Introduction to Numerical Analysis

F. Stummel & K. Hainer

Introduction to
NUMERICAL ANALYSIS

F. STUMMEL AND K. HAINER
University of Frankfurt am Main

Translated by E. R. DAWSON
University of Dundee

Edited by Professor W. N. EVERITT
University of Dundee

SCOTTISH ACADEMIC PRESS
EDINBURGH

Published by
Scottish Academic Press Ltd.
33 Montgomery Street, Edinburgh EH7 5JX

First published 1980

ISBN 7073 0130 0

Printed in Great Britain

PREFACE TO THE GERMAN EDITION

Numerical analysis deals with methods of solving the typical, fundamental, mathematical problems encountered in the various fields in which mathematics is applied in practice, and with the mathematical analysis and handling of these methods. The problems which arise in applications to the fields of science and engineering include, for example, the calculation of special functions, the approximate calculation of the derivatives, and approximate integration, of these functions, the solution of algebraic equations and of systems of linear and non-linear equations, the approximate solution of differential and integral equations, and so on. For practical purposes predominant interest attaches to those methods which make it possible to solve these problems numerically and approximately, using for this purpose the efficient electronic computers which nowadays are generally available.

In this book the knowledge of the differential and integral calculus and of linear algebra customarily acquired in the first year's work at a university is presupposed. The numerical exercises are designed so that they can be solved during a course of practical work on the computer by students who have previously taken a fortnight's course in some problem-oriented programming language such as Algol or Fortran. The numerical results for these exercises were computed in this way on the Univac 1108 computer of the computing centre of Frankfurt University, for instance. The problem-oriented programming languages of big modern computer systems make computations in the field of complex numbers possible without extra difficulty. A number of problems such as, for instance, the determination of the zeros of polynomials or the eigenvalues of matrices are, in general, completely soluble only in the field of complex numbers. For initial-value problems for ordinary differential equations the solution of the equation in the complex field is also of importance. For these reasons the methods and processes in this book have been formulated, so far as this could be done without extra trouble, for a field **K** denoting either the field **R** of the reals or the field **C** of the complex numbers.

The limited size of this textbook necessitated limitation and selection of the methods and processes to be presented. No attempt has been made in this book to present the greatest possible number of different methods, even merely as "recipes" for the individual problems. Rather we discuss the fundamental concepts and the properties of typical methods. With this background knowledge the reader will be able to follow up particular topics by means of the comprehensive books and standard works mentioned at the beginning of each section, for example. In the space available, too, the reproduction of numerical results has to be limited; the numerical exercises given at the end of each section should serve to illustrate the methods, but they cannot always pretend to characterize typical numerical properties.

This book originated from lectures and practical courses which the first-named author has given regularly for several years at the University of Frankfurt, and from lecture notes for the practical mathematical course written by both authors jointly.

A brief survey of the presentation and contents of the book follows, together with some indications of results which stem from other lecture notes and unpublished work by the first-named author. Chapter I acquaints the reader with the concepts and methods which are needed to calculate functions and their zeros, and to make best use of the relevant handbooks. Chapter II is concerned with interpolation and approximate differentiation and integration. In addition to the classical quadrature formulae, which are derived by means of the Hermite interpolation formula, we discuss Romberg integration and the general Gaussian quadrature formulae. The subjects of Chapters III and IV are approximate methods for solving systems of linear and non-linear algebraic equations and the solution of eigenvalue problems for matrices. Section 5 therefore presents the basic concepts for finite-dimensional arithmetic spaces, *viz*., norms of vectors and matrices, inner-products, and the most important properties of eigenvalues and eigenvectors of matrices. The numerical methods for solving systems of linear equations can be classified as the methods of elimination, orthogonalization, and iteration. In order to solve systems of non-linear equations by the Jacobi method or the Gauss-Seidel method, we discuss in Section 9 the method of successive approximations for contractive mappings and for monotone mappings, and also the Newton methods. In the solution of eigenvalue problems we restrict our attention to the symmetric case and the power method, the classical Jacobi method, Krylov's method with Lanczos-orthogonalization to obtain a tridiagonal matrix and a Sturm sequence for calculating the eigenvalues, to inclusion theorems and *a posteriori* error-estimates for the eigenvalues and eigenvectors. From the extensive range of numerical methods for solving differential equations, we have selected in Chapter V the one-step and multi-step methods for the numerical integration of initial-value problems for ordinary differential equations. Here we not only derive the approximation equations for these methods, but also prove the convergence and stability of the methods.

