# Numerical Methods for Unconstrained Optimization

## an introduction

M.A. Wolfe

*Department of Applied Mathematics*
*University of St. Andrews*

ISBN 0 442 30214 2 cloth
0 442 30217 7 paper



**Published by Van Nostrand Reinhold Company Ltd., Molly Millars Lane, Wokingham, Berkshire, England**

*Published in 1978 by Van Nostrand Reinhold Company A Division of Litton Educational Publishing, Inc., 450 West 33rd Street, New York, N.Y. 10001, U.S.A.*

*Van Nostrand Reinhold Limited 1410 Birchmount Road, Scarborough, Ontario, M1P 2E7, Canada*

*Van Nostrand Reinhold Australia Pty. Limited 17 Queen Street, Mitcham, Victoria 3132, Australia*

**Library of Congress Cataloging in Publication Data**

Wolfe, Michael Anthony.
Numerical methods for unconstrained optimization.

Bibliography: p.
Includes index.
1. Mathematical optimization. 2. Approximation theory. I. Title.
QA402.5.W64 519.4 77-15945
ISBN 0-442-30214-2
ISBN 0-442-30217-7 pbk.

Typeset by Alden Press Oxford, London & Northampton and printed in Great Britain by Henry Ling Ltd., the Dorset Press, Dorchester, Dorset.

# Contents

CHAPTER 1

# Fundamentals

## 1.1 Introduction

This book is an introduction to the study of numerical methods for unconstrained optimization. In order to understand the nature of unconstrained optimization and to gain some idea of its importance in industrial and scientific contexts, we consider four problems. The first problem serves to distinguish constrained optimization from unconstrained optimization and the other three problems are intended to give some idea, however slight, of the many uses to which unconstrained optimizaton can be put.

*Problem 1: A Manufacturing Process*

Suppose that in a factory a chemical is manufactured in bulk by using a given process. The cost of manufacturing unit weight of the chemical depends upon the amount of each constituent raw material required, and upon various parameters associated with the manufacturing process such as temperatures and pressures at various points in the plant. By empirical or theoretical means the cost of manufacturing unit weight of the chemical is expressible as a function $f$ of a number of parameters $x_1, \ldots, x_n$ which are measures of weights of raw materials, temperatures, pressures, and the like. In general, the ranges of values of the parameters $x_1, \ldots, x_n$ are dictated by engineering and economic considerations and cannot be arbitrary; for example, the weights of raw materials required cannot be negative. We say that *constraints* are imposed upon the values of $x_1, \ldots, x_n$. It is clearly of interest to the manufacturer of the chemical to find values of $x_1, \ldots, x_n$, within the limitations imposed by the constraints, for which the cost $f(x_1, \ldots, x_n)$ of manufacturing unit weight of the chemical has the smallest possible value. He has therefore to solve the following optimization problem.

Find values $x_1^*, \ldots, x_n^*$ of $x_1, \ldots, x_n$, which satisfy a given set of constraints, such that $f(x_1^*, \ldots, x_n^*) \leqslant f(x_1, \ldots, x_n)$ for all values of $x_1, \ldots, x_n$ which satisfy the constraints; that is, *minimize* $f$ subject to the given set of constraints.

The function $f$ to be minimized is called the *objective* function. Typical constraints are of the form

$$a_i \leqslant x_i \leqslant b_i \quad (i = 1, \ldots, n) \tag{1.1.1}$$

where the $a_i$ and the $b_i$ are given real numbers.

More generally, a problem in constrained optimization is of the following form.

Problem 1

Minimize

$$f(x_1, \ldots, x_n) \tag{1.1.2a}$$

Subject to

$$\phi_i(x_1, \ldots, x_n) = 0 \quad (i = 1, \ldots, l) \tag{1.1.2b}$$

$$\psi_j(x_1, \ldots, x_n) \leqslant 0 \quad (j = 1, \ldots, m) \tag{1.1.2c}$$

$$(x_1, \ldots, x_n) \subset D \subset R^n \tag{1.1.2d}$$

in which $\phi_i$ and $\psi_j$ are given real functions of $n$ real variables, where $l \leqslant n$, $R^n$ is the

set of all $n$-tuples of real numbers, and $D$ is a subset of $R^n$, such as the subset defined by (1.1.1) □

It is convenient to introduce vector notation in which $\mathbf{x}$, $\boldsymbol{\varphi}(\mathbf{x})$, and $\boldsymbol{\psi}(\mathbf{x})$ are column vectors defined by

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad \boldsymbol{\varphi}(\mathbf{x}) = \begin{bmatrix} \phi_1(\mathbf{x}) \\ \vdots \\ \phi_l(\mathbf{x}) \end{bmatrix} \quad \boldsymbol{\psi}(\mathbf{x}) = \begin{bmatrix} \psi_1(\mathbf{x}) \\ \vdots \\ \psi_m(\mathbf{x}) \end{bmatrix} \tag{1.1.3a}$$

or

$$\mathbf{x} = [x_1, \ldots, x_n]^T \quad \boldsymbol{\varphi}(\mathbf{x}) = [\phi_1(\mathbf{x}), \ldots, \phi_l(\mathbf{x})]^T$$
$$\boldsymbol{\psi}(\mathbf{x}) = [\psi_1(\mathbf{x}), \ldots, \psi_m(\mathbf{x})]^T \tag{1.1.3b}$$

It is also convenient to introduce vector inequalities. Let

$$\mathbf{a} = [a_1, \ldots, a_n]^T \quad \mathbf{b} = [b_1, \ldots, b_n]^T$$

Then we write $\mathbf{a} \geqslant \mathbf{b}$ if and only if $a_i \geqslant b_i$ $(i = 1, \ldots, n)$. By using the vector notation we may express the general constrained optimization problem as follows.

