to rewrite this problem with equality constraints as

$$\begin{align*}
\text{max} & \quad 3x_1 + 2x_2 \\
\text{subject to} & \quad x_1 + 2x_2 + x_3 = 3 \\
& \quad x_1 + x_2 + x_4 = 2 \\
& \quad x_1, x_2, x_3, x_4 \geq 0.
\end{align*}$$

We also introduce a variable $z$ to represent the objective value: $z = 3x_1 + 2x_2$. (Pictorially, $z$ tells us which objective contour we’re on.) These three equations constitute a linear system of the form $Ax = b$ and $c^T x - z = 0$:

$$\begin{bmatrix}
1 & 2 & 1 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 \\
3 & 2 & 0 & 0 & -1
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
x_3 \\
x_4 \\
z
\end{bmatrix} =
\begin{bmatrix}
3 \\
2 \\
0
\end{bmatrix}.$$

In analogy with the “augmented matrix” of linear algebra, we can summarize this linear system in a tableau as

$$\begin{array}{cc|c}
1 & 2 & 1 & 0 & 3 \\
1 & 1 & 0 & 1 & 2 \\
3 & 2 & 0 & 0 & 0
\end{array} \begin{array}{c}
(= -z).
\end{array}$$

This tableau represents the basic feasible solution $x_1 = x_2 = 0, x_3 = 3, x_4 = 2, z = 0$; in other words, this is the point A of Figure 8.3 and Table 8.2. Note that the tableau omits the column for $z$, because that column cannot change as we move from basis to basis. Moreover, in any tableau corresponding to a basic solution, the entry in the lower right corner always denotes the value of $-z$ for that basic solution.

Now we want to use the tableau representation of basic solutions to help us optimize the objective function. According to the above tableau, we have $\partial z / \partial x_2 = 2 > 0$. In
other words, \(z\) increases if we increase the nonbasic variable \(x_2\) upward from 0. Because we’re interested in maximizing \(z\), we pursue this possibility. For example, moving \(x_2\) from \(x_2 = 0\) to \(x_2 = 1\) gives us

\[
\begin{align*}
x_1 &= 0, \\
x_2 &= 1, \\
x_3 &= 3 - x_1 - 2x_2 = 1, \\
x_4 &= 2 - x_1 - x_2 = 1, \\
z &= 3x_1 + 2x_2 = 2,
\end{align*}
\]

versus \(z = 0\) when \(x_1 = x_2 = 0\). How far can we go in increasing \(x_2\)? Notice that \(x_3\) and \(x_4\) both dropped to 1 when \(x_2\) was increased. To maintain feasibility, we cannot let either \(x_3\) or \(x_4\) become negative. This means that we need to enforce the inequalities

\[
\begin{align*}
x_3 &= 3 - x_1 - 2x_2 \geq 0, \\
x_4 &= 2 - x_1 - x_2 \geq 0.
\end{align*}
\]

Because we’re holding the nonbasic variable \(x_1\) fixed at 0 for now, we can solve these inequalities for explicit restrictions on \(x_2\):

\[
\begin{align*}
3/2 &\geq x_2, \\
2 &\geq x_2.
\end{align*}
\]

Note that these bounds are given as the ratios of the entries in the right-hand side of the tableau to the entries in the \(x_2\)-column of the tableau. The first of these bounds is the most restrictive (see Figure 8.3), and it forces \(x_3 = 0\) and \(x_2 = 3/2\). This means that we need to swap the basis: \(x_3\) comes out and \(x_2\) goes in. To make this change, we simply perform a pivot operation on row 1, column 2 of the tableau

\[
\begin{array}{cccc|c}
1 & 2 & 1 & 0 & 3 \\
1 & 1 & 0 & 1 & 2 \\
3 & 2 & 0 & 0 & 0
\end{array}
\]

to obtain the tableau

\[
\begin{array}{cccc|c}
1/2 & 1 & 1/2 & 0 & 3/2 \\
1/2 & 0 & -1/2 & 1 & 1/2 \\
2 & 0 & -1/2 & 1 & -3
\end{array}
\].

This tableau corresponds to the point B in Figure 8.3: \(x_1 = x_3 = 0, x_2 = 3/2, x_4 = 1/2,\) and \(z = 3\).

In this new tableau, the only positive objective coefficient is in column 1. This means that only \(x_1\) can be increased if we wish to increase the objective value, so we want column 1 to enter the basis. The ratios of the right-hand entries to the entries in column 1 give the restrictions on how far \(x_1\) can be increased: \((3/2)/(1/2) = 3\) is the restriction indicated
by row 1, and \((1/2)/(1/2) = 1\) is the restriction indicated by row 2. The second of these is the tightest, so the basic variable in the second row (i.e., \(x_4\)) must leave the basis. To make this change, we pivot on row 2 and column 1 to get a new tableau:

\[
\begin{array}{cccc|c}
0 & 1 & 1 & -1 & 1 \\
1 & 0 & -1 & 2 & 1 \\
0 & 0 & 1 & -4 & -5 \\
\end{array}
\]

This tableau corresponds to the point D in Figure 8.3, with \(x_1 = 1 = x_2, x_3 = x_4 = 0\), and \(z = 5\).

We repeat the procedure of the previous two paragraphs: the only opportunity to increase the objective value is by increasing the nonbasic variable \(x_3\), because it has the only positive objective coefficient. So \(x_3\) will enter the basis. The only restriction on increasing \(x_3\) comes from row 1 (because of the positive entry in row 1, column 3), and the corresponding ratio with the right-hand side entry is \(1/1 = 1\). Hence the basic variable of row 1 (i.e., \(x_2\)) must leave the basis. Pivoting on row 1 and column 2 give us the tableau

\[
\begin{array}{cccc|c}
0 & 1 & 1 & -1 & 1 \\
1 & 1 & 0 & 1 & 2 \\
0 & -1 & 0 & -3 & -6 \\
\end{array}
\]

which corresponds to the point E in Figure 8.3, with \(x_1 = 2, x_3 = 1, x_2 = x_4 = 0\), and \(z = 6\).

Looking again for an opportunity to increase the objective, we find none: all of the objective coefficients are nonpositive. In other words, we have arrived at the optimal solution! To verify that \((x_1, x_2) = (2, 0)\) is indeed optimal for the original LP, one can check that it is feasible and that its objective value \((z = 6)\) is the same as objective value for the vector \((y_1, y_2) = (0, 3)\), which is feasible for the dual LP

\[
\begin{align*}
\text{min} & \quad 3y_1 + 2y_2 \\
\text{subject to} & \quad y_1 + y_2 \geq 3 \\
& \quad 2y_1 + y_2 \geq 2 \\
& \quad y_1, y_2 \geq 0.
\end{align*}
\]

Note that this dual solution appears in the final tableau as the negative of objective coefficients for the slack variables. □

The method illustrated above can be generalized to solve any linear program, as indicated in the next section.

### 8.5 The Simplex Method — Formal Algorithm and Theory

In this section we formalize the method illustrated in the previous section and describe its use for solving general linear programming problems. We begin with a precise statement of the core algorithm. Consider a standard-form LP

\[(\mathcal{P})\]

\[
\max \{ c \cdot x \mid Ax = b, \ x \geq 0 \}.
\]
Introducing a variable $z$ to represent the objective value, we can rewrite this as

$$\max z \text{ s.t. } \begin{bmatrix} A & 0 \\ c^T & -1 \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}, \quad x \geq 0.$$ 

The simplex method moves among basic solutions of $Ax = b$ by using pivoting operations to transform the linear system

$$Ax = b, \quad c^T x - z = 0$$

into a sequence of equivalent systems of the form

$$\bar{A}x = \bar{b}, \quad \bar{c}^T x - z = -\alpha.$$ 

As in the preceding section, these transformations can be represented in tableau omitting explicit reference to the objective variable $z$. The resulting method is as follows.

**Method 8.5.1** (Ordinary Simplex Method — tableau form). This method solves the standard-form LP $\max\{c \cdot x \mid Ax = b, \ x \geq 0\}$, beginning with a feasible basis for the matrix $A$ of full row rank.

1. Set up the initial tableau

$$\begin{array}{c|c}
A & b \\
c^T & 0 \\
\end{array}$$

Given a feasible starting basis $B$, row-reduce with respect to $B$ to obtain a basic tableau

$$\begin{array}{c|c}
\bar{A} & \bar{b} \\
\bar{c}^T & -\alpha \\
\end{array}$$

in which $c_j = 0$ for each $j \in B$, and the columns $A_j$ (for $j \in B$) are distinct coordinate vectors.

2. Choose a nonbasic column $j \in N$ whose objective coefficient $\bar{c}_j$ is positive. If there is none, then the current tableau proves optimality. [We call $j$ the *entering column.*]

3. If there are no positive entries in the entering column $j$, the current tableau proves unboundedness. Otherwise, calculate the ratios of the right-hand sides in $\bar{b}$ to the positive entries of the entering column $j$. Choose a row $i$ of $[\bar{A}, \bar{b}]$ that imposes the tightest restriction (lowest nonnegative ratio) on increasing $x_j$ and let $k \in B$ be the index of the variable $x_k$ currently determined by row $i$. [We call $k$ the *leaving column.*]

4. Pivot (in the row-reduction sense) on the entry in row $i$ and column $j$: this puts $j$ into $B$ and removes $k$ from $B$. Go to step 2.

**Observation 8.5.2** (Removing the full-rank assumption). Note that if $A$ does not have full row rank, then the equation $Ax = b$ can be replaced by an equivalent linear system that does have full row rank. In this way, the simplex method can be applied to any linear program following the necessary transformation to the standard form indicated in Method 8.5.1.
Example 8.4.1 in the previous section illustrates the simplex method with enough detail to see why it is well-defined for general linear programs; a more formal discussion is given in Appendix A.4. Eventual termination of the method can be addressed as follows.

**Observation 8.5.3** (Finite termination of the simplex method). Clearly, the objective value must increase whenever the nonbasic variable associated with the entering column takes on a positive value by becoming basic. Because there are only finitely many bases $B$, this means that the simplex method must eventually terminate, unless it “cycles” among a list of bases for which the variables of entering columns remain at zero (this is degeneracy — see Observation 8.3.4). However, we can avoid this behavior by applying Bland’s anti-cycling rule: always choose the smallest possible index $j$ in step 2 and the smallest possible index $k$ in step 3. This guarantees that no basis occurs twice, so that eventual progress is ensured. In practice, degeneracy occurs frequently, but cycling almost never does.

To apply the simplex method to general linear programs, a few more details need to be ironed out. In particular, we are confronted by three main issues: how to handle problems that are not in the standard form, how to find an initial basic solution, and how to verify optimality. The first of these can be handled directly by reformulating (as in §8.1) the given problem into the required standard form; the other two issues comprise the remainder of the current section. Examples are worked out in the next section.

In the ordinary simplex method, we assume that a basic feasible solution is known. The example in the previous section had the advantage that the given problem (before slack variables were incorporated) had a feasible origin, so that a feasible basis consisted of the slack variables themselves. In general, it’s not so easy to find an initial basis, so a procedure like the following is used.

**Method 8.5.4** (Phase 1 — finding a feasible basis). The following procedure finds an initial feasible point for a problem in the standard form

$$(P) \quad \max \{c \cdot x \mid Ax = b, \ x \geq 0 \}.$$

1. Create the “Phase 1” linear program:
   a. For each row $i$, introduce an auxiliary variable with column given by the coordinate vector $\pm e^{(i)}$, where the sign matches that of the right-hand side entry.
   b. Use an objective vector with a $-1$ for each auxiliary variable and zeroes elsewhere.

The first Phase 1 tableau takes the form

<table>
<thead>
<tr>
<th>$A$</th>
<th>$I$</th>
<th>$E$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>$-e^T$</td>
<td>0</td>
</tr>
</tbody>
</table>

2. Obtain a starting basis $B$ for the Phase 1 problem by pivoting on each auxiliary variable to zero out its objective coefficient.