In approximate methods the corresponding error-estimates are of fundamental importance. The so-called *a posteriori* error-estimates, by which the difference between the calculated approximate solution and the required solution can be estimated from the residual obtained by a trial substitution of the approximation, are of particular importance. For the solution of systems of non-linear algebraic equations we obtain in Section 9.2 an error-estimate of this kind, which does not depend on any particular method. A corresponding error-estimate for the case of a single variable is proved earlier in Section 2.2; from this result, moreover, the convergence of *regula falsi* even in the complex field can be easily proved. In exactly the same way *a posteriori* error-estimates can be given for eigenvalue problems for matrices. For the approximations to the eigenvalues there are the well-known comparison and inclusion theorems. Further, in Section 10.4, for the symmetric case, approximations to the eigenvectors are also estimated by the norm of the residual arising from a trial substitution, and this at the same time gives, by means of the square of the norm of the residual, an interesting error-estimate for the Rayleigh quotient regarded as an approximation to an eigenvalue. In several places in the book orthogonal polynomials are needed. So in Section 7.4 we present a simple, purely algebraic theory of orthogonal polynomials, with propositions about their zeros, recursion formulae, and particular representations, all of which are applicable to the classical orthogonal polynomials with the Gaussian quadrature formulae and also to the orthogonal Lanczos-polynomials of a matrix and to the minimal polynomial of a vector. The idea of the spectral radius is not introduced here for the iterative solution of a system of linear equations. In Section 8.3 we prove the convergence of the Gauss-Seidel method or, more generally, of the corresponding relaxation methods for positive-definite matrices, after defining a norm suited to the problem, by means of the method of contractive mappings. The one-step methods and multi-step methods for initial-value problems present an important and typical application of the fundamental concepts of consistency, convergence, and stability, and we show that for Lipschitz-continuous methods consistency and convergence are equivalent.

In the preparation of the manuscript Miss A. Raymond drew the diagrams, put in the formulae, and helped with the proof-reading, and we wish here to thank her for her active and valued collaboration. We also thank Miss E. Hilbricht for her careful execution of the extensive typewriting work. Finally we thank the Teubner Publishing House for their efficient production of our textbook.

F. STUMMEL, K. HAINER

Frankfurt am Main
Summer 1971

FOREWORD TO THE ENGLISH EDITION

This book is the English translation of our book *Praktische Matematik* (published in 1971). Since that time economically-priced micro-computers and mini-computers have become common. Consequently the possibilities of studying and trying out numerical methods, as well as using them in practice, have been extended to a remarkable degree. Nearly all the numerical exercises described in this book can be carried out even on pocket-sized, programmable calculators, and students will be able in this way to gain a far clearer insight into them.

The authors cordially thank the editor, Professor W. N. Everitt, for his work, Mr. E. R. Dawson for the translation, and the publishers, the Scottish Academic Press, for accepting this book into their series.

F. STUMMEL, K. HAINER

Frankfurt am Main
October 1976

CONTENTS

I CALCULATION OF FUNCTIONS AND THE ZEROS OF FUNCTIONS

In practice the task of calculating functions is, as a rule, part of some wider problem. In many fields of natural science and energineering, when problems are stated in mathematical terms, ordinary functions or higher transcendental functions appear in the mathematical formulation. This chapter presents a survey of important analytical ideas about the representation, approximation, and calculation of such functions. The object is to make the reader sufficiently well acquainted with representations by infinite series, asymptotic expansions, and continued fractions, and with Chebyshev approximations, as will be desirable if he is to use handbooks on special functions.

The calculation of the zeros of given functions is also one of the basic tasks in practical mathematics. The calculation of the zeros of polynomials, or the calculation of eigenvalues from the characteristic polynomial of a matrix, requires in the general case methods of approximation which are applicable not only in the real field but also in the complex field. Some of the approximation methods in this chapter will later be generalized in Chapter IV to systems of non-linear equations in an n-dimensional arithmetic space, and thus represent an important instance of these general methods. Particular importance has been attached to the derivation and formulation of *a priori* and *a posteriori* error-estimates in approximation methods for the calculation of zeros.