Problem 1

Minimize

$$f(\mathbf{x}) \tag{1.1.4a}$$

Subject to

$$\boldsymbol{\varphi}(\mathbf{x}) = \mathbf{0} \tag{1.1.4b}$$

$$\boldsymbol{\psi}(\mathbf{x}) \leqslant \mathbf{0} \tag{1.1.4c}$$

$$\mathbf{x} \in D \subset R^n \tag{1.1.4d}$$

□

The constraints (1.1.2b) or (1.1.4b) are *equality* constraints, while the constraints (1.1.2c) or (1.1.4c) are *inequality* constraints, and (1.1.2d) or (1.1.4d) are *set* constraints. Any point $\mathbf{x}$ which satisfies (1.1.2b), (1.1.2c), and (1.1.2d) or (1.1.4b), (1.1.4c), and (1.1.4d) is called a *feasible* point.

The problem (1.1.2) or (1.1.4) is usually referred to as the *general nonlinear programming problem*. An example is as follows.

Minimize

$$-[25 - (x_1 - 5)^2 - (x_2 - 5)^2]^{1/2}$$

Subject to

$$5x_2 - 4x_1 \leqslant 0$$

$$5x_2 + 4x_1 - 40 \leqslant 0$$

$$x_1 \geqslant 0, \quad x_2 \geqslant 0$$

This problem is trivial in that it can be solved by noticing that geometrically the surface

$$(x_1 - 5)^2 + (x_2 - 5)^2 + y^2 = 25$$

is a sphere with centre $(5, 5, 0)$ and radius 5 and that the surfaces

$$5x_2 - 4x_1 = 0$$
$$5x_2 + 4x_1 - 40 = 0$$

are planes parallel to the $y$-axis which intersect the surface of the sphere and each other at the points $(5, 4, \sqrt{24})$ and $(5, 4, -\sqrt{24})$. Hence the minimum value of the objective function which satisfies the given constraints is $-\sqrt{24}$ and this occurs at $x_1^* = 5, x_2^* = 4$.

The nonlinear programming problems which are encountered in industrial, economic, administrative and scientific contexts are seldom as trivial as the preceding problem, and cannot usually be solved analytically or geometrically. A number of numerical methods exist, however, for solving nonlinear programming problems but a satisfactory account of them is beyond the scope of this book.

*Problem 2: Fitting Data*

Suppose that some variable $y$ is thought to depend upon a variable $t$ through a formula of the form

$$y = Y(t; \mathbf{x}) \tag{1.1.5}$$

where $Y: R^1 \times R^n \to R^1$ is a known function of $t$ and of a vector $\mathbf{x}$ of $n$ parameters. (To denote that $f$ is a real-valued function of $n$ real variables $x_1, \ldots, x_n$ we write $f: R^n \to R^1$, and to denote that $\boldsymbol{\varphi}$ is a vector-valued function with $l$ components, of the $n$ real variables $x_1, \ldots, x_n$ we write $\boldsymbol{\varphi}: R^n \to R^l$. If $\mathbf{F}$ is a vector-valued function with $n$ real components, of both a vector with $l$ real components and a vector with $m$ real components, we write $\mathbf{F}: R^l \times R^m \to R^n$.) This situation arises frequently when a physically observable quantity $y$ is predicted on theoretical grounds to depend upon a variable $t$ in a certain way, but the theory, being incomplete, gives only the functional form of the dependence, which contains a number of parameters, the values of which are unknown. For example, if the length $y$ of a rod of copper is regarded as a function of its temperature $t$, then a satisfactory formula for $y$ is

$$y = Y(t; \mathbf{x}) = y_0(1 + x_1 t + x_2 t^2) \tag{1.1.6}$$

where $y_0$, which is supposed to be known, is the length of the rod when $t = 0$ and $x_1, x_2$, are parameters which are supposed to be unknown. If a number of measurements of $y$ were made corresponding to an equal number of known values of $t$ then the formula (1.1.6) could be 'fitted' to the experimental data consisting of the values of $y$ and the corresponding values of $t$, and the values of $x_1$ and $x_2$ which ensure the 'best fit' to the data could be obtained.

In the general case corresponding to (1.1.5) suppose that for $m$ values $t_1, \ldots, t_m$ of $t$, measurements $y_1, \ldots, y_m$ of $y$ are made. In practice, even if the formula (1.1.5) exactly represents the physical situation, the measured values of $y$ are usually subject to experimental errors, so that a value $\mathbf{x}^*$ of $\mathbf{x}$ such that $Y(t_i; \mathbf{x}^*)$ is exactly equal to the measured value $y_i$ of $y$ for all values of $i$ ($i = 1, \ldots, m$) is unlikely to exist. If, however, we take $m \geqslant n$ and let $f: R^n \to R^1$ be defined by

$$f(\mathbf{x}) = \sum_{i=1}^{m} [Y(t_i; \mathbf{x}) - y_i]^2 \quad (\mathbf{x} \in R^n) \tag{1.1.7}$$

then we should expect to be able to 'fit' (1.1.5) to the experimental data $y_i$ ($i = 1, \ldots, m$) by minimizing the objective function $f$. For obvious reasons, a value $\mathbf{x}^*$ of $\mathbf{x}$ obtained by minimizing $f$ defined by (1.1.7) is said to give a *least squares* fit to the experimental data. The general nonlinear least squares problem is as follows.

Problem 2.

Minimize

$f(\mathbf{x})$

Subject to

$\mathbf{x} \in R^n$

where $f: R^n \to R^1$ is defined by (1.1.7) □

Problem 2 is a special case of the general unconstrained optimization problem, which has the form

Minimize

$$f(\mathbf{x}) \tag{1.1.8a}$$

Subject to

$$\mathbf{x} \in R^n \tag{1.1.8b}$$

where $f: R^n \to R^1$ is a given objective function. □

The least squares problem corresponding to (1.1.6) is linear because $Y$ is a linear function of $x_1$ and $x_2$. In many least squares problems of practical interest $Y$ is not linear in $\mathbf{x}$ and the corresponding least squares problem cannot usually be solved analytically. A large number of numerical methods for solving such problems exist, however, and some of these are described in Chapter 7.