3. Solve the Phase 1 problem:
   a. Run the ordinary simplex method on the Phase 1 problem.
   b. If the optimal value is positive, the original problem is infeasible.
c. If the optimal value is zero, pivot any remaining auxiliary variables out of the basis. The final basis $B$ will also be basic feasible for the original problem $(\mathcal{P})$. □

**Observation 8.5.5** (Validation of the Phase 1 procedure). Note that the Phase 1 problem is always feasible and its objective value is bounded above by zero, so it must admit an optimal solution. Because a basic feasible solution is at hand, the ordinary simplex method can always be applied and must find an optimum. □

Once we have a feasible basis $B$, we can optimize the true objective by applying the ordinary simplex method. Examples are given in the next section.

**Theorem 8.5.6** (Basic solutions to LPs). If an LP of the form $(\mathcal{P})$ or $(\hat{\mathcal{P}})$ admits any feasible solutions, it must admit a basic feasible solution. Moreover, if the LP admits an optimal solution, then some optimal solution must be basic.

**Proof.** The first statement follows directly from Observation 8.5.5, while the second follows from the validity (see Appendix A.4) and finite termination of the ordinary simplex method. □

We end this section with a discussion of how to verify optimality. The simplex method stops when the objective row in the tableau indicates that further improvement is impossible (unless, of course, the problem is unbounded). In this way, the objective row “proves” optimality to the person who is performing the simplex method. But how could we use the final tableau to prove optimality of our solution to someone else? It’s unreasonable to expect another person to learn about the simplex method and then look over our shoulder as we carry it out. On the other hand, we know by duality theory that a solution of the dual linear program proves optimality by providing a tight bound on the objective value of the original. Moreover, this bound can be easily demonstrated to anyone with a high-school knowledge of algebra. Fortunately, it turns out that the objective row in the final tableau provides all the information needed to calculate an optimal dual solution.

**Theorem 8.5.7** (Recovering dual variables from the final tableau). Consider the standard form problem

$$(\mathcal{P}) \quad \max \{ c \cdot x \mid Ax = b, \ x \geq 0 \}$$

and its dual

$$(\mathcal{D}) \quad \min \{ b \cdot y \mid A^T y \geq c \}.$$

Suppose that the initial and optimal tableaus (respectively) for $(\mathcal{P})$ are given by

$$\begin{array}{c|c|c}
A & b \\
\hline
\begin{array}{c}
\text{c}^T \\
0
\end{array} & \hat{A} & \hat{b}
\end{array}, \quad \begin{array}{c|c|c}
\begin{array}{c}
\text{c}^T
\end{array} & \begin{array}{c}
\text{c}
\end{array} & \begin{array}{c}
\text{c} - \hat{c}
\end{array}
\end{array}$$

Then any vector $y$ satisfying the equation $A^T y = c - \hat{c}$ is an optimal solution of $(\mathcal{D})$. In other words, the objective row in the final tableau contains the slack variables for the dual solution.
Proof. Assume that \( y \) satisfies \( A^T y = c - \tilde{c} \). We need to prove that \( y \) is dual feasible and yields the same objective value as the point \( x \) described by the tableau. First, note that the optimality of the final tableau implies that \( \tilde{c} \leq 0 \), so \( A^T y \leq A^T y + \tilde{c} = c \). Hence \( y \) is dual feasible.

Next, recall that the initial tableau corresponds the linear system \( Ax = b \) and \( c^T x - z = 0 \), whereas the final tableau corresponds to the equivalent system \( \tilde{A}x = \tilde{b} \) and \( \tilde{c}^T x - z = -\alpha \). Now consider any optimal solution \( x \) of \((P)\). We then have

\[
b \cdot y = (Ax) \cdot y = x \cdot (A^T y) = x \cdot (c - \tilde{c}) = z - (z - \alpha) = \alpha.
\]

Hence, \( y \) and \( x \) yield the same objective value, so \( y \) is indeed optimal for \((D)\).

8.6 The Simplex Method — More Examples

In this section, we solve several linear programs with the simplex method, following the general outline in Figure 8.4.

<table>
<thead>
<tr>
<th>Two-Phase Simplex Method for Solving a General LP</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Convert the problem to standard form.</td>
</tr>
<tr>
<td>II. Find a basic feasible solution (Phase 1, if needed).</td>
</tr>
<tr>
<td>III. Find an optimal solution (Phase 2).</td>
</tr>
<tr>
<td>IV. Find a dual solution (if desired).</td>
</tr>
</tbody>
</table>

Figure 8.4

We begin by summarizing the main concerns to be dealt with in Steps I and II of Figure 8.4.

**Observation 8.6.1** (Converting the problem to standard form). The ordinary simplex method is applied to a problem in the form \( \max \{ c \cdot x \mid Ax = b, x \geq 0 \} \), which has three main requirements.

*All variables must have nonnegativity restrictions*, so (for example) the following transformations could be applied:

- Incorporate simple nonzero bounds \( x_i \leq u_i \) or \( x_i \geq l_i \) in the constraint matrix.
- Replace free variables \( x_i \) by \( x_i^+ \geq 0 \) and \( x_i^- \geq 0 \) via \( x_i = x_i^+ - x_i^- \). (In the tableau, use the coefficients of \( x_i \) for \( x_i^+ \), but negate the coefficients of \( x_i \) for \( x_i^- \).)
- Replace nonpositive variables \( x_i \leq 0 \) by their negations \( \bar{x}_i = -x_i \geq 0 \). (In the tableau, switch signs on the coefficients of \( x_i \) to get the coefficients for \( \bar{x}_i \).)

*All constraint rows must be equality constraints*, so introduce nonnegative slack variables:

- Inequality \( a_i x \leq b_i \) becomes \( a_i x + s_i = b_i, s_i \geq 0 \).
- Inequality \( a_i x \geq b_i \) becomes \( a_i x - s_i = b_i, s_i \geq 0 \).
The objective must be a maximization, so negate the objective if it’s given as a minimization. (Alternatively, an objective can be minimized directly by choosing entering variables with negative objective coefficients.)

Observation 8.6.2 (Finding an initial basic feasible solution). There are two possibilities.

For a **feasible origin** (in the non-slack variables):
1. Start with a (partial) basis $B$ consisting of all slack variables.
2. Pivot on each member of $B$ to correct the sign (we need $A_B = I$).
3. Iteratively augment $B$ by row reducing until each row contains a basic variable; start Phase 2 with $B$.

For an **infeasible origin**:
1. Introduce nonnegative auxiliary variables to create the Phase 1 problem:
   - Use a separate auxiliary variable for each infeasible row.
   - For the objective row, use $-1$ for auxiliary variables and $0$ for the rest.
2. Obtain a starting basis $B$ for the Phase 1 problem:
   a. Start with $B$ given by the auxiliary variables and the feasible slack variables.
   b. Pivot on each member of $B$ to correct the sign and clear its objective coefficient.
   c. Iteratively augment $B$ by row-reduction.
3. Now solve the Phase 1 problem.

With the details summarized above, we can see exactly what is needed to handle the important special case of

\[(\hat{P}) \quad \max\{c \cdot x \mid Ax \leq b, x \geq 0\},\]

which is precisely the same form as the example solved earlier in §8.4. As in that example, we convert any problem of the form $(\hat{P})$ into the simplex method’s standard form by adding nonnegative slack variables for the inequality rows: each inequality $a_i x \leq b_i$ is replaced by $a_i x + s_i = b_i$, $s_i \geq 0$. The converted problem has the form

\[
\max\{c \cdot x \mid Ax + s = b, x \geq 0, s \geq 0\}.
\]

To apply the simplex method to this problem, we need to find an initial basic feasible solution. There are two possibilities, depending on whether the origin $x = 0$ is a feasible point (that is, whether $s = b \geq 0$). For a feasible origin, we can start with a basis $B$ consisting of all the slack variables. More generally, we can omit auxiliary variables (during Phase 1) for any inequality constraint that can be satisfied as an equation by a nonnegative value of its corresponding slack variable. After applying Phase 2 of the simplex method, the dual variables for $(\hat{P})$ can be read directly off the final tableau. Specifically, suppose the initial (Phase-2) simplex tableau for $(\hat{P})$ is

\[
\begin{array}{ccc|c}
A & I & b \\
\hline
\end{array}
\]

\[
\begin{array}{ccc|c}
c^T & 0 & 0 \\
\hline
\end{array}
\]
and the final (optimal) tableau is
\[
\begin{array}{cc|c}
\tilde{A} & M & \tilde{b} \\
\tilde{c}^T & -y^T & -\alpha \\
\end{array}
\]

The vector \( y \) in the objective row of this tableau solves the dual (see Theorem 8.5.7) of \((\tilde{P})\), which is given by
\[
(\tilde{D}) \quad \min \{ b \cdot y \mid A^T y \geq c, y \geq 0 \}.
\]

Here is an explicit example of the simplex method carried out for such a problem. 

**Example 8.6.3** (Solving a linear program with feasible origin). 

Apply the simplex method to solve
\[
\begin{align*}
\max & \quad 5x_1 + 4x_2 + 3x_3 \\
\text{subject to} & \quad 2x_1 + 3x_2 + x_3 \leq 5 \\
& \quad 4x_1 + x_2 + 2x_3 \leq 11 \\
& \quad 3x_1 + 4x_2 + 2x_3 \leq 8 \\
& \quad x_1, x_2, x_3 \geq 0.
\end{align*}
\]

Solution: By introducing slack variables we convert this to standard form with
\[
c = \begin{bmatrix} 5 \\ 4 \\ 3 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix}, \quad A = \begin{bmatrix} 2 & 3 & 1 & 1 & 0 & 0 \\ 4 & 1 & 2 & 0 & 1 & 0 \\ 3 & 4 & 2 & 0 & 0 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 5 \\ 11 \\ 8 \end{bmatrix}.
\]

Since the origin is feasible in the original problem, we may set the non-slack variables to zero to obtain the basic feasible solution
\[
x_1 = x_2 = x_3 = 0, \ x_4 = 5, \ x_5 = 11, \ x_6 = 8.
\]

We start by entering the data for the standard form problem in a tableau:
\[
\begin{array}{cccc|c}
2 & 3 & 1 & 1 & 0 & 0 & 5 \\
4 & 1 & 2 & 0 & 1 & 0 & 11 \\
3 & 4 & 2 & 0 & 0 & 1 & 8 \\
5 & 4 & 3 & 0 & 0 & 0 & 0
\end{array}
\]

Our starting feasible basis of slack variables is \( B = \{4, 5, 6\} \). To maximize, we must choose entering variables with positive objective coefficients.

- Any of the columns 1, 2 or 3 may enter: choose 1.
- Restrictions on column 1: 5/2 by row 1, 11/4 by row 2, 8/3 by row 3.
- Pivot on column 1, row 1. (Column 4 leaves the basis.)

\[
\begin{array}{cccccc|c}
1 & 3/2 & 1/2 & 1/2 & 0 & 0 & 5/2 \\
0 & -5 & 0 & -2 & 1 & 0 & 1 \\
0 & -1/2 & 1/2 & -3/2 & 0 & 1 & 1/2 \\
0 & -7/2 & 1/2 & -5/2 & 0 & 0 & -25/2 \\
\end{array}
\]

- Only column 3 may enter.
- Restrictions on column 3: \((5/2)/(1/2) = 5\) by row 1, \((1/2)/(1/2) = 1\) by row 3.
- Pivot on column 3, row 3. (Column 6 leaves the basis.)

\[
\begin{array}{cccccc|c}
1 & 2 & 0 & 2 & 0 & -1 & 2 \\
0 & -5 & 0 & -2 & 1 & 0 & 1 \\
0 & -1 & 1 & -3 & 0 & 2 & 1 \\
0 & -3 & 0 & -1 & 0 & -1 & -13 \\
\end{array}
\]

Since no objective coefficient is positive, we cannot move to a better point: this tableau represents the optimum! We read the solution from the tableau as

\[x_1 = 2, \ x_2 = 0, \ x_3 = 1,\]

with objective value 13. (The tableau also gives the slacks as \(x_4 = 0, x_5 = 1\) and \(x_6 = 0\).)