1. Calculation of Functions

In this section we consider some methods for defining special functions and calculating them numerically. Polynomials in a real or complex variable have particularly favourable properties for numerical purposes; the Horner method is available as a simple algorithm for calculating the values of the function and its derivatives. Consequently, polynomials form an important auxiliary means of approximating functions; for instance, in the form of the initial part of a power series or of an asymptotic expansion, as a Chebyshev approximation, an interpolation polynomial, or a smoothing parabola. The great handbook of mathematical functions by Abramowitz and Stegun lists a large number of special functions which can be expanded in power series, asymptotic series, and continued fractions. By taking the initial part (of suitable length) of these expansions one can then approximate the functions and calculate their approximate values. Nevertheless in many cases one today makes use of best or almost best approximations in the Chebyshev sense in calculating special functions. In the handbook just mentioned these are listed under "polynomial approximations" and "rational approximations". Chebyshev approximations are generally most suitable for calculating functions because the error of these approximations is uniformly minimal in the interval considered.

References: Abramowitz and Stegun, Hart *et al.*

1.1. Polynomials

For calculating the function-values of polynomials and their derivatives, the algorithm known as *Horner's method* is available. By an *algorithm* we mean a recipe which lays down and prescribes each step in some calculation. In the numerical calculation of function-values of a polynomial, for instance, because of rounding errors the order in which individual steps of the calculation are carried out is by no means a matter of indifference. The Horner method is also used for separating out linear factors from a polynomial and for calculating the coefficients in the Taylor expansion of a polynomial at a given point. We use **K** to denote either the field **R** of real numbers or the field **C** of complex numbers. Let p_n be a polynomial of the nth degree, of the form

$$p_n(x) = a_0 + a_1 x + \ldots + a_n x^n, \quad x \in \mathbf{K},$$

with coefficients $a_0, \ldots, a_n \in \mathbf{K}$.

To calculate the function-value $p_n(x_0)$ at a given point $x_0 \in \mathbf{K}$ we start from the representation

$$p_n(x_0) = a_0 + x_0(a_1 + x_0(\ldots + x_0(a_{n-2} + x_0(a_{n-1} + x_0 a_n)) \ldots)).$$

So if the numbers $a_0', \ldots, a_n'$ are determined according to the rule

$$(1) \quad a_n' = a_n, \quad a_k' = a_k + x_0 a_{k+1}', \quad k = n-1, \ldots, 1, 0,$$

then the required function-value is obtained in the form $p_n(x_0) = a_0'$.

The new coefficients $a_0', \ldots, a_n'$ define a polynomial of degree $n-1$ by the equation

$$(2) \quad p_{n-1}(x) = a_1' + a_2' x + \ldots + a_n' x^{n-1}, \quad x \in \mathbf{K}.$$

As may be verified immediately, this polynomial satisfies the relation

$$(3) \quad p_n(x) = r_0 + (x - x_0) p_{n-1}(x), \quad x \in \mathbf{K}, \qquad p_n(x_0) = a_0' = r_0.$$

Accordingly, p_{n-1} arises when the linear factor $x - x_0$ is removed from p_n, and r_0 is the remainder when p_n is divided by $x - x_0$. The relation (3) also yields the representation

$$(4) \quad \frac{p_n(x) - p_n(x_0)}{x - x_0} = p_{n-1}(x), \qquad x \neq x_0, \quad x \in \mathbf{K},$$

for the difference quotient, and hence, for $x \to x_0$, the representation

$$(5) \quad \frac{dp_n}{dx}(x_0) = p_{n-1}(x_0).$$

To calculate $p_{n-1}(x_0)$ the method just described is again applied to determine the coefficients a_k'', $k = 1, \ldots, n$, and the polynomial p_{n-2} with the property

$$p_{n-1}(x_0) = a_1'' = r_1, \quad p_{n-1}(x) = r_1 + (x - x_0) p_{n-2}(x), \quad x \in \mathbf{K}.$$