*Problem 3: Solving Differential Equations*

Many applications of mathematics to physical problems result in the need to solve differential or integral equations. The following mathematical problem is typical.

Problem 3

Solve the boundary value problem

$$F(t, y, y^{(1)}, \ldots, y^{(r)}) = 0 \tag{1.1.9a}$$

where $F: [a, b] \times R^1 \times \ldots \times R^1 \to R^1$ is a given function and

$$y^{(k)}(t) = \frac{\mathrm{d}^k}{\mathrm{d}t^k} y(t) \quad (k = 1, \ldots, r)$$

subject to a given set of $r$ boundary conditions chosen from

$$y(a) = A \quad y(b) = B \tag{1.1.9b}$$

$$y^{(k)}(a) = A_k \quad y^{(k)}(b) = B_k \quad (k = 1, \ldots, r) \quad \square \tag{1.1.9c}$$

Suppose that $\phi_k: [a, b] \in R^1 \to R^1$ $(k = 1, 2, \ldots)$ are given functions which are $r$ times continuously differentiable on $[a, b]$. If the $\phi_k$ are suitable we could attempt to find a solution $Y(t; \mathbf{x}^*)$ of the boundary value problem (1.1.9) by defining $Y(t; \mathbf{x})$ by

$$Y(t; \mathbf{x}) = \sum_{k=1}^{n} x_k \phi_k(t) \quad (a \leqslant t \leqslant b) \tag{1.1.10}$$

for some integer $n \geqslant 1$ in such a way that $Y(t; \mathbf{x})$ satisfies the given set of $r$ boundary conditions selected from (1.1.9b) and (1.1.9c) for all values of $\mathbf{x}$, and then determining $\mathbf{x}^*$ by requiring that

$$F\left(t, Y(t, \mathbf{x}^*), \ldots, \frac{\mathrm{d}^r}{\mathrm{d}t^r} Y(t; \mathbf{x}^*)\right)$$

be in some sense as close as possible to zero throughout $[a, b]$. For example, we could choose $a = t_1 \leqslant t_2 \leqslant \ldots \leqslant t_s = b$ $(s \geqslant n)$ and define $f: R^n \to R^1$ by

$$f(\mathbf{x}) = \sum_{i=1}^{s} \left[ F\left( t_i, Y(t_i; \mathbf{x}), \frac{\mathrm{d}}{\mathrm{d}t} Y(t_i; \mathbf{x}), \ldots, \frac{\mathrm{d}^r}{\mathrm{d}t^r} Y(t_i; \mathbf{x}) \right) \right]^2 \tag{1.1.11}$$

Then $\mathbf{x}^* \in R^n$ could be determined by minimizing $f$ without constraints since it is supposed that we have chosen the $\phi_k$ so that the $r$ boundary conditions are satisfied for all values of $\mathbf{x}$.

If the boundary value problem (1.1.9) has a solution and if the $\phi_k$ form a complete set it is reasonable to suppose that as $n$ is increased, $Y(t; \mathbf{x})$ becomes in some sense closer to the solution. Kowalik and Osborne (1968) mention a specific problem of this sort. By using this approach, integral equations and partial differential equations may be solved in principle.

*Problem 4: The Calculus of Variations*

Let $u: [a, b] \subset R^1 \to R^1$ be continuously differentiable on $[a, b]$ and let $F: [a, b] \times R^1 \times R^1 \to R^1$ be such that $F(t, u(t), u^{(1)}(t))$ is a continuous function of $t$ on $[a, b]$. Let

$$I(u) = \int_a^b F(t, u(t), u^{(1)}(t))\, \mathrm{d}t \tag{1.1.12}$$

and let $S$ be the set of functions $u$ which are continuously differentiable on $[a, b]$ and for which $u(a) = A$, $u(b) = B$, where $A$ and $B$ are given. Then the solution of the following problem is of great importance in Physics and in Engineering.

Problem 4

Find $u^* \in S$ such that $I(u^*) \leqslant I(u)$ $(u \in S)$, where $I$ is defined by (1.1.12) □

Problem 4 is the *basic problem* of variational calculus. It can be shown that the function $u$ satisfies a certain differential equation subject to the boundary conditions $u(a) = A$, $u(b) = B$, but in general the differential equation is nonlinear. It is therefore of interest to consider a more direct approach to the solution of Problem 4.

Let $u_1, \ldots, u_n$ be $n$ continuously differentiable functions on $[a, b]$ and suppose that $u_i(a) = u_i(b) = 0$ $(i = 1, \ldots, n)$. Suppose also that $u_1, \ldots, u_n$ are linearly independent so that

$$\sum_{k=1}^{n} \alpha_k u_k(t) = 0 \quad (a \leqslant t \leqslant b)$$

implies that $\alpha_k = 0$ $(k = 1, \ldots, n)$. Let

$$\phi(t; \mathbf{x}) = \sum_{k=1}^{n} x_k u_k(t) + \frac{(A - B)}{(a - b)} t + \frac{(aB - Ab)}{(a - b)} \quad (a \leqslant t \leqslant b) \tag{1.1.13}$$

Then

$$\phi(a; \mathbf{x}) = A \quad \phi(b; \mathbf{x}) = B$$

Let

$$f(\mathbf{x}) = I(\phi(t; \mathbf{x}))$$

If we can obtain $\mathbf{x}^*$ such that $f(\mathbf{x}^*) \leqslant f(\mathbf{x})$ $(\mathbf{x} \in R^n)$ then we have an approximation to $u^*$ given by

$$u_n^*(t) = \phi(t; x^*)$$

As $n$ increases it is reasonable to suppose that $u_n^*$ is in some sense an increasingly good approximation to $u^*$.

If $I(u)$ cannot be evaluated analytically then it may be estimated by using a numerical quadrature formula so that Problem 4 may still be solved in principle by minimizing $f$. Burley (1974) has given several elementary examples of problems similar to Problems 3 and 4.