Finally, as mentioned just before this example, the objective row indicates dual variables for the matrix inequalities with values

\[y_1 = 1, \ y_2 = 0, \ y_3 = 1.\]

It is easily verified that this is a dual feasible point whose objective value is also 13. □

Next, we illustrate the two-phase simplex method on an example that is already in the required standard form.

**Example 8.6.4** (Solving a linear program with infeasible origin). Apply the simplex method to solve

\[
\begin{align*}
\text{max} & \quad 3x_1 - 2x_2 + x_3 \\
\text{subject to} & \quad x_1 + x_2 + x_3 = 2 \\
& \quad 2x_1 - x_2 = 1 \\
& \quad x_1, x_2, x_3 \geq 0.
\end{align*}
\]

Solution: This problem is already in the standard form required by the simplex method, but a feasible starting basis is perhaps not apparent. We therefore set up a Phase 1 tableau with auxiliary variables and a temporary objective to measure the infeasibilities:

\[
\begin{array}{cccccc|c}
1 & 1 & 1 & 1 & 0 & 2 \\
2 & -1 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & -1 & -1 & 0 \\
\end{array}
\]

108
To begin the simplex method on this Phase 1 problem, we first pivot the auxiliary variables into the basis and thereby zero-out their objective entries. This requires pivoting on row 1, column 4, and then on row 2, column 5. The resulting tableau is

\[
\begin{array}{ccccc|c}
1 & 1 & 1 & 1 & 0 & 2 \\
2 & -1 & 0 & 0 & 1 & 1 \\
3 & 0 & 1 & 0 & 0 & 3 \\
\end{array}
\]

which shows that the origin \( x_1 = x_2 = x_3 = 0 \) has total infeasibilities of \(-z = 3\). Either of columns 1 or 3 may enter the basis:

- Choose column 1 to enter.
- Restrictions on column 1: \( 2/1 = 2 \) by row 1, \( 1/2 \) by row 2.
- Pivot on column 1, row 2. (Column 5 leaves the basis.)

The next tableau is

\[
\begin{array}{ccccc|c}
0 & 3/2 & 1 & 1 & -1/2 & 3/2 \\
1 & -1/2 & 0 & 0 & 1/2 & 1/2 \\
0 & 3/2 & 1 & 0 & -3/2 & 3/2 \\
\end{array}
\]

which corresponds to \((x_1, x_2, x_3) = (1/2, 0, 0)\) and total infeasibilities of \(3/2\). Now we can let either of columns 2 or 3 enter the basis:

- Choose column 2 to enter.
- Restriction on column 2: \((3/2)/(3/2) = 1\) by row 1.
- Pivot on column 2, row 1. (Column 4 leaves the basis.)

This gives us

\[
\begin{array}{ccccc|c}
0 & 1 & 2/3 & 2/3 & -1/3 & 1 \\
1 & 0 & 1/3 & 1/3 & 1/3 & 1 \\
0 & 0 & 0 & -1 & -1 & 0 \\
\end{array}
\]

which corresponds to \((x_1, x_2, x_3) = (1, 1, 0)\) with no infeasibilities. In particular, the auxiliary variables are all nonbasic. This concludes Phase 1: a feasible basis for the original problem consists of \( B = \{1, 2\} \).

We start Phase 2 by entering the data for the standard form problem in a tableau:

\[
\begin{array}{ccccc|c}
1 & 1 & 1 & | & 2 \\
2 & -1 & 0 & | & 1 \\
3 & | & -2 & 1 & 0 \\
\end{array}
\]

Next, we row-reduce the tableau using the given basis \( B = \{1, 2\} \):

\[
\begin{array}{ccccc|c}
1 & 0 & 1/3 & | & 1 \\
0 & 1 & 2/3 & | & 1 \\
0 & 0 & 4/3 & | & -1 \\
\end{array}
\]
which represents the point \( x = (1,1,0) \) with objective value \(-(-1) = 1\). The presence of a positive entry among the objective coefficients indicates that this point may not be optimal, so we apply an iteration of the simplex method:

- Only column 3 may enter.
- Restrictions on column 3: \(1/(1/3) = 3\) by row 1, \(1/(2/3) = 3/2\) by row 2.
- Pivot on column 3, row 2. (Column 2 leaves the basis.)

This yields the updated tableau

\[
\begin{array}{ccc|c}
1 & -1/2 & 0 & 1/2 \\
0 & 3/2 & 1 & 3/2 \\
0 & -2 & 0 & -3 \\
\end{array}
\]

All of the objective coefficients are nonpositive, so this tableau indicates optimality: the solution represented by the tableau is \( x = (1/2,0,3/2) \) with optimal value 3. Note that this \( x \) is indeed feasible for the original LP.

We can find the solution of the dual LP

\[
\begin{align*}
\text{min} & \quad 2y_1 + y_2 \\
\text{subject to} & \quad y_1 + 2y_2 \geq 3 \\
& \quad y_1 - y_2 \geq -2 \\
& \quad y_1 \geq 1
\end{align*}
\]

by using Theorem 8.5.7. That result tells us to solve \( A^T y = c - \tilde{c} \) for \( y \), where \( A \) is the original coefficient matrix, \( c \) is the original objective vector, and \( \tilde{c} \) is the objective row in the optimal Phase 2 tableau. This yields \( y_1 = y_2 = 1 \).

We conclude this section with an example illustrating most of details needed by the simplex method for problems that aren’t in any standard form.

**Example 8.6.5** (Converting and solving a linear program). Solve the following LP using the simplex method:

\[
\begin{align*}
\text{max} & \quad 3x_1 + 4x_2 + 5x_3 \\
\text{subject to} & \quad -4x_1 + 3x_2 + 5x_3 \leq 18 \\
& \quad 2x_1 - x_2 - 2x_3 = 6 \\
& \quad 5x_1 + 3x_2 + x_3 \geq 11 \\
& \quad x_2, x_3 \geq 0.
\end{align*}
\]

Solution: In this example, we explicitly follow the steps indicated in Figure 8.4 and the elaborations made in Observations 8.6.1–8.6.2.

Step I. (Convert the problem to standard form). We must

- replace the free variable \( x_1 \) by its positive and negative parts;
- introduce slack variables for the first and third rows.
This leads to a standard form problem with
\[
c = \begin{bmatrix}
3 \\
-3 \\
4 \\
5 \\
0
\end{bmatrix}, \quad x = \begin{bmatrix}
x_1^+ \\
x_2 \\
x_3 \\
x_4 \\
x_5
\end{bmatrix}, \quad A = \begin{bmatrix}
-4 & 4 & 3 & 5 & 1 & 0 \\
2 & -2 & -1 & -2 & 0 & 0 \\
5 & -5 & 3 & 1 & 0 & -1 \\
\end{bmatrix}, \quad b = \begin{bmatrix}
18 \\
6 \\
11
\end{bmatrix}.
\]

Step II. (Find a basic feasible solution). Setting the non-slack variables to zero we obtain the infeasible point
\[
x_1^+ = x_1^- = x_2 = x_3 = 0, \quad x_4 = 18, x_5 = -11.
\]

Introduce auxiliary variables for the second and third rows, and change the objective to measure the infeasibilities. The Phase 1 problem is an LP in standard form with
\[
c = \begin{bmatrix}
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
-1 \\
-1
\end{bmatrix}, \quad x = \begin{bmatrix}
x_1^+ \\
x_1^- \\
x_2 \\
x_3 \\
x_4 \\
x_5 \\
x_6 \\
x_7
\end{bmatrix}, \quad A = \begin{bmatrix}
-4 & 4 & 3 & 5 & 1 & 0 & 0 & 0 \\
2 & -2 & -1 & -2 & 0 & 0 & 1 & 0 \\
5 & -5 & 3 & 1 & 0 & -1 & 0 & 1 \\
\end{bmatrix}, \quad b = \begin{bmatrix}
18 \\
6 \\
11
\end{bmatrix}.
\]

We enter our Phase 1 problem as the following tableau:
\[
\begin{array}{cccccccc|c}
-4 & 4 & 3 & 5 & 1 & 0 & 0 & 0 & 18 \\
2 & -2 & -1 & -2 & 0 & 0 & 1 & 0 & 6 \\
5 & -5 & 3 & 1 & 0 & -1 & 0 & 1 & 11 \\
\hline
0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0
\end{array}
\]

Taking a basis of slacks and auxiliaries gives us \( B = \{5, 7, 8\} \). We pivot on each of these columns to correct the signs and clear their columns: column 5, row 1; column 7, row 2; column 8, row 3. This leaves us with the following starting tableau for Phase 1:
\[
\begin{array}{cccccccc|c}
-4 & 4 & 3 & 5 & 1 & 0 & 0 & 0 & 18 \\
2 & -2 & -1 & -2 & 0 & 0 & 1 & 0 & 6 \\
5 & -5 & 3 & 1 & 0 & -1 & 0 & 1 & 11 \\
\hline
7 & -7 & 2 & -1 & 0 & -1 & 0 & 0 & 17
\end{array}
\]

As always, our entering variables should have positive objective coefficients.

- Either 1 or 3 may enter: take column 1.
• Restrictions on column 1: row 1—none, row 2—$6/2 = 3$, row 3—$11/5 = 2.2$. Therefore pivot on column 1, row 3. (Column 8 leaves basis.)

\[
\begin{array}{cccccccc}
0 & 0 & 27/5 & 29/5 & 1 & -4/5 & 0 & 4/5 & 134/5 \\
0 & 0 & -11/5 & -12/5 & 0 & 2/5 & 1 & -2/5 & 8/5 \\
1 & -1 & 3/5 & 1/5 & 0 & -1/5 & 0 & 1/5 & 11/5 \\
\end{array}
\]

• Only column 6 may enter.

• Only restriction on column 6 is row 2: pivot on that entry. (Column 7 leaves the basis.)

\[
\begin{array}{cccccccc}
0 & 0 & 27/5 & 29/5 & 1 & -4/5 & 0 & 4/5 & 134/5 \\
0 & 0 & -11/5 & -12/5 & 0 & 2/5 & 1 & -2/5 & 8/5 \\
1 & -1 & 3/5 & 1/5 & 0 & -1/5 & 0 & 1/5 & 11/5 \\
\end{array}
\]

This shows that the original problem is feasible, and the standard form problem has a feasible basis given by $B = \{1, 5, 6\}$. This corresponds to the point

\[
x_1^+ = 3, \quad x_1^- = x_2 = x_3 = 0, \quad x_4 = 30, \quad x_5 = 4.
\]

Step III. (Find an optimal solution). Now we turn to Phase 2. We start by entering the data for the original problem (in standard form):

\[
\begin{array}{ccccc}
-4 & 4 & 3 & 5 & 1 & 0 & 18 \\
2 & -2 & -1 & -2 & 0 & 0 & 6 \\
5 & -5 & 3 & 1 & 0 & -1 & 11 \\
3 & -3 & 4 & 5 & 0 & 0 & 0 \\
\end{array}
\]

Our starting feasible basis is $B = \{1, 5, 6\}$, so we pivot to row-reduce the system to this basis: column 1, row 2; column 5, row 1; column 6, row 3. (This can be done in any order which works—the order used here happens to make the simplest changes.)

\[
\begin{array}{cccccccc}
0 & 0 & 1 & 1 & 1 & 0 & 30 \\
1 & -1 & -1/2 & -1 & 0 & 0 & 3 \\
0 & 0 & -11/2 & -6 & 0 & 1 & 4 \\
\end{array}
\]

Note that this tableau represents the same point as the final tableau in Phase 1. Because we are maximizing, entering variables must have positive objective coefficients.

• Either of the columns 3 or 4 may enter: choose 4.