By continuing to take out linear factors $(x - x_0)$ we obtain generally a finite sequence of polynomials $p_n, p_{n-1}, p_{n-2}, \ldots, p_0$ with the property

$$(6) \quad p_{n-j}(x) = r_j + (x - x_0) p_{n-j-1}(x), \quad x \in \mathbf{K}, \quad j = 0, 1, \ldots, n-1,$$

and $p_0 = r_n$,

and so we have the representation

$$p_n(x) = r_0 + r_1(x - x_0) + \ldots + r_n(x - x_0)^n, \quad x \in \mathbf{K}.$$

This is the Taylor expansion of the given function p_n at the point x_0. Hence we have for the derivatives of p_n

$$r_j = \frac{1}{j!} \frac{\mathrm{d}^j p_n}{\mathrm{d}x^j}(x_0), \quad j = 0, \ldots, n.$$

If we let $a_k^{(j)}$ denote the coefficients of the polynomial p_{n-j} in the form

$$p_{n-j}(x) = a_j^{(j)} + a_{j+1}^{(j)} x + \ldots + a_n^{(j)} x^{n-j}, \quad x \in \mathbf{K}, \quad j = 0, 1, \ldots, n,$$

then the general rule for calculating these coefficients reads:

$$(7) \quad a_n^{(j+1)} = a_n^{(j)}, \quad a_k^{(j+1)} = a_k^{(j)} + x_0 a_{k+1}^{(j+1)}, \quad k = n-1, \ldots, j,$$

for $j = 0, \ldots, n$ or $n - 1$, and the following relation holds:

$$(8) \quad r_j = p_{n-j}(x_0) = \frac{1}{j!} \frac{\mathrm{d}^j p_n}{\mathrm{d}x^j}(x_0) = a_j^{(j+1)}, \quad j = 0, \ldots, n.$$

The coefficients $a_k^{(j)}$ form the general Horner scheme with $a_k^{(0)} = a_k$, $k = 0, \ldots, n$, in the form

$$\begin{array}{cccccc}
a_n^{(0)} & a_{n-1}^{(0)} & \ldots & a_2^{(0)} & a_1^{(0)} & a_0^{(0)} \\
a_n^{(1)} & a_{n-1}^{(1)} & \ldots & a_2^{(1)} & a_1^{(1)} & r_0 \\
a_n^{(2)} & a_{n-1}^{(2)} & \ldots & a_2^{(2)} & r_1 & \\
\cdot & \cdot & & & & \\
\cdot & \cdot & & & & \\
\cdot & \cdot & & & & \\
a_n^{(n)} & r_{n-1} & & & & \\
r_n & & & & &
\end{array}$$

In order to illustrate the method of working of the Horner scheme, we give finally the scheme of coefficients $a_k^{(j)}$ for the polynomial $p(x) = x^n$.

Example

1	0	0	$\ldots$	0	0	0
1	x_0	x_0^2	$\ldots$	x_0^{n-2}	x_0^{n-1}	x_0^n
1	$2x_0$	$3x_0^2$	$\ldots$	$(n-1)x_0^{n-2}$	nx_0^{n-1}	
1	$3x_0$	$6x_0^2$	$\ldots$	$\frac{n(n-1)}{2}x_0^{n-2}$		
.	.					
.	.					
.	.					
1	$(n-1)x_0$	$\frac{(n-1)n}{2}x_0^2$				
1	nx_0					
1						

1.2. Infinite Series

Convergent series can be approximated by their partial sums. The corresponding error-estimates obtained from convergence criteria should also be given. In the Leibniz criterion for alternating series, as is well-known, the error is given by the first term neglected of the series. But with ratio tests, too, we can easily give a bound for the remainder term by means of the first term neglected of the series.

Let $a_0, a_1, a_2, \ldots$ be an infinite sequence of real or complex numbers. Then the corresponding infinite series

$$(1) \quad a_0 + a_1 + a_2 + \ldots = \sum_{k=0}^{\infty} a_k$$

converges if the sequence of partial sums s_n of the series converges to a number s, and hence the corresponding remainder terms r_n form a null-sequence,

$$(2) \quad s_n = \sum_{k=0}^{n} a_k \to s \iff r_n = s - s_n = \sum_{k=n+1}^{\infty} a_k \to 0 \quad (n \to \infty).$$

1.2.1. Error Estimate from the Leibniz Criterion. The Leibniz criterion indicates the convergence of alternating series of real terms $a_0, a_1, a_2, \ldots$ whose absolute values form a monotone null-sequence

$$(3) \quad a_k a_{k+1} \leqslant 0, \quad |a_{k+1}| \leqslant |a_k|, \quad |a_k| \to 0 \quad (k \to \infty).$$

Since questions of convergence do not arise for series of finitely many terms, the Leibniz criterion can be stated as follows.