The three applications of unconstrained optimization considered in this section illustrate only a small part of the whole field of applications, but it is hoped that the reader will gain from studying them a little insight into the nature of unconstrained optimization and its great importance in industrial and scientific contexts. A description of some applications of nonlinear programming to a wide diversity of problems has been given by Bracken and McCormick (1968). See also Dixon (1976).

In all of the problems considered in this section, no efficient general analytical methods for minimizing the objective functions exist. A great deal of effort has, however, been expended since 1960 in constructing efficient numerical methods for solving unconstrained and constrained minimization problems. Many of these methods are easy to implement on a computer using a high level language such as FORTRAN or ALGOL and much can be said about their behaviour without resorting to advanced mathematical analysis. Certain numerical methods for optimization have proved repeatedly to be not only robust but also efficient; that is to say, they seldom fail to obtain a sufficiently accurate estimate of the solution of a given optimization problem and do so with tolerably few evaluations of the objective function and of its derivatives if the latter are required. This is not to say, however, that these methods are universally applicable. There is no panacea for all optimization problems.

In the subsequent chapters of this book some methods for unconstrained optimization are described in sufficient detail to permit them to be implemented on a computer. Computer programs are not given, because many have been published, and some of these are readily available from sources which are given where appropriate in the text. The reader is, however, urged to write programs and to run them in order to gain that insight into the methods which are described in this book which can be acquired only from numerical experience.

## 1.2 Fundamental Linear Algebra

In this section some fundamental ideas from linear algebra which are used subsequently are collected, and some notation is described. For greater detail than it is possible to give here the reader should consult, for example, Hadley (1969), Noble (1969), or Franklin (1968). It is not necessary for the reader to understand the content of this section in detail if it is required only to *use* the methods which are described in subsequent chapters, but an understanding of this section is, however, necessary to appreciate the theory which underlies these methods.

By $R^n$ is meant the set of $n \times 1$ vectors $\mathbf{x} = [x_1, \ldots, x_n]^T$ in which $x_i \in R^1$ $(i = 1, \ldots, n)$ where $R^1$ is the set of real numbers. If we admit the usual rules for adding $n \times 1$ vectors $\mathbf{x}, \mathbf{y} \in R^n$ and multiplying them by real numbers $\alpha$, namely

$$\mathbf{x} + \mathbf{y} = [x_1 + y_1, \ldots, x_n + y_n]^T \tag{1.2.1}$$

and

$$\alpha\mathbf{x} = [\alpha x_1, \ldots, \alpha x_n]^T \tag{1.2.2}$$

then $R^n$ together with the operations defined by (1.2.1) and (1.2.2) constitute a linear space. We shall subsequently refer to the linear space as $R^n$. A function $\|.\|: R^n \to R^1$ such that

$$\|\mathbf{x}\| \geqslant 0, \quad (\forall \mathbf{x} \in R^n)$$

$$(\|\mathbf{x}\| = 0) \Leftrightarrow (\mathbf{x} = \mathbf{0})$$

$$\|\alpha\mathbf{x}\| = |\alpha| \, \|\mathbf{x}\| \quad (\forall \mathbf{x} \in R^n; \forall \alpha \in R^1)$$

and

$$\|\mathbf{x} + \mathbf{y}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\| \quad (\forall \mathbf{x}, \mathbf{y} \in R^n)$$

is called a *vector norm* on $R^n$. Commonly used vector norms on $R^n$ are the *maximum* norm $\|.\|_M$ defined by

$$\|\mathbf{x}\|_M = \max_{1 \leqslant i \leqslant n} \{|x_i|\} \tag{1.2.3}$$

the *taxicab* norm $\|.\|_T$ defined by

$$\|\mathbf{x}\|_T = \sum_{i=1}^{n} |x_i| \tag{1.2.4}$$

and the *Euclidean* norm $\|.\|_E$ defined by

$$\|\mathbf{x}\|_E = \left( \sum_{i=1}^{n} x_i^2 \right)^{1/2} \tag{1.2.5}$$

where $\mathbf{x} = [x_1, \ldots, x_n]^T$

It can be shown that all vector norms on $R^n$ are equivalent in the sense that if $\|.\|_A$ and $\|.\|_B$ are any two such norms then there exist $\alpha, \beta \in (0, \infty)$ such that

$$\alpha\|\mathbf{x}\|_A \leqslant \|\mathbf{x}\|_B \leqslant \beta\|\mathbf{x}\|_A \quad (\forall \mathbf{x} \in R^n)$$

The linear space $R^n$ together with a vector norm $\|.\|$ on $R^n$ is called a *normed linear space*; we denote it by $\{R^n, \|.\|\}$.

The *scalar product* $\mathbf{x}^T\mathbf{y}$ of the vectors $\mathbf{x}, \mathbf{y} \in R^n$ is defined by

$$\mathbf{x}^T\mathbf{y} = \sum_{i=1}^{n} x_i y_i \tag{1.2.6}$$

In the normed linear space $\{R^n, \|.\|_E\}$ we have Schwarz' inequality, namely

$$|\mathbf{x}^T\mathbf{y}| \leqslant \|\mathbf{x}\|_E \, \|\mathbf{y}\|_E \quad (\forall \mathbf{x}, \mathbf{y} \in R^n) \tag{1.2.7}$$

with equality if an only if there exists $\lambda \in R^1$ such that $\mathbf{y} = \lambda\mathbf{x}$.