• Only restriction on column 4 is row 1: pivot on that entry. (Column 5 leaves the basis.)

\[
\begin{array}{cccccccc}
0 & 0 & 1 & 1 & 1 & 0 & 30 \\
1 & -1 & 1/2 & 0 & 1 & 0 & 33 \\
0 & 0 & 1/2 & 0 & 6 & 1 & 184 \\
0 & 0 & -5/2 & 0 & -8 & 0 & -249 \\
\end{array}
\]

112
Since no objective coefficient is positive, we cannot move to a better point: this tableau represents the optimum! We read the solution from the tableau as

\[ x_1^+ = 33, \; x_1^- = 0, \; x_2 = 0, \; x_3 = 30, \; x_4 = 0, \; x_5 = 184. \]

In the original variables, the solution is \( x_1 = 33, \; x_2 = 0, \; x_3 = 30 \) with optimal value 249.

Step IV. (Find a dual solution). The dual of the given LP is

\[
\begin{align*}
\text{min} & \quad 18y_1 + 6y_2 + 11y_3 \\
\text{subject to} & \quad -4y_1 + 2y_2 + 5y_3 \geq 3 \\
& \quad 3y_1 - y_2 + 3y_3 \geq 4 \\
& \quad 5y_1 - 2y_2 + y_3 \geq 5 \\
& \quad y_1 \geq 0, \; y_3 \leq 0.
\end{align*}
\]

To obtain a solution \( y \), we use the final tableau to calculate a solution \( u \) for the dual of the standard-form problem, and then convert \( u \) into \( y \). In this case, the dual of the standard-form problem is

\[
\begin{align*}
\text{min} & \quad 18u_1 + 6u_2 + 11u_3 \\
\text{subject to} & \quad -4u_1 + 2u_2 + 5u_3 \geq 3 \\
& \quad 4u_1 - 2u_2 - 5u_3 \geq -3 \\
& \quad 3u_1 - u_2 + 3u_3 \geq 4 \\
& \quad 5u_1 - 2u_2 + u_3 \geq 5 \\
& \quad u_1 \geq 0 \\
& \quad -u_3 \geq 0.
\end{align*}
\]

We see that the correspondence between \( u \) and \( y \) is trivial: we can simply take \( y_i = u_i \). To calculate \( u \) from the final tableau in the preceding Step, we solve \( A^T u = c - \tilde{c} \) for \( u \).

This system has the form

\[
\begin{bmatrix}
-4 & 2 & 5 \\
4 & -2 & -5 \\
3 & -1 & 3 \\
5 & -2 & 1 \\
1 & 0 & 0 \\
0 & 0 & -1
\end{bmatrix}
\begin{bmatrix}
u_1 \\
u_2 \\
u_3
\end{bmatrix}
= \begin{bmatrix}3 \\
-3 \\
4 \\
5 \\
0 \\
0
\end{bmatrix} - \begin{bmatrix}0 \\
0 \\
-5/2 \\
0 \\
-8 \\
0
\end{bmatrix},
\]

which gives us \( u_1 = 8, \; u_2 = 35/2, \; u_3 = 0 \), so the desired vector \( y \) is \( (y_1, y_2, y_3) = (8, 35/2, 0) \). It is easily verified that this is dual feasible with objective value 249, which validates the primal solution obtained in Step III above.
8.7 Primal-Dual Interior Point Methods for Linear Programs

We now outline another approach, known as *interior point methods* (IPM), to solving linear programming problems. This general class of methods seems to be the most effective for solving very large problems with up to hundreds of millions of variables and constraints. Consider the linear program

\[(\mathcal{P}) \quad \min \{ c \cdot x \mid Ax = b, \ x \geq 0 \} \]

and its dual

\[(\mathcal{D}) \quad \max \{ b \cdot y \mid A^T y + s = c, \ s \geq 0 \}. \]

Note that we explicitly include the dual slack variables \(s\).

**Theorem 8.7.1 (KKT for LP).** The KKT conditions for \((\mathcal{P})\) and \((\mathcal{D})\) can be stated as

\[A^T y + s = c, \quad Ax = b, \quad XSe = 0, \quad (x, s) \geq 0, \]

where \(X = \text{diag}(x_1, \ldots, x_n)\) and \(S = \text{diag}(s_1, \ldots, s_n)\) are diagonal matrices defined in terms of the vectors \(x\) and \(s\).

**Observation 8.7.2 (IPM notation).** In the above theorem we rewrite the familiar complementary slackness condition

\[x_i s_i = 0, \ \forall i, \]

in the form \(XSe = 0\), where \(e\) is the \(n\)-vector of all ones. Note that \(x = Xe\) and \(s = Se\). Because \((x, s)\) is nonnegative, the complementarity condition is equivalent to \(x \cdot s = 0\). \(\Box\)

**Proof of Theorem 8.7.1.** First, consider problem \((\mathcal{P})\). Let the vector \(\eta\) denote the multipliers for the constraints \(b - Ax = 0\) and let \(\sigma\) denote the multipliers for the constraints \(-Ix \leq 0\). The KKT conditions for optimality of \(x\) in \((\mathcal{P})\) are that \((\eta, \sigma)\) exists so that \((x, \eta, \sigma)\) satisfy:

\[x \leq 0, \quad \sigma \geq 0, \quad \sigma \cdot (-Ix) = 0, \quad c - A^T \eta - I\sigma = 0.\]

These can be equivalently restated as

\[Ax = b, \quad x \geq 0, \quad \sigma \geq 0, \quad \sigma \cdot x = 0, \quad A^T \eta + \sigma = c.\]

Second, consider problem \((\mathcal{D})\). Let the multiplier vectors for the constraints

\[c - [A^T, I] \begin{bmatrix} y \\ s \end{bmatrix} = 0 \quad \text{and} \quad [0, -I] \begin{bmatrix} y \\ s \end{bmatrix} \leq 0 \]

be denoted by \(\xi\) and \(\zeta\), respectively. The KKT conditions for optimality of \((y, s)\) in \((\mathcal{D})\) are that \((\xi, \zeta)\) exists so that \((y, s, \xi, \zeta)\) satisfy:

\[c - [A^T, I] \begin{bmatrix} y \\ s \end{bmatrix} = 0, \quad [0, -I] \begin{bmatrix} y \\ s \end{bmatrix} \leq 0, \quad \zeta \leq 0, \quad \zeta \cdot \left( [0, -I] \begin{bmatrix} y \\ s \end{bmatrix} \right) = 0, \]

\[\begin{bmatrix} b \\ 0 \end{bmatrix} - \begin{bmatrix} A \\ I \end{bmatrix} \xi + \begin{bmatrix} 0 \\ -I \end{bmatrix} \zeta = 0.\]
Equivalently, we need for \((y, s, \xi)\) to satisfy
\[
A^T y + s = c, \quad s \geq 0, \quad \xi \geq 0, \quad \xi \cdot s = 0, \quad A\xi = b.
\]

Now it is clear that the KKT conditions are equivalent to each other by the correspondence
\[
x \longleftrightarrow \xi, \quad y \longleftrightarrow \eta, \quad s \longleftrightarrow \sigma.
\]

Moreover, by using these substitutions and replacing the complementary slackness condition \(s \cdot x = 0\) by \(SXe = 0\), the above conditions become the same as those given in the statement of the theorem.

The equations in the KKT conditions form a nearly linear system, so it would be reasonable to attempt solving it with Newton’s method. An iteration of Newton’s method would then consist of solving the linear system

\[
(N\text{-KKT}) \begin{bmatrix}
0 & A^T & I \\
A & 0 & 0 \\
S^{(k)} & 0 & X^{(k)}
\end{bmatrix} \begin{bmatrix}
\Delta x^{(k)} \\
\Delta y^{(k)} \\
\Delta s^{(k)}
\end{bmatrix} = \begin{bmatrix}
c - A^T y^{(k)} - s^{(k)} \\
b - Ax^{(k)} \\
-X^{(k)} S^{(k)} e
\end{bmatrix}
\]

and setting
\[
\begin{bmatrix}
x^{(k+1)} \\
y^{(k+1)} \\
s^{(k+1)}
\end{bmatrix} = \begin{bmatrix}
x^{(k)} \\
y^{(k)} \\
s^{(k)}
\end{bmatrix} + \begin{bmatrix}
\Delta x^{(k)} \\
\Delta y^{(k)} \\
\Delta s^{(k)}
\end{bmatrix}.
\]

For the following discussion we assume that \(A\) is \(m \times n\) with full row rank, so that the coefficient matrix on the left-hand side of \((N\text{-KKT})\) is actually invertible.

The majority of solutions for the KKT equations of a linear program do not satisfy the nonnegativity conditions for \((x, s)\), and most feasible starting points for Newton’s method lead eventually to one of these unsatisfactory points. To compensate for this, Newton’s method can be “truncated”: rather than taking the full Newton step, move just far enough to avoid introducing negative components in \((x, s)\). However, the iterates are still drawn toward an infeasible solution of the KKT equations, so they tend to pile up near the boundary. To fully deal with this difficulty, an interior point method “biases” the Newton step away from the boundary by slightly modifying the KKT conditions: we replace \(XSe = 0\) by \(XSe = \tau e\), where \(\tau\) is a small positive real number, and require \(x\) and \(s\) to remain positive. This leads to a new system:

\[
(C_\tau) \quad A^T y + s = c, \quad Ax = b, \quad XSe = \tau e, \quad (x, s) > 0.
\]

A solution \((x_\tau, y_\tau, s_\tau)\) to this system is a strictly feasible primal-dual point for \((P)\) and \((D)\). It is easy to see that \(x \cdot s = n\tau\) and to use this to show that \(n\tau\) is actually the duality gap for \((x, y, s)\). It can be proved, in fact, that if strictly feasible primal-dual points exist then for each \(\tau > 0\) there is a unique point satisfying the system \((C_\tau)\). Moreover, \(\{(x_\tau, y_\tau, s_\tau) : \tau > 0\}\) defines a smooth parametric arc of strictly feasible primal-dual points, known as the central path. As \(\tau \to 0^+\) the central path approaches some KKT point.
The idea of a primal-dual interior point method is to follow the central path toward a KKT solution. Variants of Newton’s method are used to move toward a point on the path, combined with some means of decreasing the parameter \( \tau \) to zero. Typically, at a given point \((x, y, s)\) with \((x, s) > 0\) none of the equations defining the central path are satisfied. To move toward the path without increasing the value of \(x \cdot s\), we calculate a Newton step for solving \( C_{\tau} \) with a suitably chosen value of \( \tau \in [0, 1/n \cdot x \cdot s] \). To avoid losing the strict positivity of \((x, s)\), we either have to truncate the step or else stay close enough to the central path to allow full Newton steps. Here is a general framework for such methods.

**Method 8.7.3 (General primal-dual algorithm for LP)**. Given \((x^{(0)}, y^{(0)}, s^{(0)}) > 0\), generate a sequence of iterates \((x^{(k)}, y^{(k)}, s^{(k)})\) according to the following procedure:

1. (Search Direction) Choose \(\tau^{(k)} \in [0, 1/n \cdot x^{(k)} \cdot s^{(k)}]\) and calculate the Newton step for the nonlinear system \(C_{\tau^{(k)}}\) by solving the linear system

\[
(N-C_{\tau^{(k)}}) \begin{bmatrix} 0 & A^T & I \\ A & 0 & 0 \\ S^{(k)} & 0 & X^{(k)} \end{bmatrix} \begin{bmatrix} \Delta x^{(k)} \\ \Delta y^{(k)} \\ \Delta s^{(k)} \end{bmatrix} = \begin{bmatrix} c - A^T y^{(k)} - s^{(k)} \\ b - A x^{(k)} \\ \tau^{(k)} e - X^{(k)} S^{(k)} e \end{bmatrix}.
\]

2. (Step-size) Choose \(t^{(k)} \in (0, \alpha^{(k)})\), where

\[
\alpha^{(k)} = \max\{ \alpha \geq 0 : (x^{(k)}, s^{(k)}) + \alpha(\Delta x^{(k)}, \Delta s^{(k)}) \geq 0 \}.
\]

3. (Update) Calculate

\[
\begin{bmatrix} x^{(k+1)} \\ y^{(k+1)} \\ s^{(k+1)} \end{bmatrix} = \begin{bmatrix} x^{(k)} \\ y^{(k)} \\ s^{(k)} \end{bmatrix} + t^{(k)} \begin{bmatrix} \Delta x^{(k)} \\ \Delta y^{(k)} \\ \Delta s^{(k)} \end{bmatrix}.
\]

Replace \(k\) by \(k + 1\) and go to step 1.