(4) *A real infinite series* (1) *the terms of which alternate from some index* $N+1$ *onwards and are such that their absolute values form a monotone null-sequence will converge to a number* s*; an estimate for the error is given by*

$$|r_n| = |s - s_n| = \Big|\sum_{k=n+1}^{\infty} a_k\Big| \leqslant |a_{n+1}|, \quad n \geqslant N.$$

Proof. We consider first the case of an alternating series. Without loss of generality we may assume that $a_0 = |a_0|, a_1 = -|a_1|, a_2 = |a_2|, \ldots$. Under the above hypothesis the partial sums then satisfy the relation

$$s_{2j+1} \leqslant s_{2j+3} \leqslant s_{2j+2} \leqslant s_{2j}, \quad j = 0, 1, 2, \ldots,$$

and

$$s_{2j} - s_{2j+1} = |a_{2j+1}| \to 0 \quad (j \to \infty).$$

The two sub-sequences of partial sums s_{2j}, s_{2j+1} $(j = 0, 1, 2, \ldots)$ are therefore monotone convergent to the same limit s, and so the series (1) also converges to s.

We also have the two inequalities

$$s = a_0 + a_1 + a_2 + a_3 + \ldots = (|a_0| - |a_1|) + (|a_2| - |a_3|) + \ldots \geqslant 0$$

and

$$s = a_0 + a_1 + a_2 + a_3 + \ldots = a_0 - (|a_1| - |a_2|) - \ldots \leqslant |a_0|,$$

and hence the estimate $|s| \leqslant |a_0|$.

In the general case with an index N, for each $n \geqslant N$, the sequence $a_{n+1}, a_{n+2}, \ldots$ is alternating and, in absolute value, is a monotone null-sequence. So the series $s' = \sum_{k=n+1}^{\infty} a_k$ converges and satisfies the estimate $|s'| \leqslant |a_{n+1}|$.

Hence, finally, the given series (1) converges, and the remainder of the series $r_n = s'$ satisfies the stated inequality. □

Example. *The Bessel function* J_ν has the series expansion (see *Abramowitz and Stegun*, No. 9)

$$(5) \quad J_\nu(x) = \sum_{k=0}^{\infty} \frac{(-1)^k (x/2)^{2k+\nu}}{k!\,\Gamma(k+1+\nu)}.$$

For real $x \in \mathbf{R}$ and real indices $\nu \geqslant 0$, this is an alternating series. For $k = 0, 1, 2, \ldots$ with $k + 1 \geqslant |x|/2$, the absolute values of the terms a_k of the series form a monotone null-sequence, since

$$|a_{k+1}| = \frac{1}{(k+1)(k+1+\nu)} \left(\frac{x}{2}\right)^2 |a_k| \leqslant |a_k|.$$

The above theorem therefore permits the representation

$$(6) \quad J_\nu(x) = \sum_{k=0}^{n-1} \frac{(-1)^k (x/2)^{2k+\nu}}{k!\,\Gamma(k+1+\nu)} + r_{n-1}(x)$$

with the error estimate

$$(7) \quad |r_{n-1}(x)| \leqslant \frac{1}{n!\,\Gamma(n+1+\nu)} \left(\frac{x}{2}\right)^{2n+\nu}, \quad \left|\frac{x}{2}\right| \leqslant n + 1.$$

For example, for $\nu = 0, 1, 2, \ldots$ and $n = 7$, we obtain the estimate

$$(8) \quad |r_6(x)| \leqslant 4 \cdot 10^{-8}, \quad |x| \leqslant 2.$$ □

1.2.2. Error-Estimate from a Ratio-test. For an infinite series of the form (1) with real or complex terms we can also give an easily calculable error-estimate derived from the ratio-test.