The vectors $\mathbf{x}, \mathbf{y} \in R^n$ are *orthogonal* if and only if

$$\mathbf{x}^T\mathbf{y} = 0 \tag{1.2.8}$$

and the vector $\mathbf{x} \in R^n$ is *normalized* (to unity) if and only if

$$\mathbf{x}^T\mathbf{x} = 1 \tag{1.2.9}$$

The set of vectors $\{\mathbf{x}_i \in R^n \mid i = 1, \ldots, m; m \leqslant n\}$ is an *orthonormal* set if and only if

$$\mathbf{x}_i^T\mathbf{x}_j = \delta_{ij} \quad (i, j = 1, \ldots, m) \tag{1.2.10}$$

where the Kronecker delta $\delta_{ij}$ is defined by

$$\delta_{ij} = \begin{cases} 1 \ (i = j) \\ 0 \ (i \neq j) \end{cases} \tag{1.2.11}$$

The set of vectors $\{\mathbf{x}_i \in R^n \mid i = 1, \ldots, m; m \leqslant n\}$ is a *linearly independent* set if and only if

$$\sum_{i=1}^{m} c_i \mathbf{x}_i = \mathbf{0} \tag{1.2.12}$$

implies that $c_i = 0$ $(i = 1, \ldots, m)$, and is *linearly dependent* otherwise. A set of vectors $\{\mathbf{u}_i \in R^n \mid i = 1, \ldots, m; m \geqslant n\}$ *spans* or *generates* the linear space $R^n$ if an only if every vector $\mathbf{x} \in R^n$ is expressible in the form

$$\mathbf{x} = \sum_{i=1}^{m} c_i \mathbf{u}_i \tag{1.2.13}$$

A *basis* for the linear space $R^n$ is a linearly independent set of vectors in $R^n$ which spans $R^n$. The set of vectors $\{\mathbf{e}_i \in R^n \mid i = 1, \ldots, n\}$ where

$$\mathbf{e}_i = [0, \ldots, 0, \overset{(i)}{1}, 0, \ldots, 0]^T \quad (i = 1, \ldots, n) \tag{1.2.14}$$

in which $\mathbf{e}_i$ has all components equal to zero save the $i$th which is equal to unity, clearly forms a basis for $R^n$ since if $\mathbf{x} = [x_1, \ldots, x_n]^T$ then

$$\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{e}_i \tag{1.2.15}$$

It can be shown that every basis for $R^n$ contains exactly $n$ vectors, and that every set of $n$ orthogonal vectors in $R^n$ excluding $\mathbf{0}$ forms a basis for $R^n$. An *orthonormal* basis for $R^n$ is a set of $n$ orthonormal vectors. For example, $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ defined by (1.2.14) is an orthonormal basis for $R^n$.

Given any basis for $R^n$, an orthonormal basis for $R^n$ may be constructed by using the Gram–Schmidt procedure. Let $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ be a basis for $R^n$ and let $\mathbf{v}_i$, $\mathbf{w}_i$ $(i = 1, \ldots, n)$ be generated recursively from

$$\mathbf{v}_1 = \mathbf{u}_1 \quad \mathbf{w}_1 = \frac{\mathbf{v}_1}{(\mathbf{v}_1^T \mathbf{v}_1)^{1/2}} \tag{1.2.16a}$$

$$\mathbf{v}_i = \mathbf{u}_i - \sum_{k=1}^{i-1} (\mathbf{w}_k^T \mathbf{u}_k)\mathbf{w}_k \quad \mathbf{w}_i = \frac{\mathbf{v}_i}{(\mathbf{v}_i^T \mathbf{v}_i)^{1/2}} \quad (i = 2, \ldots, n) \tag{1.2.16b}$$

Then it is easy to show (Exercise 1.2.7) that

$$\mathbf{w}_i^T \mathbf{w}_j = \delta_{ij} \quad (i, j = 1, \ldots, n) \tag{1.2.17}$$

so that $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$ is an orthonormal basis for $R^n$.

A set $S \subset R^n$ is a *subspace* of the linear space $R^n$ if and only if for each pair $\mathbf{x}'$, $\mathbf{x}''$ of vectors in $S$, the vector $\alpha\mathbf{x}' + \beta\mathbf{x}''$ is also in $S$ where $\alpha, \beta$ are any real numbers. As with $R^n$, the dimensionality $m$ of a subspace $S$ of $R^n$ is the number of vectors in any basis for $S$. Clearly $m \leqslant n$. Let $\{\mathbf{u}_i \in R^n \mid i = 1, \ldots, m\}$ be a basis for the $m$-dimensional subspace $S$ of $R^n$. Then clearly an orthonormal basis for $S$ can be constructed by using the Gram–Schmidt procedure (1.2.16), where $i$ runs from 2 to $m$.

Let $S$ be a subspace of $R^n$ and let $\hat{\mathbf{x}} \in R^n$ be given. Then the set $V$ defined by

$$V = \{\mathbf{v} \in R^n | \mathbf{v} = \hat{\mathbf{x}} + \mathbf{y}, \mathbf{y} \in S\}$$

is called a *linear variety*. It is convenient to use the notation

$$V = \hat{\mathbf{x}} + S$$

In conformity with the notation used to represent column vectors, the *transpose* $\mathbf{A}^T$ of the $m \times n$ matrix $\mathbf{A}$ with elements $a_{ij}$ $(i = 1, \ldots, m; j = 1, \ldots, n)$ is the $n \times m$ matrix $\mathbf{B}$ with elements $b_{ij}$ $(i = 1, \ldots, n; j = 1, \ldots, m)$ such that

$$b_{ij} = a_{ji} \quad (i = 1, \ldots, n; j = 1, \ldots, m)$$

An $n \times n$ matrix $\mathbf{A}$ is *symmetric* if and only if

$$\mathbf{A}^T = \mathbf{A} \tag{1.2.18}$$

and is *skew-symmetric* or *anti-symmetric* if and only if

$$\mathbf{A}^T = -\mathbf{A} \tag{1.2.19}$$

Any $n \times n$ matrix $\mathbf{A}$ may be expressed as the sum of a symmetric matrix $\mathbf{B}$ and a skew-symmetric matrix $\mathbf{C}$ where

$$\mathbf{B} = \tfrac{1}{2}(\mathbf{A} + \mathbf{A}^T) \quad \mathbf{C} = \tfrac{1}{2}(\mathbf{A} - \mathbf{A}^T) \tag{1.2.20}$$

If $\mathbf{A}$ and $\mathbf{B}$ are any $n \times n$ matrices, then it follows from the definitions of matrix multiplication and transpose that

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T \tag{1.2.21}$$

We write

$$\mathbf{A} = [a_{ij}] \quad (i = 1, \ldots, m; j = 1, \ldots, n)$$

to denote that $\mathbf{A}$ is an $m \times n$ matrix with elements $a_{ij}$, and we write

$$\mathbf{A} = \text{Diag}(a_{11}, \ldots, a_{nn})$$

to denote that $\mathbf{A}$ is the diagonal $n \times n$ matrix with diagonal elements $a_{ii}$ $(i = 1, \ldots, n)$.