The above algorithm can take many forms, depending on the specific rules used to determine \(\tau^{(k)}\) and \(t^{(k)}\). Here is a simple variant that is provably convergent, provided that the initial guess \((x^{(0)}, y^{(0)}, s^{(0)})\) is primal-dual feasible:

Use \(t^{(k)} = 1\) and

\[
\tau^{(k)} = \begin{cases} 
\left(1 - \frac{0.4}{\sqrt{n}}\right) \mu^{(k)}, & \text{if } \|X^{(k)} S^{(k)} e - \mu^{(k)} e\| \leq 0.4 \mu^{(k)} , \\
\mu^{(k)}, & \text{otherwise},
\end{cases}
\]

where we define \(\mu^{(k)} = x^{(k)} \cdot s^{(k)}/n\).

Rules for specifying \(\tau^{(k)}\) and \(t^{(k)}\) are more complicated when the current iterate is not primal and dual feasible. However, if a full Newton step is ever taken (i.e., if \(t^{(k)} = 1\)), then all subsequent iterates will be primal-dual feasible.
Note that Method 8.7.3 generates an infinite sequence of primal-dual points, so it must be terminated when deemed sufficiently close to a solution. A typical test for numerical convergence is to stop when the following quantities are all less than a specified small positive number (e.g., $10^{-8}$):

\[
\frac{\|Ax^{(k)} - b\|}{\|b\|}, \quad \frac{\|ATy^{(k)} + s^{(k)} - c\|}{\|c\|}, \quad \frac{|c \cdot x^{(k)} - b \cdot y^{(k)}|}{1 + |c \cdot x^{(k)}|}.
\]

Unlike the simplex method, an interior point method will not find an exact solution. However, once an iterate is sufficiently close to the set of KKT points, it can be projected onto an actual KKT point by solving a linear least squares problem. If desired, a basic optimal solution can then be found by performing a few simplex-type pivoting operations.
A. APPENDIX — Additional Theory and Proofs

The appendix provides extended proofs omitted from the main text.

A.1 Positive Definite Matrices

This section fills in the theoretical details needed to justify the procedures of chapter 2 for determining the definiteness of a matrix.

Lemma A.1.1 (Block-matrix representation of determinant). Suppose

\[
A = \begin{bmatrix}
P & u \\
u^T & q
\end{bmatrix}
\]

is an \(n \times n\) matrix, where \(P\) is a nonsingular \((n-1) \times (n-1)\) matrix, \(u\) and \(v\) are (column) vectors in \(\mathbb{R}^{n-1}\), and \(q\) is a scalar. Then \(\det(A) = \det(P)(q - v \cdot P^{-1}u)\).

Proof. Let \(S = \begin{bmatrix} I & 0 \\ -v^TP^{-1} & 1 \end{bmatrix}\), so that \(\det(S) = 1\). Then

\[
\det(A) = \det(SA) = \det\begin{bmatrix} P & u \\ 0 & q - v \cdot P^{-1}u \end{bmatrix} = \det(P)(q - v \cdot P^{-1}u),
\]

where the last equality follows by expanding the determinant along the bottom row.

Theorem A.1.2 (Characterizations of positive definiteness). Let \(A\) be a symmetric matrix. The following are equivalent:

(a) \(A\) is positive definite;

(b) the leading principal minors of \(A\) are positive;

(c) \(A = U^TU\) for some upper triangular matrix \(U\) whose diagonal entries are positive.

Proof. (a)⇒(b): We prove this by induction on the order of the matrix \(A\). It is clearly true for \(1 \times 1\) matrices. Suppose it is valid for all \((n-1) \times (n-1)\) symmetric matrices, and consider a symmetric matrix \(A \in \mathbb{R}^{n \times n}\) of the form

\[
A = \begin{bmatrix} P & u \\ u^T & q \end{bmatrix},
\]

for a symmetric matrix \(P \in \mathbb{R}^{(n-1) \times (n-1)}\), a vector \(u \in \mathbb{R}^{n-1}\) and a scalar \(q \in \mathbb{R}\). Assume \(A\) is positive definite. For any nonzero vector \(x \in \mathbb{R}^{n-1}\) we have

\[
0 < \begin{bmatrix} x \\ 0 \end{bmatrix} \cdot \begin{bmatrix} P & u \\ u^T & q \end{bmatrix} \begin{bmatrix} x \\ 0 \end{bmatrix} = x \cdot Px,
\]

so \(P\) is positive definite. By the inductive hypothesis, all leading minors of \(P\) are positive: thus all leading minors of \(A\) (except possibly \(\det(A)\)) are positive. Lemma 1 tells us that \(\det(A) = \det(P)(q - u^TPu)\). Since \(A\) is positive definite we have

\[
0 < \begin{bmatrix} -P^{-1}u \\ 1 \end{bmatrix} \cdot \begin{bmatrix} P & u \\ u^T & q \end{bmatrix} \begin{bmatrix} -P^{-1}u \\ 1 \end{bmatrix} = q - u^TP^{-1}u.
\]
Therefore $\det(A) > 0$, as desired.

(b)$\Rightarrow$(c): We prove this by induction on the order of the matrix $A$. It is clearly true for $1 \times 1$ matrices. Suppose it is valid for all $(n-1) \times (n-1)$ symmetric matrices, and consider a symmetric matrix $A \in \mathbb{R}^{n \times n}$ of the form

$$A = \begin{bmatrix} P & u \\ u^T & q \end{bmatrix},$$

for a symmetric matrix $P \in \mathbb{R}^{(n-1) \times (n-1)}$, a vector $u \in \mathbb{R}^{n-1}$ and a scalar $q \in \mathbb{R}$. Assume that $A$ has all positive leading principal minors. Then $P$ has all positive leading principal minors, so the inductive hypothesis tells us that $P = M^T M$ for some upper triangular matrix $M$ with positive diagonal entries. At the same time, $0 < \det(A) = \det(P)(q - u^T P^{-1} u)$ so $q - u^T P^{-1} u > 0$. If we define

$$U = \begin{bmatrix} M & (M^{-1})^T u \\ 0 & \sqrt{q - u^T P^{-1} u} \end{bmatrix},$$

then it is not difficult to verify that $A = U^T U$.

(c)$\Rightarrow$(a): Since $U$ is nonsingular, we see that $Ux \neq 0$ whenever $x \neq 0$. Consequently

$$x \cdot Ax = x \cdot U^T Ux = (Ux) \cdot (Ux) = \|Ux\|^2 > 0$$

for any $x \neq 0$. Thus $A$ is positive definite. This completes the proof of the theorem.

**Theorem A.1.3** (Determining definiteness from a block-triangular factorization). Let $A$ be a symmetric matrix factored as

$$A = \begin{bmatrix} L & 0 \\ M & I \end{bmatrix} \begin{bmatrix} U & N \end{bmatrix},$$

where $L$ is an invertible matrix. Then the following correspondences hold:

- $A$ is PD $\iff$ both $LU$ and $Q$ are PD;
- $A$ is PSD $\iff$ both $LU$ and $Q$ are PSD;
- $A$ is ND $\iff$ both $LU$ and $Q$ are ND;
- $A$ is NSD $\iff$ both $LU$ and $Q$ are NSD.

In particular, $A$ is indefinite if either $LU$ or $Q$ are indefinite.

**Proof.** Calculating the product defining $A$ gives

$$A = \begin{bmatrix} LU & LN \\ MU & MN + Q \end{bmatrix}.$$

By symmetry of $A$ we have $LU = (LU)^T$, $LN = (MU)^T$, and $MN + Q = (MN + Q)^T$. These can be rewritten, respectively, as

$$U(L^T)^{-1} = L^{-1}U^T,$$
$$N = L^{-1}U^T M^T,$$
$$Q = Q^T + (MN)^T - MN.$$
Note that (1) proves the symmetry of $U(L^T)^{-1}$. Combining (1) and (2) yields $N = U(L^T)^{-1}M^T$. This guarantees the symmetry of $MN = ML^{-1}U^TM^T$, which in turn, by (3), and proves the symmetry of $Q$. Putting all this together allows us to factor $A$ as

$$A = \begin{bmatrix} L & 0 \\ M & I \end{bmatrix} \begin{bmatrix} U(L^T)^{-1} & 0 \\ 0 & Q \end{bmatrix} \begin{bmatrix} L^T & M^T \\ 0 & I \end{bmatrix}.$$ 

Here the right and left factors are invertible, whereas the middle factor is symmetric. The desired correspondences now follow directly.

**Proposition A.1.4** (Determining definiteness from an LU decomposition). Suppose that $A = LU$, where $A$ is symmetric, $L$ is lower triangular with ones on its diagonal, and $U$ is upper triangular. Then the following correspondences hold:

- $A$ is PD $\iff$ all diagonal entries of $U$ have $d_{ii} > 0$;
- $A$ is PSD $\iff$ all diagonal entries of $U$ have $d_{ii} \geq 0$;
- $A$ is ND $\iff$ all diagonal entries of $U$ have $d_{ii} < 0$;
- $A$ is NSD $\iff$ all diagonal entries of $U$ have $d_{ii} \leq 0$.

Consequently, if any pair of nonzero diagonal entries of $U$ have opposite sign, then $A$ is indefinite.

**Proof.** Let $D$ be the diagonal matrix whose diagonal entries are the same as those in $U$. It is easy to show that the symmetry of $A$ implies that $U = DL^T$. Since $L$ is nonsingular, the result follows.

**Proposition A.1.5** (Semidefiniteness guarantees LU factorization). Suppose that $A$ is a symmetric, semidefinite matrix. Then $A = LU$, where $L$ is lower triangular with ones on the diagonal and $U$ is upper triangular.

**Proof.** It suffices to prove the result for positive semidefinite matrices $A$. We proceed by induction. Assume the result is true for $k \times k$ matrices with $k < n$, and consider an $n \times n$ positive semidefinite matrix $A$. Then $A$ has the form

$$A = \begin{bmatrix} P & u \\ u^T & \alpha \end{bmatrix}$$

where $P$ is positive semidefinite. By the inductive hypothesis, $P = MN$ for some lower triangular matrix $M$ with ones on the diagonal and some upper triangular matrix $N$. We now show that there must exist a vector $v$ satisfying $Pv = u$. Suppose that no such vector existed. Then $u$ is not in the range of $P$, so $u$ cannot be orthogonal to the null-space of $P$. Hence there is a vector $x$ satisfying $x \cdot u = 1$ and $Px = 0$. For any real number $\lambda$ we have

$$\begin{bmatrix} \lambda x \\ 1 \end{bmatrix} \cdot A \begin{bmatrix} \lambda x \\ 1 \end{bmatrix} = \begin{bmatrix} \lambda x \\ 1 \end{bmatrix} \cdot \begin{bmatrix} P & u \\ u^T & \alpha \end{bmatrix} \begin{bmatrix} \lambda x \\ 1 \end{bmatrix} = 2\lambda + \alpha.$$
This, however, violates the semidefiniteness of $A$. Consequently, there is indeed a vector $v$ with $Pv = u$. It is easily verified that $A = LU$ for the choice

$$L = \begin{bmatrix} M & 0 \\ v^T M & 1 \end{bmatrix}, \quad U = \begin{bmatrix} N & M^{-1}Pv \\ 0 & \alpha - v^T Pv \end{bmatrix}. $$

This completes the proof.

Here is an alternative proof that positive definiteness is “preserved locally”, using the same idea as in Proposition 3.2.1(c) on the coercivity of positive definite quadratic forms.

**Proposition 2.3.15(d)** (Local preservation of definiteness). *If $A$ is positive definite, then there exists $\delta > 0$ so that $B$ is positive definite whenever $|a_{ij} - b_{ij}| < \delta$ for all $i, j$.***

**Proof.** Take $\epsilon = \tilde{z} \cdot A \tilde{z} > 0$, where $\tilde{z}$ is a global minimizer of the continuous function $z \cdot Az$ over the closed and bounded set of $z$ with $\|z\| = 1$. Then we have $x \cdot Ax \geq \epsilon \|x\|^2$ for all $x$. (To see why, simply consider $x \neq 0$ and $z = x/\|x\|$; then $\|z\| = 1$, so $z \cdot Az \geq \epsilon$.)

Next, define $\delta = \epsilon/n$, where $n$ is the (column) dimension of $A$. Consider any $n \times n$ matrix $B$ for which $|b_{ij} - a_{ij}| < \delta$ for all $i, j$, and let $x \neq 0$. Let $[a^{(i)}]^T$ and $[b^{(i)}]^T$ denote the $i^{th}$ rows of $A$ and $B$. Using the Schwarz inequality twice, we obtain

$$|x \cdot (B - A)x| \leq \|x\| \|(B - A)x\| = \|x\| \left\| \left[ \begin{array}{c} (b^{(1)} - a^{(1)}) \cdot x \\ \vdots \\ (b^{(n)} - a^{(n)}) \cdot x \end{array} \right] \right\| = \|x\| \sqrt{\sum_{i=1}^{n} |(b^{(i)} - a^{(i)}) \cdot x|^2}$$

$$\leq \|x\| \sqrt{\sum_{i=1}^{n} \|b^{(i)} - a^{(i)}\|^2 \|x\|^2} = \|x\|^2 \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} (b_{ij} - a_{ij})^2} < \|x\|^2 \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} \delta = \|x\|^2 n\delta = \epsilon \|x\|^2}. $$

Thus $|x \cdot (B - A)x| < \epsilon \|x\|^2$, which gives us

$$x \cdot Bx = x \cdot Ax + x \cdot (B - A)x \geq x \cdot Ax - |x \cdot (B - A)x| > \epsilon \|x\|^2 - \epsilon \|x\|^2 = 0. $$

This completes the proof.

**Corollary A.1.7** (Local preservation of full rank). *If $A$ has full column rank, then there exists $\delta > 0$ so that $B$ has full column rank whenever $|a_{ij} - b_{ij}| < \delta$ for all $i, j$.***

**Proof.** This follows from Proposition 2.3.8, the above Proposition, and the continuity of the entries in $B^TB$ with respect to the entries in $B$. \qed
A.2 Coercivity

This section gives proofs of key concepts presented in Chapter 3 regarding the verification of coercivity, along with statements of more general results needed to prove those concepts. The results and proofs here remain valid if all occurrences of the Euclidean norm \( \|x\| \) are replaced by \( \max_i \{|x_i|\} \). We shall assume the following easily verified facts.

**Proposition A.2.1** (Finite combinations of coercive functions). If \( f_1, \ldots, f_k \) are all coercive on \( \mathbb{R}^n \), then \( \sum_i f_i \), \( \min_i \{ f_i \} \), and \( \max_i \{ f_i \} \) are also coercive on \( \mathbb{R}^n \).

**Proposition A.2.2** (Lim-inf form of coercivity). The function \( f : \mathbb{R}^n \to \mathbb{R} \) is coercive if and only if
\[
\lim_{s \to \infty} \inf_x \{ f(x) : \|x\| \geq s \} = \infty.
\]

**Proposition A.2.3** (Linear composition with coercive function). The function \( f(x) = g(Ax) \) is coercive if \( g \) is coercive and \( A \) has full column rank.

**Theorem A.2.4** (Sum of coercive functions — variables partitioned into two sets). If \( g : \mathbb{R}^k \to \mathbb{R} \) and \( h : \mathbb{R}^{n-k} \to \mathbb{R} \) are coercive and bounded below, then \( f : \mathbb{R}^n \to \mathbb{R} \) defined by \( f(x, y) = g(x) + h(y) \) is coercive.

**Proof.** Let \( P \) and \( Q \) be lower bounds for \( f \) and \( g \), respectively. By the definition of coercivity, we need to prove that for each \( M \in \mathbb{R} \) there is a number \( N \) so that \( \|(x, y)\| < N \) whenever \( f(x, y) < M \). Consider \( M \in \mathbb{R} \) and choose \( N \) so that \( \|x\| < N/2 \) whenever \( g(x) < M - Q \), and \( \|y\| < N/2 \) whenever \( h(y) < M - P \). Now suppose that \( (x, y) \) satisfies \( f(x, y) < M \). Then \( g(x) + h(y) < M \), so \( g(x) < M - Q \) and hence \( \|x\| < N/2 \); likewise, \( h(y) < M - P \) and so \( \|y\| < N/2 \). Thus
\[
\|(x, y)\| = \|(x, 0) + (0, y)\| \leq \|(x, 0)\| + \|(0, y)\| = \|x\| + \|y\| < N,
\]
as desired. \( \square \)

**Corollary A.2.5** (Sum of coercive functions — variables partitioned into \( k \) sets). Assume that \( \phi_i : \mathbb{R}^{n_i} \to \mathbb{R} \) and \( f(x^{(1)}, \ldots, x^{(k)}) = \sum_{i=1}^k \phi_i(x^{(i)}) \). If each \( \phi_i \) is coercive and bounded below, then \( f \) is coercive.

**Proof.** Use induction on \( k \) and apply Theorem A.2.4 in the inductive step. \( \square \)

**Corollary A.2.6** (Sum of coercive functions — overlapping variables). Suppose that \( f : \mathbb{R}^n \to \mathbb{R} \) can be written as a sum of expressions in the variables \( x_1, \ldots, x_n \) such that each variable appears in at least one summand and each summand is coercive and bounded below in the variables that appear in it. Then \( f \) is coercive.

**Proof.** Assume there are \( k \) summands and the \( i \)th summand defining \( f \) can be expressed as the composition of a coercive and lower-bounded function \( y^{(i)} \mapsto \phi_i(y^{(i)}) \) with the projection \( x \mapsto A^{(i)}x \) of \( x = (x_1, \ldots, x_n) \) onto the variables appearing in the \( i \)th summand.
By Corollary A.2.5, the mapping $g : (y^{(1)}, \ldots, y^{(k)}) \mapsto \sum_i \phi_i(y^{(i)})$ is coercive, whereas the matrix
\[ A = \begin{bmatrix} A^{(1)} \\ \vdots \\ A^{(k)} \end{bmatrix} \]
has full column rank. Thus $f(x) = g(Ax)$ is coercive.

**Example A.2.7** (Need for lower bound on summands). This example shows that a sum of coercive functions on distinct coordinates need not be coercive if the summand functions are *unbounded* below. Define
\[ g(t) = \begin{cases} \ln |t| & \text{if } t \neq 0, \\ 0 & \text{if } t = 0. \end{cases} \]
Then $\lim_{|t| \to \infty} g(t) = \infty$, so $g$ is coercive. However, $f(x, y) = g(x) + g(y)$ is not coercive, as can be determined by considering the sequence $(1/k, k)$.

**Theorem A.2.8** (Max of coercive functions — variables partitioned into two sets). If $g : \mathbb{R}^k \to \mathbb{R}$ and $h : \mathbb{R}^{n-k} \to \mathbb{R}$ are coercive, then $f : \mathbb{R}^n \to \mathbb{R}$ defined by $f(x, y) = \max\{g(x), h(y)\}$ is coercive.

**Proof.** Consider $M \in \mathbb{R}$. Choose $N$ so that $\|x\| < N/2$ whenever $g(x) < M$ and $\|y\| < N/2$ whenever $h(y) < M$. If $(x, y)$ satisfies $f(x, y) < M$, then $\|x\| < N/2$ and $\|y\| < N/2$, so $\|(x, y)\| < N$ (as in the proof of Theorem A.2.4).

**Corollary A.2.9** (Max of coercive functions — variables partitioned into $k$ sets). Suppose that $\phi_i : \mathbb{R}^{n_i} \to \mathbb{R}$ and $f(x_1, \ldots, x_k) = \max \{ \phi_1(x_1), \ldots, \phi_k(x_k) \}$. If each $\phi_i$ is coercive, then $f$ is coercive.

**Proof.** Use induction on $k$ and apply Theorem A.2.8 in the inductive step.

**Corollary A.2.10** (Sum of coercive functions — overlapping variables). Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ can be written as a point-wise maximum of expressions in the variables $x_1, \ldots, x_n$ such that each variable appears in at least one operand and each operand is coercive in the variables that appear in it. Then $f$ is coercive.

**Proof.** Assume there are $k$ summands and the $i^{th}$ summand defining $f$ can be expressed as the composition of a coercive function $y^{(i)} \mapsto \phi_i(y^{(i)})$ with the projection $x \mapsto A^{(i)}x$ of $x = (x_1, \ldots, x_n)$ onto the variables appearing in the $i^{th}$ summand. By Corollary A.2.9, the mapping $g : (y^{(1)}, \ldots, y^{(k)}) \mapsto \max_i \phi_i(y^{(i)})$ is coercive, whereas the matrix
\[ A = \begin{bmatrix} A^{(1)} \\ \vdots \\ A^{(k)} \end{bmatrix} \]
has full column rank. Thus $f(x) = g(Ax)$ is coercive.
Lemma A.2.11 (Lim-inf coercivity of a sum). The function $f(x) = g(x) + h(x)$ is coercive if
\[
\lim_{t \to \infty} \left[ \inf_x \{g(x) : \|x\| \geq t\} + \inf_x \{h(x) : \|x\| \leq t\} \right] = \infty.
\]

**Proof.** We use the fact that $f$ is coercive if and only if
\[
\lim_{s \to \infty} \inf_y \{f(y) : \|y\| \geq s\} = \infty.
\]
In other words, given $M \in \mathbb{R}$ we need to show that there exists $T$ so that
\[
\inf_y \{f(y) : \|y\| \geq s\} \geq M
\]
whenever $s \geq T$. By the hypothesis, we may choose $T$ so that
\[
\inf_x \{g(x) : \|x\| \geq t\} + \inf_x \{h(x) : \|x\| \leq t\} \geq M
\]
whenever $t \geq T$. Now consider $s \geq T$. For any $y$ satisfying $\|y\| \geq s$ we have
\[
f(y) = g(y) + h(y) \\
\geq \inf_x \{g(x) : \|x\| \geq \|y\|\} + \inf_x \{h(x) : \|x\| \leq \|y\|\} \\
\geq M,
\]
where the last inequality is implied by $\|y\| \geq T$. Because $y$ was arbitrary, this shows that
\[
\inf_y \{f(y) : \|y\| \geq s\} \geq M,
\]
as desired. \qed

**Theorem A.2.12** (Domination of polynomials by separably coercive functions). Consider $f(x_1, \ldots, x_n) = \sum_{i=1}^n p_i(x_i) + q(x_1, \ldots, x_n)$. Suppose that
(a) each $p_i$ is differentiable on $\mathbb{R}$ and has at most a finite number of critical points;
(b) $q$ is a polynomial on $\mathbb{R}^n$;
(c) for each $i$, the function $s \mapsto p_i(s) - \hat{q} \cdot s^{\deg q}$ is coercive on $\mathbb{R}$, where $\hat{q}$ denotes the sum of the magnitudes of the coefficients in $q$.

Then $f$ is coercive on $\mathbb{R}^n$.