It is assumed that the reader is familiar with the more elementary properties of the determinant, $\text{Det}(\mathbf{A})$, of an $n \times n$ matrix $\mathbf{A}$, and with the elementary properties of the inverse $\mathbf{A}^{-1}$ of $\mathbf{A}$. In particular, if $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ matrices then

$$\text{Det}(\mathbf{A}^T) = \text{Det}(\mathbf{A}) \tag{1.2.22}$$

and

$$\text{Det}(\mathbf{AB}) = \text{Det}(\mathbf{A})\,\text{Det}(\mathbf{B}) \tag{1.2.23}$$

Also, $\mathbf{A}$ is singular, so that $\mathbf{A}^{-1}$ does not exist, if and only if $\text{Det}(\mathbf{A}) = 0$, and if $\mathbf{A}$ and $\mathbf{B}$ are nonsingular then

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1} \tag{1.2.24}$$

It is also assumed that the reader is familiar with the idea of *partitioning* of matrices. For example, if $\mathbf{A}$ is an $m \times n$ matrix with elements $a_{ij}$ $(i = 1, \ldots, m; j = 1, \ldots, n)$ and columns $\mathbf{a}_j$ $(j = 1, \ldots, n)$ then we write

$$\mathbf{A} = \left[\mathbf{a}_1 \vdots \cdots \vdots \mathbf{a}_n\right]$$

to denote the partitioning of $\mathbf{A}$ into columns. More generally, to denote partitioning of

$\mathbf{A}$ into submatrices $\mathbf{A}_{ij}$ $(i = 1, \ldots, p \leqslant m; j = 1, \ldots, q \leqslant n)$ we use a similar notation. For example, if $p = 2$ and $q = 3$ we write

$$\mathbf{A} = \left[\begin{array}{c|c|c} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{A}_{13} \\ \hline \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{A}_{23} \end{array}\right]$$

A real symmetric $n \times n$ matrix $\mathbf{A}$ has $n$ orthonormal eigenvectors $\mathbf{u}_i$ $(i = 1, \ldots, n)$ corresponding to $n$ real eigenvalues $\lambda_i$ $(i = 1, \ldots, n)$ (which need not be distinct) so that

$$\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i \quad (i = 1, \ldots, n) \tag{1.2.25}$$

and

$$\mathbf{u}_i^T \mathbf{u}_j = \delta_{ij} \quad (i, j = 1, \ldots, n) \tag{1.2.26}$$

An $n \times n$ matrix $\mathbf{B}$ is *orthogonal* if and only if

$$\mathbf{B}\mathbf{B}^T = \mathbf{B}^T\mathbf{B} = \mathbf{I} \tag{1.2.27}$$

It is easy to see that if $\mathbf{A}$ is an $n \times n$ real symmetric matrix then there exists an orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}^T\mathbf{A}\mathbf{P}$ is a diagonal matrix. Indeed, $\mathbf{P}$ is the matrix with columns $\mathbf{u}_1, \ldots, \mathbf{u}_n$ where the $\mathbf{u}_i$ are the normalized eigenvectors of $\mathbf{A}$ which satisfy (1.2.25) and (1.2.26).

Let $\mathbf{A}$ be an $n \times n$ matrix and let $\mathbf{x}$ be an $n \times 1$ vector. Then, irrespective of whether $\mathbf{A}$ is symmetric,

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i a_{ij} x_j$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} x_i a_{ji} x_j$$

so that

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\mathbf{B}\mathbf{x}$$

where the symmetric matrix $\mathbf{B}$ is defined in (1.2.20). Therefore when considering quadratic forms $\mathbf{x}^T\mathbf{A}\mathbf{x}$ we may always suppose that the corresponding matrix $\mathbf{A}$ is symmetric.

An $n \times n$ matrix $\mathbf{A}$ is *positive definite* if and only if

$$\mathbf{x}^T\mathbf{A}\mathbf{x} > 0 \quad (\forall \mathbf{x} \in R^n, \mathbf{x} \neq \mathbf{0}) \tag{1.2.28a}$$

and is *positive semi-definite* if and only if

$$\mathbf{x}^T\mathbf{A}\mathbf{x} \geqslant 0 \quad (\forall \mathbf{x} \in R^n) \tag{1.2.29a}$$

Similarly $\mathbf{A}$ is *negative* definite if and only if

$$\mathbf{x}^T\mathbf{A}\mathbf{x} < 0 \quad (\forall \mathbf{x} \in R^n, \mathbf{x} \neq \mathbf{0}) \tag{1.2.28b}$$

and $\mathbf{A}$ is *negative semi-definite* if and only if

$$\mathbf{x}^T\mathbf{A}\mathbf{x} \leqslant 0 \quad (\forall \mathbf{x} \in R^n) \tag{1.2.29b}$$

Setting $\mathbf{x} = \mathbf{e}_i$ where $\mathbf{e}_i$ is defined by (1.2.14), it follows immediately from (1.2.28a) that if $\mathbf{A}$ is positive definite then $a_{jj} > 0$ $(i = 1, \ldots, n)$. Similar results hold when $\mathbf{A}$ is positive semi-definite, negative semi-definite, and negative definite, respectively.