**Proof.** Note that hypothesis (c) implies that each $p_i$ is coercive on $\mathbb{R}$; by continuity, each $p_i$ attains an infimum on any closed interval. Without loss of generality, we may therefore assume that each $p_i$ is nonnegative. Furthermore, by hypothesis (a), we may choose $T > 0$ sufficiently large that no $p_i$ has a critical point $s$ with $|s| > T$. Consequently, for $t > T$, the unique minimizers of each $p_i(s)$ on the intervals $(-\infty, -t]$ and $[t, \infty)$ must occur at $s = -t$ and $s = t$, respectively. Thus, for $t > T$, we have
\[
\inf_s \{p_i(s) : |s| \geq t\} = \min\{p_i(t), p_i(-t)\}. \tag{1}
\]
Now consider any \( x \in \mathbb{R}^n \) for which \( \max_i \{ |x_i| \} \geq t > T \) and choose an index \( j \) giving \( |x_j| \geq t \). Combining the nonnegativity of each \( p_i \) with (1) gives us

\[
\sum_i p_i(x_i) \geq p_j(x_j) \geq \min_i \{p_j(t), p_j(-t)\} \geq \min_i \{\min(p_i(t), p_i(-t))\}.
\]

Because \( x \) is arbitrary, this yields

\[
\inf_x \left\{ \sum_i p_i(x_i) : \max_i \{ |x_i| \} \geq t \right\} \geq \min_i \{\min_i \{\min(p_j(t), p_j(-t))\}\} \quad (2)
\]

whenever \( t > T \).

Next, hypothesis (b) implies that \( q \) can be written as

\[
q(x_1, \ldots, x_n) = \sum_{j=1}^m q_j x_1^{l_1(j)} \cdots x_n^{l_n(j)}.
\]

It is clear that \( q(x) \geq -\sum_j |q_j| t^\deg q = -\hat{q} t^\deg q \) whenever \( \max_i \{ |x_i| \} \leq t \); thus we have

\[
\inf_x \left\{ q(x) : \max_i \{ |x_i| \} \leq t \right\} \geq -\hat{q} t^\deg q
\]

Combining this with (2) and hypothesis (c), we see that \( f \) satisfies the hypotheses of Lemma A.2.11 and must therefore be coercive.

**Corollary A.2.13 (Coercivity of polynomials in many variables).** Consider the function \( f(x_1, \ldots, x_n) = \sum_{i=1}^n p_i(x_i) + q(x_1, \ldots, x_n) \), where \( q \) is a polynomial on \( \mathbb{R}^n \) and each \( p_i \) is a coercive polynomial on \( \mathbb{R} \). If \( \deg q < \min_i \{\deg p_i\} \), then \( f \) is coercive on \( \mathbb{R}^n \).

### A.3 Spectral Theorem for Symmetric Matrices

We now use the existence theory of Chapter 3 to prove that every symmetric real matrix admits an eigenvalue decomposition, as needed for the proof of Proposition 2.3.15(e). Recall that we call \( \lambda \) an *eigenvalue* for \( A \) if \( Ax = \lambda x \) for some vector \( x \neq 0 \).

**Lemma A.3.1 (Existence of eigenvalues).** Every real symmetric matrix has at least one (real) eigenvalue.

**Proof.** Consider \( A = A^T \in \mathbb{R}^{n \times n} \). If \( A = 0 \) we’re done. Assume that \( A \neq 0 \) and define

\[
M = \max_{\|x\|=1} |x \cdot Ax|, \quad N = \max_{\|x\|=1} \|Ax\|.
\]

Note that \( N > 0 \) because \( A \neq 0 \). It is an exercise to show that the definition of \( N \) implies

\[
\|Ax\| \leq N \|x\|.
\]
Combined with the Schwarz inequality $|x \cdot y| \leq \|x\| \|y\|$, this yields $M \leq N$. For arbitrary $x, y \in \mathbb{R}^n$ we have

\[
(x + y) \cdot A(x + y) = x \cdot Ax + 2x \cdot Ay + y \cdot Ay,
\]
\[
(x - y) \cdot A(x - y) = x \cdot Ax - 2x \cdot Ay + y \cdot Ay.
\]

Subtracting these gives

\[
4x \cdot Ay = (x + y) \cdot A(x + y) - (x - y) \cdot A(x - y)
\leq M(\|x + y\|^2 + \|x - y\|^2) = 2M(\|x\|^2 + \|y\|^2),
\]

where the inequality follows directly from the definition of $M$. Consequently, $x \cdot Ay \leq M$ whenever $\|x\| = \|y\| = 1$, so that (by substituting $-y$ for $y$, if necessary) we also have $|x \cdot Ay| \leq M$ whenever $\|x\| = \|y\| = 1$. Taking the maximum over all $y$ with $\|y\| = 1$ gives $\|Ax\| \leq M$ whenever $\|x\| = 1$. Hence we have $N \leq M$; because $M \leq N$, this yields $M = N > 0$. Because the unit sphere is closed and bounded, there exists $x$ with $\|x\| = 1$ and $M = |x \cdot Ax|$. For $\lambda = x \cdot Ax$ we have

\[
0 \leq \|(A - \lambda I)x\|^2 = \|Ax\|^2 - 2\lambda x \cdot Ax + \lambda^2 \|x\|^2
\leq M^2 - 2\lambda^2 + \lambda^2 = \lambda^2 - \lambda^2 = 0.
\]

Therefore $\|(A - \lambda I)x\|^2 = 0$, so $Ax = \lambda x$.

**Theorem A.3.2** (Spectral theorem for real symmetric matrices). Every symmetric matrix $A \in \mathbb{R}^{n \times n}$ can be decomposed as

\[
A = QDQ^T,
\]

where $Q$ is an orthogonal matrix (i.e. $Q^T = Q^{-1}$) and $D$ is a diagonal matrix. Equivalently, $Q^TAQ = D$. Moreover, the columns of $Q$ form an orthonormal set of eigenvectors for $A$ with corresponding eigenvalues given by the diagonal elements of $D$.

**Proof.** We proceed by induction on $n$. The decomposition is immediate for each $1 \times 1$ (symmetric) matrix. Suppose it is also possible for arbitrary $(n - 1) \times (n - 1)$ symmetric matrices, and consider a symmetric matrix $A \in \mathbb{R}^{n \times n}$. By Lemma A.3.1, there exist an eigenvector $x$ of $A$ of unit length and a corresponding (real) eigenvalue $\lambda$. Define

\[
H = I - \frac{2}{v^Tv}vv^T,
\]

where $v = x - \|x\|e_1$ and $e_1$ is the first coordinate vector. It is an exercise to verify that $H$ is a symmetric matrix for which

\[
H^T = H^{-1}, \quad Hx = \|x\|e_1.
\]

Now substitute $x = H\|x\|e_1$ into $Ax = \lambda x$ and rearrange terms to get

\[
(H^TAH)e_1 = \lambda e_1.
\]
Since $H^T A H$ is symmetric, it has the form
\[
H^T A H = \begin{bmatrix}
\beta & b^T \\
b & B
\end{bmatrix},
\]
where $B$ is a symmetric $(n - 1) \times (n - 1)$ matrix. In addition, we must have
\[
(H^T A H)e_1 = \begin{bmatrix}
\beta & b^T \\
b & B
\end{bmatrix}
\begin{bmatrix}
1 \\
0
\end{bmatrix} = \begin{bmatrix}
\lambda \\
0
\end{bmatrix},
\]
which implies that $b = 0$ and $\beta = \lambda$. By the inductive hypothesis there is an orthogonal matrix $R \in \mathbb{R}^{(n-1) \times (n-1)}$ and a diagonal matrix $C$ so that $R^T B R = C$. Now define
\[
Q = H \begin{bmatrix}
1 & 0 \\
0 & R
\end{bmatrix}, \quad D = \begin{bmatrix}
\lambda & 0 \\
0 & C
\end{bmatrix}.
\]
It is readily verified that $Q^T = Q^{-1}$. Also, we see that
\[
Q^T A Q = \begin{bmatrix}
1 & 0 \\
0 & R^T
\end{bmatrix} H^T A H \begin{bmatrix}
1 & 0 \\
0 & R
\end{bmatrix} = \begin{bmatrix}
\lambda & 0 \\
0 & R^T B R
\end{bmatrix} = D.
\]
Therefore the decomposition is valid for $n \times n$ symmetric matrices. The statement about the eigenvectors and eigenvalues follows by rewriting the decomposition as $AQ = QD$, because this says that $[Aq_1, \ldots, Aq_n] = [d_{11}q_1, \ldots, d_{nn}q_n].$

A.4 Matrix View of the Simplex Method

The validity of the simplex method is easily established when viewed from a matrix perspective.

**Definition A.4.1** (Submatrices and basic solutions). Consider an $m \times n$ matrix $A$.

(a) Given set of indices $J \subseteq \{1, \ldots, n\}$, we write $A_J$ to denote the submatrix of $A$ consisting of columns with indices in $J$. Likewise, for a vector $x \in \mathbb{R}^n$ we write $x_J$ to denote the subvector of $x$ with indices in $J$.

(b) We say that $B \subseteq \{1, \ldots, n\}$ is a basis for $A$ if rank($A_B$) = rank($A$) and the columns of $A_B$ are linearly independent.

(c) A solution $x \in \mathbb{R}^n$ of $Ax = b$ is said to be basic if the indices $\{1, \ldots, n\}$ can be partitioned into two sets $B$ and $N$ so that $B$ is a basis for $A$ and $x_N = 0$. In this case, the variables $x_j$ with $j \in B$ are called basic and those with $j \in N$ are nonbasic.

Suppose $B$ is a basis for the matrix $A$. If $N$ is the complement of $B$ in $\{1, \ldots, n\}$, then $Ax = A_N x_N + A_B x_B$ and $c \cdot x = c_N \cdot x_N + c_B \cdot x_B$ for all $x$. Moreover, if $A$ has full
row rank then \( A_B \) is invertible and \( Ax = b \) if and only if \( x_B = A_B^{-1}(b - A_N x_N) \). In this case, we also have \( c \cdot x = (c_N^T - c_B^T A_B A_B^{-1} A_N) x_N + (c_B^T A_B^{-1} b) \).

Now consider a standard-form LP

\[
\begin{align*}
\text{(P)} & \quad \max \{ c \cdot x \mid Ax = b, \ x \geq 0 \}.
\end{align*}
\]

Introducing a variable \( z \) to represent the objective value, we can rewrite this as

\[
\begin{align*}
\max z & \quad \text{s.t.} \quad \begin{bmatrix} A & 0 \\ c^T & -1 \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}, \quad x \geq 0.
\end{align*}
\]

Given a basis \( B \) for \( A \), this becomes

\[
\begin{align*}
\max z & \quad \text{s.t.} \quad \begin{bmatrix} A_N & A_B \\ c_N^T & c_B^T \end{bmatrix} \begin{bmatrix} x_N \\ x_B \\ z \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}, \quad \begin{bmatrix} x_N \\ x_B \end{bmatrix} \geq 0.
\end{align*}
\]

If \( A \) has full row rank, solving for \( x_B \) as in the preceding paragraph amounts to rewriting this as

\[
\begin{align*}
\max z & \quad \text{s.t.} \quad \begin{bmatrix} A_B^{-1} A_N & I & 0 \\ c_N^T & c_B^T A_B A_B^{-1} A_N & 0 & -1 \end{bmatrix} \begin{bmatrix} x_N \\ x_B \\ z \end{bmatrix} = \begin{bmatrix} A_B^{-1} b \\ -c_B^T A_B^{-1} b \end{bmatrix}, \quad \begin{bmatrix} x_N \\ x_B \end{bmatrix} \geq 0.
\end{align*}
\]

In this way, each basis \( B \) for the matrix \( A \) gives an LP of the form

\[
\begin{align*}
\text{(PB)} & \quad \max z & \quad \text{s.t.} \quad \begin{bmatrix} \tilde{A}_N & I & 0 \\ \tilde{c}_N^T & 0 & -1 \end{bmatrix} \begin{bmatrix} x_N \\ x_B \\ z \end{bmatrix} = \begin{bmatrix} \tilde{b} \\ -\alpha \end{bmatrix}, \quad \begin{bmatrix} x_N \\ x_B \end{bmatrix} \geq 0.
\end{align*}
\]

which is equivalent to the original problem \( (P) \). Moreover, each such basis represents a basic solution for \( Ax = \tilde{b} \) given by \((x_N, x_B) = (0, \tilde{b})\) with objective value \( z = \alpha \).

Because \( x_B \) is entirely determined from \( x_N \), any motion from a basic feasible solution to another feasible point necessarily increases \( x_j \) away from zero for some \( j \in N \). In particular, the objective value cannot be improved unless \( \tilde{c}_j > 0 \) for at least one such \( j \in N \). On the other hand, increasing any \( x_j \) may require decreasing \( x_k \) for one or more \( k \in B \): to maintain feasibility, we need to have \( x_B = \tilde{b} - \tilde{A}_N x_N \geq 0 \).

**Method A.4.2** (Ordinary Simplex Method — matrix form). This method solves the standard-form LP \( \max \{ c \cdot x \mid Ax = b, \ x \geq 0 \} \), when started from a feasible basis for the matrix \( A \) of full row rank.

1. Given a feasible basis \( B \), calculate a reformulation of the form \( (PB) \).
2. Choose a nonbasic column \( j \in N \) whose objective coefficient \( \tilde{c}_j \) is positive. If there is none, then the current basic feasible solution is optimal.
3. If there are no positive entries in column $j$ of $\tilde{A}$, then the problem is unbounded: the variable $x_j$ can be increased without bound. Otherwise, calculate the ratios of the entries in $\tilde{b}$ to the corresponding positive entries in column $j$ of $\tilde{A}$. Choose a row $i$ which imposes the tightest restriction (lowest nonnegative ratio) on increasing $x_j$.

4. Put $j$ into $B$ and remove the index of whichever variable was basic in row $i$. Go to step 1.

Observation A.4.3 (Validation of the simplex method). The comments prior to Method A.4.2 explain why the ratio test in step 3 preserves feasibility and why the conclusions of optimality and unboundedness in steps 2 and 3 are valid. The algorithm is therefore well-defined so long as the choice of entering and leaving columns preserves the basis assumption on $B$. To see why this is true, let $j \in N$ and $k \in B$ denote the indices of the entering and leaving columns, let $R = B \setminus \{k\}$ be the columns that remain in the basis, and let $w \in \mathbb{R}^m$ denote the $j$th column of $A = A_B^{-1}A$. In other words, $A_j = A_R w_R + A_k w_k$ and the choice of $j$ and $k$ guarantees that $w_k$ is nonzero. We must show that $[A_R, A_j]$ has full column rank, so consider a vector $z$ with $0 = A_R z_R + A_j z_j$. This gives us $0 = A_R z_R + (A_R w_R + A_k w_k) z_j = A_R (z_R + w_R z_j) + A_k (w_k z_j)$. Because $[A_R, A_k]$ has full column rank, this implies that $z_R + w_R z_j = 0$ and $w_k z_j = 0$; the latter tells us that $z_j = 0$, so the former implies $z_R = 0$. Hence $z = 0$, so replacing $R \cup \{k\}$ by $R \cup \{j\}$ does indeed yield a basis.

As noted in Observation 8.5.3, the simplex method is guaranteed to terminate after finitely many steps if an “anti-cycling” rule (such as Bland’s rule) is employed.
active constraints, 79
anti-cycling rule (in simplex method), 105
attained, 4; also see existence
ball (open, closed), 8
basic optimal solutions (to linear programs), 99, 106
basic solution (for simplex method), 97, 99, 130
basis (of a linear subspace), 62
  existence, 63
Bland’s anti-cycling rule, 104
block-matrix representation of determinants, 120
block-triangular factorization (to determine definiteness), 121
bound
  lower, 3
  upper 4
boundary, 9
bounded
  above, 4
  below, 3
  set, 11
  and coercivity, 31, 35
box (open, closed), 8
Broyden-Fletcher-Goldfarb-Shanno (BFGS), 56
coercive function, 30
chain rule, 7
Cholesky factorization (of positive definite matrices), 22
Clairaut’s theorem, 19
closed
  ball, 8
  box, 8
  set, 9
coercive functions, 31–35
  combining, 124–128
and bounded sets, 31, 35
column space (of a matrix), 62
combining functions
  coercive, 42, 124–128
  convex, 42
complementary slackness, 79
compositions of convex functions, 42
concave function, 37
constraint qualification, 74
continuity, of a convex function, 47
continuous function, 10
continuously differentiable, 6
convergence
  of a sequence, 10
  of Newton’s method, 54
converting general linear programs to simplex standard form, 107
convex combination, 46
convex functions, 37
  combinations, 42
  compositions, 42
[convex functions]
  continuity, 47
  Hessian criterion, 41
  maximizing, 47
  point-wise maximum, 42
  secant criterion, 38
  strictly, 39
  sufficient conditions, 39, 41
  sums, 42
  tangent (gradient) criterion, 40
convex hull, 46
convex inequalities, 38
convex sets, 36
intersections 36

critical point, 13

curve fitting 59
  \( l_1 \)-fit 60
  least-squares fit 60
  minimax fit 60
definite (matrix), 19
degeneracy in linear programming, 99
derivative matrix, 6
descent direction, 49
descent method, 49
determinants
  block-matrix representation, 120
  positive definiteness, 22, 120
diagonal test (for indefiniteness), 23
differentiable, continuously, 6
direction
  descent, 49
  steepest descent, 50
dot product, 5
dual
  of a dual linear program, 92
  of a linear program, 90
duality
  optimality conditions, 94
  weak, 94
    and Karush-Kuhn-Tucker conditions, 94, 116
eigenvalues, 25, 129
epigraph, 37
equality-constrained linear least squares, 68
existence
  by coercivity, 30
  by Weierstrass theorem, 28
in linear least squares, 63
in linear programming, 95
of basis (for a linear subspace), 63
extreme value theorem (one-dimensional), 28
Farkas lemma, 71
Fermat, 13
first-derivative test, sufficiency at endpoints, 15
Gauss-Newton method, 64
global minimizer, 2
global sufficient condition
convexity and critical point, 41
convexity and local minimizer, 39
second derivatives on \( \mathbb{R}^n \), 19
gradient, 6
criterion (for convexity), 40
Hessian, 18
criterion (for convexity), 41
hypograph, 37
increasing, see nondecreasing or strictly increasing
indefinite (matrix), 20
diagonal test, 23
inequality
convex, 38
Jensen, 47
linear, 88
Schwarz, 8
triangle, 8
inexact linesearch, 50
infimum, 3
interior, 9
interior point method, 116–119
intersection of convex sets, 36
Jacobian, 7
Jensen’s inequality, 47
Karush-Kuhn-Tucker
  point, 85
  theorem, 79
  and linear programming duality, 94, 116
kernel (of a matrix), 62
$l_1$-fit, 60
Lagrange multiplier theorem, 72–73
Lagrangian
  function 85
  sufficiency, 86
Lawson-Hanson method, 70
leading principal minor, 22
least squares — see linear least squares or nonlinear least squares
least-squares fit, 60
linear equation, 88
linear function, 88
linear independence, 62
linear inequality, 88
linear least squares, 60
  equality-constrained, 68
  existence, 63
  nonnegative, 69–71
  normal equations, 61
linear program, 88
  converting to simplex standard form, 107–108
  dual, 90
  dual variables, recovering, 106, 109, 112, 115
[linear program]
  existence of solutions, 95
  existence of basic optimal solutions, 100, 106
  Karush-Kuhn-Tucker conditions, 94, 116
  Newton’s method, 117–119
reflexive duality, 92
standard form, 88–89
weak duality, 94
linear subspace, 62
linesearch, 50
inexact, 50
local minimizer, 2
local preservation of definiteness, 123
local preservation of full rank, 124
local sufficient condition, second derivatives on $\mathbb{R}^n$, 26
lower bound, 3
LU factorization (to determine definiteness), 124
matrix transpose, 5
rules, 6
maximizer, 3
global 3
local 3
of a convex function, 47
strict 3
maximum, 4
attained, 4
attained using Weierstrass theorem, 28
maximum, (point-wise) of convex functions, 42
minimax fit, 60
minimizer 2
global 2
local 2, 9
strict 3
minimizing sequence, 4
minimum, 4
attained, 4
attained using Weierstrass theorem, 28
minimum-norm solution (of linear system), 67
monotonicity, 16

\( n \times n \) test for definiteness (using row-reduction), 24

nearest point
  among solutions of a linear system, 67
  in a linear subspace, 64
  in null space of a matrix, 66
  in range of a matrix, 66

necessary condition, 13
  duality-based, 94
  multidimensional, 14
  one-dimensional, 13

negative definite, 19

negative semidefinite, 19

neighborhood, 9

Newton’s method
  convergence, 54

[Newton’s method]
  for linear programming, 117–119
  for minimization, 53
  for systems of equations, 57

nondecreasing, 16

nonlinear least squares, 63

nonnegative linear least squares, 69–71

norm (of a vector), 7
  properties, 8

normal equations, 61

null space (of a matrix), 63

open
  ball, 8
  box, 8
  set, 9

orthogonal vectors, 5

orthogonal complement, 62
of the image of a nullspace, 68
of orthogonal complement, 65
of range and null space, 65
orthogonal projection, 65
Phase 1 (of simplex method), 105, 108, 111, 113
validation, 106
point-wise maximum of convex functions, 42
positive definite, 19
  block-triangular factorization, 121
  characterizations, 120
[positive definite]
  Cholesky factorization , 22
eigenvalues, 25
local preservation, 123
LU factorization, 122
row-reduction test, 24
some additional facts, 25
symmetric factorization, 22
2 × 2 test, 22
positive semidefinite, 19
principal minor, 22
projection matrix, 65
pseudo-inverse, 61
quadratic convergence (of Newton’s Method), 54
quadratic form, 20
quasi-Newton method, 56
range (of a matrix), 62
reflexive duality, 92
regular point, 73, 79
row-reduction test (for definiteness), 24
saddle point, 26
scalar product, 5
Schwarz inequality, 8
secant criterion (for convexity), 38
second-derivative test
    global sufficiency on $\mathbb{R}^n$, 19
    local sufficiency on $\mathbb{R}^n$, 26
[second-derivative test]
    sufficiency on intervals, 16
semidefinite, 19
sequences and closed sets, 11
simplex method
    Bland’s anti-cycling rule, 105
    dual variables, recovering, 103, 106–107, 109–110, 112, 115
    finite termination, 104–105
    matrix form, 131
    Phase 1, 105, 108, 111, 113
    tableau form, 101, 104
    two-phase 107
    validation, 131
    validation of Phase 1, 106
slackness (complementary) 79
span, 62
spectral theorem, 124
standard-form linear program, 88–89
steepest descent direction, 50
steepest descent method, 51
    zig-zag behavior, 52
strictly convex, 39
strictly increasing, 16
strict minimizer, 3
sublevel boundedness (one-dimensional), 28
subsequence, 11
    convergent, see Bolzano-Weierstrass theorem
sufficient condition, 14
    at endpoints, 15
convexity and critical point, 41
convexity and local minimizer, 39
duality-based, 94
Lagrangian, 86
second derivatives on intervals, 16
second derivatives on $\mathbb{R}^n$, 19, 26
sums of convex functions, 42
supremum, 4
symmetric factorization (of positive definite matrices), 22
symmetric matrix, 5
symmetric part (of a matrix), 21
tableau (for simplex method), 101, 104
tangent criterion (for convexity), 40
Taylor’s formula, 5
transpose matrix, 5
rules, 6
triangle inequality, 8
$2 \times 2$ test for definiteness, 22
two-phase simplex method, 107
unbounded
below, 3
above, 4
set, 11
upper bound, 4
valid inequality (for linear program), 90
$v$-restriction (of a function), 18
weak duality, 94
weak linear inequality, 88
Weierstrass theorem, 28
zig-zag behavior (of steepest descent method), 52