It can be shown that if **A** is symmetric positive definite, then there exists a symmetric positive definite matrix **M** such that

$$\mathbf{M}^2 = \mathbf{A} \tag{1.2.30}$$

Furthermore, a matrix **A** is positive definite if and only if all of its eigenvalues are positive.

Let **B** be an $m \times n$ matrix. Then the *rank* of **B**, which we write $r(\mathbf{B})$, is the maximum number of linearly independent columns of **B**. The $n \times n$ matrix **A** is nonsingular if and only if $r(\mathbf{A}) = n$. Hence if $r(\mathbf{A}) = n$, then $\mathbf{Au} = \mathbf{0}$ if and only if $\mathbf{u} = \mathbf{0}$. Furthermore if **C** is an $m \times n$ real matrix with $r(\mathbf{C}) = n$ then $\mathbf{Cu} = \mathbf{0}$ if and only if $\mathbf{u} = \mathbf{0}$, while if $r(\mathbf{C}) < n$, then $\mathbf{Cu} = \mathbf{0}$ has a nonzero solution. If $m < n$ then necessarily $r(\mathbf{C}) < n$ and so $\mathbf{Cu} = \mathbf{0}$ always has a nonzero solution if $m < n$. Finally in connection with the concept of rank, we note that if **A** and **B** are any two matrices such that the product **AB** exists, then

$$r(\mathbf{AB}) \leqslant \min\{r(\mathbf{A}), r(\mathbf{B})\} \tag{1.2.31}$$

Let $\{R^n, \|.\|\}$ be a normed linear space, and let **A** be any $n \times n$ matrix. Then the *matrix norm* $\|.\|$ corresponding to the vector norm $\|.\|$ is defined by

$$\|\mathbf{A}\| = \max_{\mathbf{x} \neq \mathbf{0}} \frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|} \tag{1.2.32}$$

which implies that

$$\|\mathbf{Ax}\| \leqslant \|\mathbf{A}\| \, \|\mathbf{x}\| \quad (\forall \mathbf{x} \in R^n) \tag{1.2.33}$$

and that if **A** and **B** are any $n \times n$ matrices then

$$\|\mathbf{AB}\| \leqslant \|\mathbf{A}\| \, \|\mathbf{B}\| \tag{1.2.34}$$

Note that the same symbol $\|.\|$ is used for a vector norm and its corresponding matrix norm, but in practice no confusion results because which norm is intended is always clear from the context in which it appears.

The matrix norms corresponding to the vector norms $\|.\|_{\mathbf{M}}$, $\|.\|_{\mathbf{T}}$, and $\|.\|_{\mathbf{E}}$ are given by

$$\|\mathbf{A}\|_{\mathbf{M}} = \max_{1 \leqslant i \leqslant n} \left\{ \sum_{j=1}^{n} |a_{ij}| \right\} \tag{1.2.35}$$

$$\|\mathbf{A}\|_{\mathbf{T}} = \max_{1 \leqslant j \leqslant n} \left\{ \sum_{i=1}^{n} |a_{ij}| \right\} \tag{1.2.36}$$

$$\|\mathbf{A}\|_{\mathbf{E}} = \left( \max_{1 \leqslant i \leqslant n} \{\lambda_i\} \right)^{1/2} \tag{1.2.37}$$

in which **A** is an $n \times n$ matrix with elements $a_{ij}$ $(i, j = 1, \ldots, n)$ and $\lambda_1, \ldots, \lambda_n$ are the $n$ real eigenvalues of the matrix $\mathbf{A}^{\mathbf{H}}\mathbf{A}$, where $\mathbf{A}^{\mathbf{H}}$ is the Hermitian conjugate matrix of **A** defined by

$$\mathbf{A}^{\mathbf{H}} = (\bar{\mathbf{A}})^T$$

$\bar{\mathbf{A}}$ denoting the complex conjugate of **A**, so that $\bar{\mathbf{A}}$ has elements $\bar{a}_{ij}$ $(i, j = 1, \ldots, n)$. If **A** is real symmetric then $\mathbf{A}^{\mathbf{H}} = \mathbf{A}$ and so the eigenvalues of $\mathbf{A}^{\mathbf{H}}\mathbf{A}$ are $\mu_1^2, \ldots, \mu_n^2$ where

$\mu_1, \ldots, \mu_n$ are the eigenvalues of $\mathbf{A}$, which are all real. Hence when $\mathbf{A}$ is an $n \times n$ real symmetric matrix then

$$\|\mathbf{A}\|_E = \max_{1 \leqslant i \leqslant n} \{|\mu_i|\} \tag{1.2.38}$$

Finally, let $\|.\|$ be any matrix norm. Then we have the following lemma, which is used in Chapter 7.

*Lemma 1.2.1*

If 1. $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ matrices;
2. $\mathbf{A}$ is nonsingular and $\|\mathbf{A}^{-1}\| \leqslant \alpha$;
3. $\|\mathbf{B} - \mathbf{A}\| \leqslant \beta$;
4. $\alpha\beta < 1$,

then $\mathbf{B}$ is nonsingular and

$$\|\mathbf{B}^{-1}\| \leqslant \frac{\alpha}{(1 - \alpha\beta)}$$

□

For a proof of Lemma 1.2.1 see Exercise 1.2.14.

## 1.3 Fundamental Analysis

In this section we collect some results from the calculus of functions of several variables and describe some notation which is used subsequently. For more detail than it is possible to give here the reader should consult, for example, Russell (1970), Burkill (1970), and Mangasarian (1969). As with Section 1.2 it is not necessary to understand in detail the content of this section if the reader is not primarily interested in the theory underlying the numerical methods which are described subsequently in the text.

It is very convenient to express many mathematical ideas connected with the theory of optimization in a geometrical form. We speak of $\mathbf{x}, \mathbf{y} \in R^n$ as *points* in the linear space $R^n$, and we define a *distance* $\rho(\mathbf{x}, \mathbf{y})$ between the points $\mathbf{x}, \mathbf{y}$ in the normed linear space $\{R^n, \|.\|\}$ by

$$\rho(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| \tag{1.3.1}$$

We generalize the idea of open and closed intervals in $R^1$ by means of open and closed *balls* in $\{R^n, \|.\|\}$.

*Definition 1.3.1*

The set of points $\mathbf{x} \in R^n$ such that $\|\mathbf{x} - \hat{\mathbf{x}}\| \leqslant r$, where $r > 0$ is a given real number and $\hat{\mathbf{x}}$ is a given point in $R^n$ is called the *closed ball* $B[\hat{\mathbf{x}}, r]$ with centre $\hat{\mathbf{x}}$ and radius $r$. Thus

$$B[\hat{\mathbf{x}}, r] = \{\mathbf{x} \in R^n \mid \|\mathbf{x} - \hat{\mathbf{x}}\| \leqslant r\}$$

The *open ball* $B(\hat{\mathbf{x}}, r)$ with centre $\hat{\mathbf{x}}$ and radius $r$ is defined by

$$B(\hat{\mathbf{x}}, r) = \{\mathbf{x} \in R^n \mid \|\mathbf{x} - \hat{\mathbf{x}}\| < r\}$$

□

Using the idea of distance between points in the normed linear space $\{R^n, \|.\|\}$ we can generalize the idea of convergence of a sequence $\{\mathbf{x}^{(k)}\}$ in $\{R^1, |.|\}$.

*Definition 1.3.2*

The sequence $\{\mathbf{x}^{(k)}\}$ in the normed linear space $\{R^n, \|.\|\}$ *converges* to $\mathbf{x}^* \in R^n$ if and only if

$$\lim_{k \to \infty} \|\mathbf{x}^{(k)} - \mathbf{x}^*\| = 0 \qquad \square$$

Because, as we have seen in Section 1.2, all norms on $R^n$ are equivalent, we have

$$\left(\lim_{k \to \infty} \|\mathbf{x}^{(k)} - \mathbf{x}^*\| = 0\right) \Leftrightarrow \left(\lim_{k \to \infty} \|\mathbf{x}^{(k)} - \mathbf{x}^*\|' = 0\right)$$

where $\|.\|$ and $\|.\|'$ are any two vector norms on $R^n$. Hence convergence of $\{\mathbf{x}^{(k)}\}$ to $\mathbf{x}^*$ with respect to any norm on $R^n$ implies convergence of $\{\mathbf{x}^{(k)}\}$ to $\mathbf{x}^*$ with respect to any other norm on $R^n$.

In conformity with the definition of a Cauchy sequence of real numbers, we say that the sequence $\{\mathbf{x}^{(k)}\}$ in the normed linear space $\{R^n, \|.\|\}$ is a *Cauchy sequence* if and only if given $\epsilon > 0$, there exists an integer $l$ such that $\|\mathbf{x}^{(j)} - \mathbf{x}^{(k)}\| < \epsilon$ whenever $j > l$ and $k > l$ simultaneously. It can be shown that if $\{\mathbf{x}^{(k)}\}$ is a Cauchy sequence in $\{R^n, \|.\|\}$ then $\{\mathbf{x}^{(k)}\}$ has a limit $\mathbf{x}^* \in R^n$.

We can generalize the idea of continuity of $f: R^1 \to R^1$ as follows.

*Definition 1.3.3*

A function $\mathbf{f}: D \subseteq R^n \to R^m$ is *continuous* at $\hat{\mathbf{x}} \in D$ if and only if

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\hat{\mathbf{x}})\|_m \to 0$$

whenever

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_n \to 0$$

where $\|.\|_m$ is a norm on $R^m$ and $\|.\|_n$ is a norm on $R^n$. $\square$

By definition 1.3.3, $f: R^n \to R^1$ is continuous at $\hat{\mathbf{x}} \in R^n$ if $|f(\mathbf{x}) - f(\hat{\mathbf{x}})| \to 0$ whenever $\|\mathbf{x} - \hat{\mathbf{x}}\| \to 0$, where $\|.\|$ is any norm on $R^n$.

We shall denote the first partial derivatives of $f: R^n \to R^1$ by $\partial_i f(\mathbf{x})$ $(i = 1, \ldots, n)$ so that

$$\partial_i f(\mathbf{x}) = \frac{\partial}{\partial x_i} f(\mathbf{x}) \quad (i = 1, \ldots, n) \tag{1.3.2}$$

We denote the second partial derivatives of $f$ by $\partial_i \partial_j f(\mathbf{x})$ $(i, j = 1, \ldots, n)$ so that

$$\partial_i \partial_j f(\mathbf{x}) = \frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x}) \quad (i, j = 1, \ldots, n) \tag{1.3.3}$$

Let $f: R^n \to R^1$ have first partial derivatives in $R^n$. Then the *gradient vector* $\mathbf{g}(\mathbf{x})$ of $f$ at $\mathbf{x} \in R^n$ is defined by

$$\mathbf{g}(\mathbf{x}) = [\partial_1 f(\mathbf{x}), \ldots, \partial_n f(\mathbf{x})]^T. \tag{1.3.4}$$

It is sometimes convenient to use the notation $\nabla f(\mathbf{x})$ for the gradient vector of $f$ at $\mathbf{x}$.

Let $f: R^n \to R^1$ have second partial derivatives in $R^n$. Then the *Hessian matrix* $\mathbf{G}(\mathbf{x})$ of $f$ at $\mathbf{x}$ is defined by

$$\mathbf{G}(\mathbf{x}) = [\partial_i \partial_j f(\mathbf{x})] \quad (i, j = 1, \ldots, n) \tag{1.3.5}$$

If $f$ has continuous second partial derivatives in $R^n$ then

$$\partial_i \partial_j f(\mathbf{x}) = \partial_j \partial_i f(\mathbf{x}) \quad (i, j = 1, \ldots, n)$$